applied sciences

MDPI

Review

A Review of Transformer-Based Approaches for Image Captioning

Oscar Ondeng *, Heywood Ouma and Peter Akuon

Department of Electrical and Information Engineering, University of Nairobi, Nairobi P.O. Box 30197-00100, Kenya; houma@uonbi.ac.ke (H.O.); akuon@uonbi.ac.ke (P.A.)
* Correspondence: oscaror@uonbi.ac.ke

Abstract: Visual understanding is a research area that bridges the gap between computer vision and natural language processing. Image captioning is a visual understanding task in which natural language descriptions of images are automatically generated using vision-language models. The transformer architecture was initially developed in the context of natural language processing and quickly found application in the domain of computer vision. Its recent application to the task of image captioning has resulted in markedly improved performance. In this paper, we briefly look at the transformer architecture and its genesis in attention mechanisms. We more extensively review a number of transformer-based image captioning models, including those employing vision-language pre-training, which has resulted in several state-of-the-art models. We give a brief presentation of the commonly used datasets for image captioning and also carry out an analysis and comparison of the transformer-based captioning models. We conclude by giving some insights into challenges as well as future directions for research in this area.

Keywords: computer vision; convolutional neural networks; image captioning; MS COCO; CIDEr; natural language processing; feature extraction and representation; general attention; self-attention; transformers; vision-language pre-training; multimodal alignment

check for updates

Citation: Ondeng, O.; Ouma, H.; Akuon, P. A Review of Transformer-Based Approaches for Image Captioning. *Appl. Sci.* **2023**, *13*, 11103. https://doi.org/10.3390/app131911103

Academic Editor: Andrea Prati

Received: 23 June 2023
Revised: 18 August 2023
Accepted: 29 August 2023
Published: 9 October 2023

1. Introduction

Given the abundance of visual information in the form of images and videos, image processing techniques are useful in automating tasks which would otherwise be very difficult for humans to carry out due to the sheer abundance of the visual information. Visual information understanding is a task that humans perform effortlessly. This is often manifested in our ability to capture visual information and express or communicate it through natural language.

In an effort to bridge the gap between image processing algorithms and natural language understanding, various visual understanding tasks have received significant attention in recent years. Image captioning is one such visual understanding task. Image captioning refers to the visual understanding task of expressing images using natural language. Similarly to other areas of machine learning, the major improvements in image captioning can, to a large extent, be attributed to the recent developments in deep learning—bigger and more refined datasets, faster hardware, especially graphical processing units, and better algorithms.

However, image captioning still remains a challenging task. Captioning in the wild and out-of-domain captioning are difficult tasks which show that the generated captions have a lot of room for improvement. Captioning, therefore, continues to receive a lot of research attention in a bid to refine the related algorithms. This is because image captioning has many potential applications such as in surveillance and security, robotics, self-driving cars, assistance to the visually impaired, etc.

Many of the earlier approaches to image captioning consisted of template-based methods [1–3] and composition-based methods [4,5]. The template-based methods hinged on generating templates for captions. The slots of the templates were then completed

based on the results of object detection, attribute classification and scene recognition. The composition-based methods employed existing image-caption databases to extract components of related captions and compose them together to come up with new descriptions. Advances in neural networks led to neural-based methods. These capitalized on the use of convolutional neural networks (CNN) to carry out the feature extraction. The use of CNNs was inspired by their success in the tasks of image classification [6–12] and object detection [13–16]. The work of [17] involved an early use of neural networks for caption generation, employing a feed-forward neural network that uses a given image and previous words to predict the following word. The feed-forward neural network was then replaced by a recurrent neural network (RNN) [18] and later the limitations of the RNN in regard to gradient propagation led to the use of the LSTM RNN [19] to decode the extracted image features into a string of words that form a caption [20,21].

In the neural-based methods, a common framework that emerged was the encoder-decoder framework. The encoder-decoder framework was born of language modeling [22–25]. Kiros et al. [26] were among the first to employ the encoder-decoder framework in image captioning. They were soon followed by a number of other authors who used CNNs and LSTMs for encoding and decoding [20,21,27]. The early use of the encoder-decoder framework in captioning consisted in encoding an image into an embedding space such that it could then be used as an input to downstream decoder to generate textual tokens. Many models used a variant of CNNs and LSTMs for encoding and decoding, respectively. This architecture has proven to be very powerful and many of the current captioning systems employ a variant of it.

A significant advancement in image captioning was the introduction of attention mechanisms to the encoder-decoder framework. The attention models drew inspiration from attention as used in machine translation [22] and then object detection [28,29]. Through the attention mechanisms, captioning models learn to attend to various aspects of the input image as the captions are being generated. Attention mechanisms have been widely employed in a number of captioning models [30–35]. In the model of Lu et al. [33], they develop an adaptive attention mechanism which learns whether or not to attend to the image depending on the context and word token being generated. Anderson et al. [35] designed a bottom-up top-down design, which extracts visual features based on object detectors rather than the last convolutional layers of a CNN as in previous feature extractors. This model was influential for several other successive designs.

The initial attention mechanisms gave rise to self-attention, which was the underlying concept that was then employed in the development of multi-head attention in the transformer model [36]. The transformer results from abstracting the multi-head self-attention operations into a self-contained unit. Stacking such units provides the necessary non-linearity and representational power to model complicated functions [37]. These recent developments were initially applied to machine translation and were thereafter transferred to the visual domain. Since the initial presentation of the transformer model in 2017 [36], it has proven to be a powerful basis on which many of the current state-of-the-art models have been designed [38–40].

In this paper, we review a number of transformer-based image captioning models leading up to the current state-of-the-art. Other reviews of image captioning models have been completed previously [41–43] but none of these have focused on transformers, which are a more recent invention. Many of the previous reviews have for instance focused on deep learning per se as applied to image captioning [42] or on the prior concepts such as attention [41,43]. Khan et al. [44] look broadly at transformers in vision but do not focus on the task of image captioning. Stefanini et al. [45] also give a review that is closely related to our work. They explore a broad array of captioning methods. They deal with earlier approaches and as well as the more recent approaches in image captioning. Our review, in contrast, deals specifically with transformer-based approaches given that transformers have been the cornerstone of the current state-of-the-art in image captioning as well as other computer vision and NLP tasks. Furthermore, the earlier approaches have already been

amply covered in previous surveys. Our focused approach, therefore, allows us to delve deeper into the more recent transformer-based approaches making up the current state-of-the-art in image captioning and assesses the challenges and possible future directions in this specific field. Compared to [45], we offer a wider range of insights into future directions.

This paper is organized as follows: we first look at the common datasets and evaluation metrics that are relevant for image captioning. We then review the early vanilla transformer-based approaches that were used for image captioning. We then look at developments in vision-language pre-training and discuss how this approach has been used in the context of transformer models and image captioning. We then look at a number of specific transformer-based models involving vision-language pre-training. This is followed by a discussion and analysis that assesses the major contributions of the various methods studied as well as a comparison of those methods based on their performance results. The paper concludes with a discussion on the open challenges and future directions of the field. This organization helps appreciate the impact that transformers have had in image captioning as well as the effectiveness of vision-language pre-training.

2. Datasets and Evaluation

In this section we take a look at the datasets that have been employed by the various transformer-based image captioning models which we review in Sections 4 and 5. We also look at the main evaluation metrics used to gauge and compare different models.

2.1. Datasets

The datasets have been used for training, validation and evaluation. The data is a crucial factor of the performance of many of the machine learning algorithms. The availability of vast amounts of data has resulted in better-trained models with significantly improved levels of accuracy.

Table 1 shows a listing of the main datasets that have been used in relation to image captioning. As can be seen from the sizes, the general trend has been towards larger and larger datasets in terms of number of images and the language features (captions, questions and other textual annotations). The collection and curation of captions and other language features of the large-scale datasets tends to be automated rather than human-annotated, given the high cost of human annotation. The textual description (image-text pairs) for these large-scale datasets are largely acquired through the alt-text HTML attribute of images from the internet. Despite the fact that these large-scale datasets are noisier and less clean, they have proven to be effective in vision-language pre-training. For the smaller datasets, the annotation is commonly created through human annotators via services such as the Amazon Mechanical Turk [46].

The categories of the datasets locate inspiration to a large extent from WordNet [47,48], together with its hierarchical structure and synsets. The synsets are sets of cognitive synonyms, each expressing a distinct concept [47]. Some of the categories that occur in many of the datasets are taken from wordnet subtrees and include *mammal, bird, fish, reptile, amphibian, vehicle, furniture, musical instrument, geological formation, tool, flower, fruit* [49]. Datasets such as MS COCO [50] tend to have a high proportion of common objects such as people, animals, furniture and vehicles. The large-scale datasets, such as OpenImages [51], ALT200M [52], WebImageText [53] and ALIGN [54] have a higher variety of categories due to their sheer scale and method of collection (automated, from web images). WebLI [55] ramps up the scaling further and is made up of 10 billion images and the corresponding textual descriptions.

Three datasets worthy of particular mention are ImageNet [49,56], MS COCO [50] and Nocaps [57]. ImageNet spurred a lot of research in deep learning models for computer vision and it contributed significantly to better models related to vision tasks such as image classification and object detection. ImageNet was the basis for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [7], which saw convolutional neural networks

(CNN) come of age and achieve and surpass human-level performance in some of the vision tasks.

Table 1. The major datasets used in relation to image captioning. For the primary domains: IC is image classification, VLP is vision-language pre-training, OD is object detection, IS is image segmentation, Cap is image captioning, D-Cap is dense captioning, ARM is attribute and relationship modeling, VQA is visual question answering, ISR is image-sentence retrieval.

Dataset Name	Size (Images)	Categories	Language Features	Primary Domain	Human-Annotated
ImageNet [49]	14.1 M	21,841	14 M image-level labels	IC, OD, VLP	✓
MS COCO [50]	300,000	91	413,915 captions	IC, OD, IS, Cap	✓
Visual Genome [58]	108,000	33,877	5.4 M region descriptions	ARM, D-Cap, VQA, VLP	✓
CC3M [59]	3.3 M	16,000	3 M image-text pairs	IC, VLP	✗
CC12M [60]	12 M	-	12 M image-text pairs	VLP	✗
Flickr 8 K [61]	8000	-	40,460 captions	Cap, ISR	✓
Flickr 30 K[62]	30,000	-	15,8915 captions	Cap, ISR	✓
Nocaps [57]	15,100	513	166,100 captions	Out-of-domain captioning	✓
VQA [63]	254,721	-	760 K questions	VQA, VLP	✓
VQAv2 [64]	265,016	-	1.1 M questions	VQA, VLP	✓
GQA [65]	113,000	-	22 M questions	VR, VLP	✗
SBU Captions [66]	1 M	89	1,000,000	Cap	✗
OpenImages [51]	9 M	600	61.4 M image-level labels	IC, OD, ARM, VLP	Partially
Objects365 [67]	600,000	365	10 M object labels	OD, VLP	✓
ALT200M [52]	200 M	-	200 M image-text pairs	VLP	✗
WebImageText [53]	400 M	-	400 M image-text pairs	VLP	✗
ALIGN [54]	1.8 B	-	1.8 B image-text pairs	VLP	✗
WebLI [55]	10 B	-	10 B image-text pairs	VLP, Cap	✗

MS COCO [50] has also been a major driver in the development of computer vision applications. It has fostered state-of-the-art algorithm development in areas such as image classification, object detection, image segmentation and image captioning. Many of the image captioning algorithms developed since 2014 have used this dataset as a basis for evaluation and validation. In developing their method for deep visual-semantic alignments for generating image captions, Karpathy and Fei-Fei [27] used a subset of the MS COCO Captions 2014 dataset and divided the subset into training, validation and evaluation subset. This division has come to be called the *Karpathy split* and has been used by many subsequent models as a basis for offline comparison for different algorithms. In the Karpathy split, the validation set of the MS COCO 2014 dataset is divided into a 'val' split and a 'test' split, each with 5000 images, and a 'restval' split with 30,504 images. The 'train' split has 82,783 images. The 'restval' and 'train' splits can be combined to make a total of 113,287 training images. The online evaluation server associated with the MS COCO dataset has also been useful for algorithm development. The server is hosted by CodaLab [68] and has served as a tool for different state-of-the-art models to be objectively evaluated and compared against each other. The leaderboard [69] has been particularly instrumental as a basis of comparison.

The Novel Object Captioning at Scale (Nocaps) [57] is a dataset that enables novel captioning at scale. It consists of 15,100 images each with 11 human-generated captions resulting in 166,100 captions. The images are split into 4500 and 10,600 images for the validation and test sets, respectively. The images are taken from the Open Images object detection validation and test sets. The training data consists of MS COCO image-caption pairs. Open Images has more object classes than MS COCO and so when training is conducted on MS COCO, the test set of Nocaps has object classes not seen during training. Nocaps encourages models to learn a large variety of visual concepts from alternative and diverse data sources; these visual concepts may not be in the actual training data. It

therefore makes possible the design of models which can run much better in the wild with novel object categories.

2.2. Evaluation Metrics

In this subsection we take a look at the main evaluation metrics that are used to gauge and compare models. The metrics are summarized in Table 2. The evaluation metrics provide an avenue for different researchers to compare their works. Since captioning has the natural language component as a major output, the evaluation metrics are largely taken from the language domain, where they are for instance applicable in other tasks such as machine translation.

Table 2. The main evaluation metrics used in image captioning. The 'Correlation' column refers to human-judgment correlation.

Metric	Basis	Correlation	Original Domain	Operation
BLEU [70]	Precision-based	Low	Summarizing and translation	Co-occurrence of n-grams
ROUGE [71]	Recall-based	Moderate	Summarizing and translation	Longest common sequence (ROUGE-L)
METEOR [72]	Precision and recall	Moderate	Summarizing and translation	Matching and comparison of n-grams
CIDEr [73]	Precision and recall	High	Image captioning	Cosine similarity of n-grams
SPICE [74]	Precision and recall	High	Image captioning	Sentence comparison using scene graphs

The BiLingual Evaluation Understudy (BLEU) metric [70] was originally designed for sentences resulting from machine translations. It is precision-based and it analyzes the co-occurrence of n-grams between the generated sentence and the reference sentence and calculates an error metric. The matches are position-independent so that a test sentence with more matches with the reference sentence is deemed as a better match than one with fewer matches. The BLEU scores are usually reported in terms of the cumulative scores, which are calculated based on the individual n-gram scores at all orders from 1 to n and weighted using the geometric mean. Thus, if n is 4, cumulative scores for BLEU-1, BLEU-2, BLEU-3 and BLEU-4 are given.

ROUGE [71] refers to Recall-Oriented Understudy for Gisting Evaluation. It is an evaluation algorithm that was originally designed to evaluate text summarization algorithms. ROUGE is recall-based and it counts the number of overlapping units such as word sequences, n-grams and word pairs between generated captions and ground truth captions. A number of variants or ROUGE exist, namely ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S. ROUGE-N measures the n-gram recall between a candidate text and a set of reference texts. ROUGE-L measures the longest common subsequence between the candidate and reference texts. ROUGE-W is a version of ROUGE-L that incorporates weighting. ROUGE-S measures the skip-bigram co-occurrence statistics. Skip-bigrams are all pairs of ordered words in a piece of text, sampled with an arbitrary distance between them. ROUGE-L is the variant commonly used for image captioning since it has a higher correlation with human judgements [71].

METEOR [72] stands for Metric for Evaluation of Translation with Explicit ORdering and is an automatic metric originally designed for machine translation. It was designed to overcome a number of weaknesses observed in the BLEU metric that preceded it such as the lack of recall and the lack of explicit word matching [72]. METEOR matches different aspects of n-grams from different sentences to find alignments and then evaluates a score. The matching of the n-grams can be based on aspects such as surface forms, stemmed forms and meanings. METEOR uses a combination of unigram-precision and unigram-recall. METEOR has shown better correlation than BLEU and ROUGE to judgments by human subjects.

CIDEr [73] refers to Consensus-based Image Description Evaluation. CIDEr makes an effort to capture human judgment of consensus. CIDEr measures how similar a sentence is to the consensus description of an image, i.e., to how an image is described by most people. It is based on evaluations (average cosine similarity) performed on n-grams taken

from the generated captions and the ground truth captions. CIDEr has been shown to have a higher correlation with human judgments than BLEU, ROUGE and METEOR [72]. A version of CIDEr called CIDEr-D is used in the MS COCO evaluation server. CIDEr-D is an advanced version of CIDEr and is designed to make it more resistant to gaming, which refers to a situation in which a sentence that is poorly judged by humans scores highly when evaluated using automatic metrics.

SPICE [74] stands for Semantic Propositional Image Caption Evaluation. It primarily tries to overcome the sensitivity of the other captioning evaluation metrics to n-gram overlap, which can give misleading evaluations. SPICE takes advantage of the semantic structure of scene descriptions and gives preference to nouns. It makes use of scene graphs to compare sentences. Scene graphs are graph-based semantic representations of sentences which help abstract away lexical and syntactic idiosyncrasies. Experiments performed showed that evaluations by SPICE had a higher correlation to human judgments than the other metrics [74].

3. Method Selection

In this section, we describe the strategy that we used to select the models reviewed. There have been many approaches to carrying out image captioning. However, in this review we are mainly interested in transformer-based approaches since these constitute the current state-of-the-art. We, therefore, do not include in our review models which were developed prior to the design of the transformer model in 2017 [36]. Furthermore, there have already been other reviews which have amply studied the previous approaches of image captioning [41–43].

We divide the transformer-based approaches into two broad categories: those prior to vision-language pre-training (vanilla transformer-based models) and those involving vision-language pre-training (VLP models). The former methods were developed earlier in time (shortly after the initial design of the transformer model). The latter models arose after the recent advancements in pre-training. Vision-language pre-training has had a significant impact on multi-modal models involving the visual and language domains and the current state-of-the-art models employ some form of vision-language pre-training. To illustrate this, we first study a number of vanilla transformer-based models and thereafter we assess a number of VLP models.

We are interested in models that employ publicly available datasets, especially MS COCO [75], since this has for a long time served as a good basis for comparing various works. We are particularly interested in models that have performance metrics based on the online and public MS COCO leaderboard [69] since this adds an additional level of objectivity. We are also interested in models that at least give performance using the CIDEr metric [73], due to its high correlation with human judgements. Nocaps [57] is a more recent dataset and interesting for out-of-domain captioning but we preferred to leave the study wider and include methods that do not necessarily report on Nocaps.

Given the plethora of models that have sprung up and the clear limitation of not being able to include them all in our review, we mainly focus on those with a significant design component over and above the initial design of the transformer. This means that we have not included models that have only very slight adjustments, for instance mere changes in the parameters such as size.

The selection strategy has resulted in the models reviewed being representative of the main recent approaches of transformer-based models used in image captioning.

4. Transformers in Captioning

Following the success of transformers in language modeling, they have been applied to several areas of computer vision. In image captioning, transformers are the backbone of the current high-performing models. Figure 1 highlights the evolution of captioning systems from the traditional encoder-decoder models to the basic transformer-based captioning systems to the transformer-based captioning models that employ vision-language pre-

training (VLP). The VLP transformer-based models are the basis for the current state-of-the-art models. The approach of vision-language pre-training is characterized by the use of large-scale automatically-annoteated datasets to pre-train the transformers prior to application in downstream vision-language tasks. It is interesting to note that the image feature extractor has been based on convolutional neural networks (CNN) for a long time, but recently there has been a shift towards Vision Transformers (ViT) for this function as well. This makes possible the realization of all-transformer models. The pipelines for each category shown in Figure 1 are paradigmatic and as such, most of the models we look at in this paper follow a similar flow. However, different models have incorporated various other features as well as changes in the fundamental architecture, which have improved the performance.

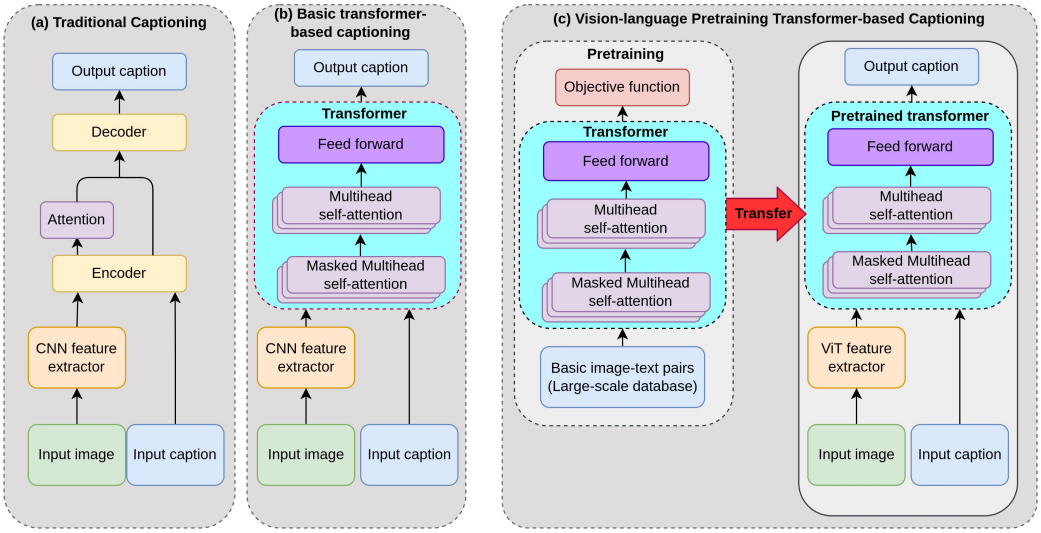

Figure 1. Transformers in image captioning: evolution from (**a**) traditional encoder-decoder-based captioning to (**b**) basic transformer-based captioning models to (**c**) captioning transformers based on vision-language pre-training.

In the rest of this section, we review the basic transformer-based captioning models and the models employing vision-language pre-training. We do not focus on the prior traditional captioning models (encoder-decoder-based), since those have been sufficiently reviewed in other works, such as [41,43].

4.1. Conceptual Captions

Sharma et al. [59] present a novel dataset of image caption annotations called Conceptual Captions. It contains an order of magnitude more images than the MS-COCO dataset [75], which has been a benchmark dataset for image captioning. The Conceptual Captions dataset is created using a pipeline that programmatically acquires images and captions from billions of internet web pages. In the modeling of the captioning system, they use a feature extractor based on Inception-ResNet-v2 [76]. A transformer is used as the decoder and they show that the transformer model achieves better results than a Recurrent Neural Network, which was previously the dominant mode for caption generation. Their encoder and decoder essentially adhere to the generic pipeline shown in Figure 1b. Whereas their main contribution is the Conceptual Captions dataset, they are among the first to employ the transformer model in image captioning.

4.2. Captioning Transformer with Stacked Attention Modules

Zhu et al. [77] apply a standard transformer to perform image captioning. The encoder is a ResNext CNN [78] and the image features are taken from the final layers. These are then used as the keys and values in the decoder. Thus, the encoder is a CNN and the

decoder is a transformer, with a design and parameters closely following those used by Vaswani et al. [36].

The decoder design envisions a stacking of several individual decoder layers; the overall output is taken as a combination of the outputs of the different layers. The decoder layers are separated by dropout layers to avoid overfitting.

Based on this stacking, they introduce multi-level supervision to take advantage of multi-layer outputs of the transformer. Every layer can be used to generate the current word. During training, the outputs of each layer are passed through a linear transformation and are then taken together to calculate the cross-entropy loss of the model as shown in Figure 2. The training objective is to minimize the cross-entropy loss.

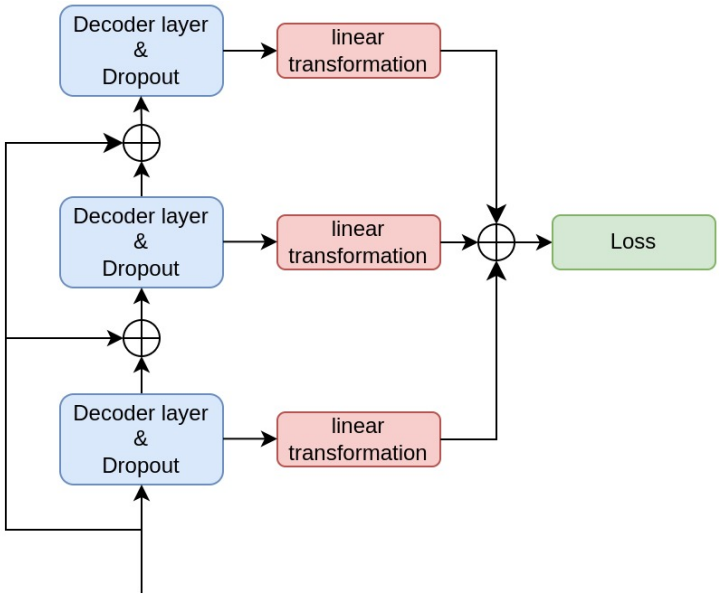

Figure 2. Architecture of the captioning transformer with stacked attention modules and multi-level supervision.

4.3. Image Captioning: Transforming Objects into Words

Herdade et al. [79] introduce the object relation transformer. This transformer employs the concept of geometric attention to incorporate information about the spatial relationships between various objects in an image. Geometric attention had earlier been used for object detection [80]. The approach of the authors involves utilizing the size ratio of the bounding boxes of different objects and the difference of the bounding box coordinates to deduce object relationship features. Their model is shown in Figure 3.

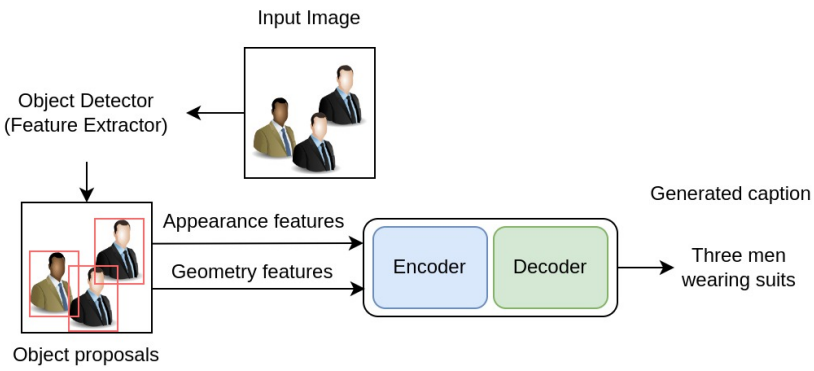

Figure 3. Object relation transformer. Appearance and geometry features are extracted from the image based on the object detector's regions.

The feature detector is a Faster R-CNN object detector [15], which is used to extract appearance and geometry features, similar to the approach of Hu et al. [80]. The transformer decoder is similar to that introduced by Vaswani et al. [36]. However, for each encoding layer, the attention scores are modified by multiplying by the geometric attention weights,

$$\mathbf{\Omega}_A = \frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}} \tag{1}$$

where $\boldsymbol{Q}$ and $\boldsymbol{K}$ are the queries and keys, respectively, and $\mathbf{\Omega}_A$ is an $N \times N$ attention weight matrix, whose elements ω_A^{mn} are the appearance-based attention weights between the m^{th} and n^{th} token. Relative geometry is incorporated by multiplying the appearance-based attention weights by a learned function of their relative position and size. The geometric attention weights, ω_G^{mn}, are first calculated based on the geometric features of the bounding boxes (center coordinates, widths and heights). The combined attention weights are then given by

$$\omega^{mn} = \frac{\omega_G^{mn}\exp(\omega_A^{mn})}{\sum_{l=1}^{N}\omega_G^{ml}\exp(\omega_A^{ml})} \tag{2}$$

where ω_A^{mn} and ω_G^{ml} are the elements of the attention and geographic attention weights, respectively.

Yang et al. [81], who followed a similar approach, used a scene graph representation for the encoding, which helps capture object relationships. Yao et al. [82] introduced a Graph Convolutional Network plus LSTM (GCN-LSTM) architecture that incorporates semantic and spatial object relationships into the image encoder. Yao et al. [83] develop a HIerarchy Parsing (HIP) architecture that parses images into multi-level structure consisting of the global level, regional level features and instance level features based on semantic segmentation. The hierarchical structure is fed into a Tree-LSTM to generate captions. However, none of these models ([81–83]) make use of transformers.

Herdade et al. [79] report better performance than Yang et al. [81] and Yao et al. [82] on the CIDEr-D metric. However, the Herdade et al. [79] model uses a transformer whereas [81] and [82] do not. The use of a transformer is a significant factor contributing to the better performance. It is worth noting that [83] outperforms [79] on the CIDEr metric, which highlights the effectiveness of their hierarchical parsing approach.

4.4. Attention on Attention

Huang et al. [84] introduce an attention on attention module (AoA) as illustrated in Figure 4. In the encoder they extract feature vectors of objects in the image and apply self attention. The AoA module is then applied to determine how the objects are related to each other. Essentially, self-attention is used to model the relationships among objects in the input image. In the decoder, the AoA module helps determine to what extent the attention results are related to the queries. The model first generates an information vector and an attention gate by employing the attention result and the context vector. A second attention module is then added through an element-wise multiplication of the attention gate and the information vector to yield the final attended information. Thus, the irrelevant or misleading results are filtered out to keep only the useful ones.

The attention on attention mechanism is formulated as the element-wise multiplication of an attention gate $\boldsymbol{g}$ and an information vector $\boldsymbol{i}$ as shown in Equation (3). The attention gate and information vector are results of two linear transformations, which are both conditioned on the context vector (the query) and the attention result. In the decoder, the AoA module is used, coupled to an LSTM module.

$$\mathrm{AoA}(f_{att}, \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \sigma(\boldsymbol{W}_q^g\boldsymbol{Q} + \boldsymbol{W}_v^g f_{att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + \boldsymbol{b}^g) \odot (\boldsymbol{W}_q^i\boldsymbol{Q} + \boldsymbol{W}_v^i f_{att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) + \boldsymbol{b}^i) \tag{3}$$

where f_{att} is an attention function that operates on the queries, keys and values, denoted by the matrices $\boldsymbol{Q}, \boldsymbol{K}$ and $\boldsymbol{V}$, respectively; $\boldsymbol{W}_q^i, \boldsymbol{W}_v^i, \boldsymbol{W}_q^g$ and $\boldsymbol{W}_v^g$ are learnable weight matrices;

$\boldsymbol{b}^g$ and $\boldsymbol{b}^i$ are bias vectors of the attention gate and information vector, respectively; σ is the sigmoid activation function and $\odot$ denotes element-wise multiplication.

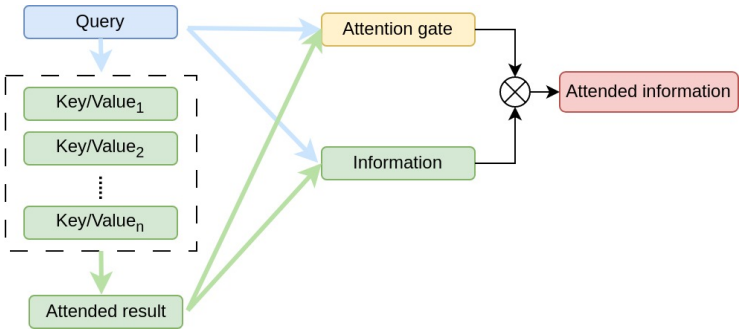

Figure 4. Attention on attention model.

4.5. Entangled Transformer

Li et al. [85] introduce the Entangled Attention (ETA) transformer that tries to exploit the semantic and visual information simultaneously and thus bridge the semantic gap. The semantic gap arises due to difficulties in attention mechanisms identifying accurately the equivalent visual signals, especially when predicting highly abstract words. With their entangled attention, semantic information is injected into the visual attention process; similarly, visual information is injected into the semantic attention process, hence the name "entangled".

They also presented the Gated Bilateral Controller (GBC) which is a gating mechanism that controls the path through which information flows. The GBC controls the interactions between the visual and semantic information. The overall attention model includes a visual sub-encoder and a semantic sub-encoder and a multimodal decoder, as shown in Figure 5. Each sub-encoder consists of N identical blocks, each with a multi-head self-attention and feed-forward layer.

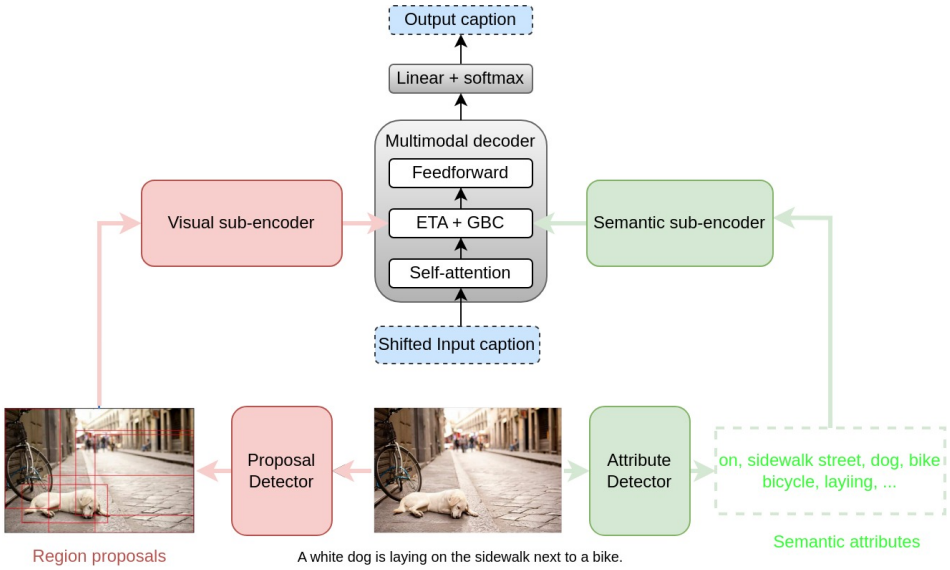

Figure 5. Overall architecture of the entangled transformer.

4.6. Meshed-Memory Transformer

The model of Cornia et al. [86] learns multi-level representations of the relationships between image regions such that low-level and high-level relations are represented. For this, *a priori* knowledge on relationships between image regions is encoded using persistent memory vectors, which results in memory-augmented attention. During the decoding, the

low-level and high-level relationships are used instead of employing a single visual mode representation. This is performed using a learned gating mechanism that weights the multi-level contributions at each stage. Figure 6 shows the general flow of their model which they refer to as a meshed-memory transformer due to the mesh connectivity between the encoder and decoder layers. Apart from using the MS COCO dataset [75], they validate the performance of their model on the novel object captioning using the Nocaps dataset [57].

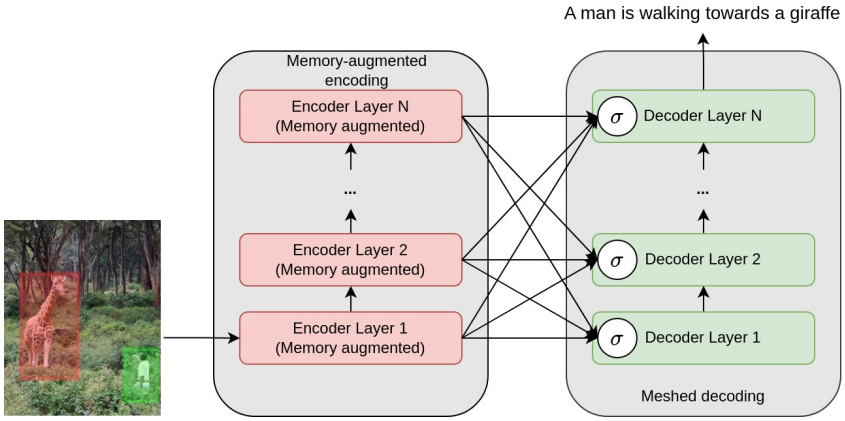

Figure 6. Architecture of the meshed-memory transformer. Multi-level encodings are connected to the decoder through a meshed and learnable connectivity. Each decoder layer inputs are controlled by a gating mechanism (σ) that weights the multilevel contributions at each stage.

4.7. X-Linear Attention Networks

Pan et al. [87] employ a method based on bilinear pooling [88]. Bilinear pooling entails getting the outer product of two input vectors. It is employed in a bid to more fully capture the relationships between two vectors. However, it is computationally expensive since the resulting number of parameters is $O(n^2)$. The approach can be made compact by dimensionality reduction. Bilinear pooling is an option to combine two vectors instead of other approaches such as concatenation, element-wise vector summing or element-wise vector multiplication.

A problem of conventional attention mechanisms that Pan et al. [87] try to overcome is the fact that the attention weights are essentially derived from the linear combination of the query and the key via an element-wise summation. This only exploits the first-order feature interactions between the textual domain and the visual domain. Whereas typical attention approaches are additive attention or dot-product attention, they propose a spatial and channel-wise attention mechanism based on bilinear pooling to exploit the higher order feature interactions. They package their attention mechanism into a block that they name X-linear attention block. The X-linear attention block uses a feature extraction backbone based on the Squeeze-and-Excitation Networks (SENet) of Hu et al. [89]. The final model is the X-Linear Attention Network (X-LAN) shown in Figure 7. It incorporates the X-linear attention block into the encoding and decoding operations.

Similar to AoANet, the authors confirm that improving attention measurement is an effective way of improving the interactions between the visual and textual domains. Exploiting rich semantic information in images (such as scene graphs and visual relations) leads to improved performance.

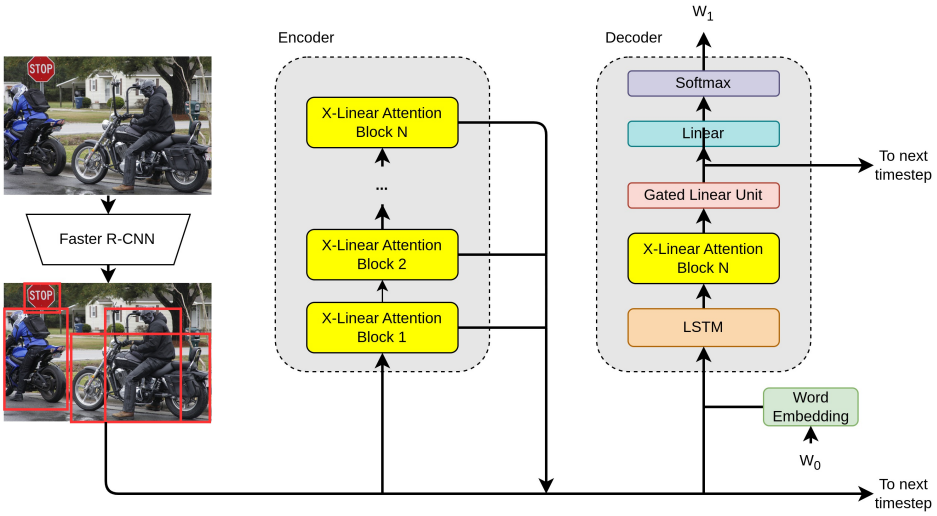

Figure 7. Simplified X-Linear Attention Network. The encoder and decoder utilize the X-Linear Attention Block. The decoder also relies on an LSTM. Decoder outputs at each time step therefore feed into the decoding operation of the next time step.

4.8. Image Transformer

He et al. [90] proposed an image transformer, whose core idea is to increase the width of the original transformer layer, designed for machine translation, and make it more suitable for the structure of images. In their image transformer, each layer has several sub-transformers that capture the spatial relationships between the image regions. The encoding method makes use of a visual semantic graph as well as a spatial graph. They use a transformer layer to combine them without external relationship or attribute detectors.

He et al. [90] distinguish between single-stage and two-stage attention-based methods. The single-stage methods are those in which attention is only applied at the decoding step with the decoder attending to the most relevant regions. Two-stage methods use bottom-up attention and top-down attention [35]. The bottom-up uses object detection based methods to select the most relevant regions; top-down attention then attends to those detected regions. Although the two-stage methods improve on the single-stage methods, they have the limitation that the detected regions are isolated and their relationships are not modeled. This limitation is tackled by scene graph based models. However, the scene graph models use auxiliary or external models to detect and build the scene graphs. He et al. [90] introduce a spatial graph encoding transformer layer, which considers the spatial relationships between the various detected regions in an image.

The model considers three categories of spatial relationships, namely parent, neighbor and child relationships between the various regions of an image. These categories are based on the amount of overlap between the regions. Neighbors of a query region are those regions with no overlap or with overlap below a set threshold; a parent region contains a query region, whereas a child region is contained by the query region. The spatial relationships between region pairs are captured using graph adjacency matrices. For any two regions, l and m, the graph adjacency matrices are defined as represented in Equation (4). $\mathbf{\Omega}_p$, $\mathbf{\Omega}_n$ and $\mathbf{\Omega}_c$ are the parent, neighbor and child node adjacency matrices, respectively, and ϵ is a given threshold.

$$Area_l = \frac{Area(l \cap m)}{Area(l)} \text{ and } Area_m = \frac{Area(l \cap m)}{Area(m)}$$

$$\mathbf{\Omega}_p[l,m] = \begin{cases} 1, \text{ if } Area_l \geqslant \epsilon \text{ and } Area_l > Area_m \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

$$\mathbf{\Omega}_c[l,m] = \mathbf{\Omega}_p[m,l]$$

$$\text{with} \sum_{i\in\{p,n,c\}} \mathbf{\Omega}_i[l,m] = 1$$

As shown in Figure 8, the original transformer layer is widened by adding three parallel sub-transformer layers, each being responsible for a subcategory of spatial relationships. The decoder incorporates an LSTM and its structure is also widened to correspond to the encoder; its output is obtained via a gated linear unit.

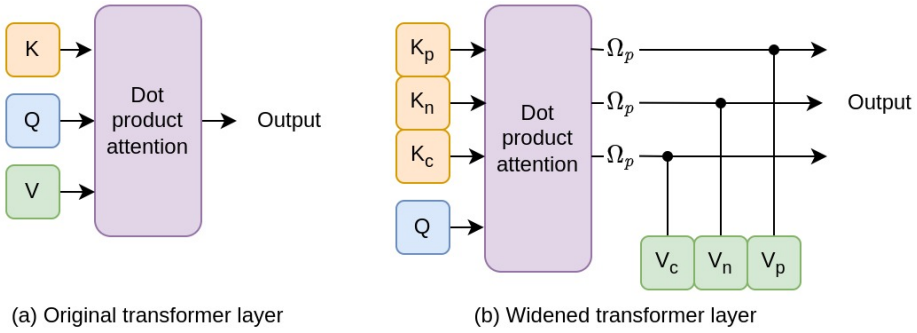

Figure 8. (**a**) The original transformer layer and (**b**) the widened encoding transformer layer of the image transformer. Each sub-transformer of the widened layer is responsible for a category of spatial relationship; all share the same query. $\mathbf{\Omega}_p$, $\mathbf{\Omega}_n$ and $\mathbf{\Omega}_c$ are the parent, neighbor and child node adjacency matrices.

Since they are incorporating graph information into the transformer, the model is similar to other graph-extraction techniques. However, they use a transformer, unlike other graph-extraction techniques [81,82]. The authors point out that their model is more computationally efficient since the other scene graph extracting models fuse semantic and spatial scene graphs, and require auxiliary models to first build the scene graph.

4.9. Comprehending and Ordering Semantics

Li et al. [91] develop a model, called COS-Net, that aims at comprehending the rich semantics in images and ordering them in linguistic order so as to generate visually grounded and linguistically coherent captions for the images. Their architecture entails four primary components: cross-modal retrieval, a semantic comprehender, a semantic ranker and a sentence decoder. The cross-modal retrieval serves to generate semantic cues. The retrieval uses CLIP [53] to search for all the relevant sentences related to images. The words of these sentences are then used as the semantic cues. The semantic comprehender filters out irrelevant semantic cues while at the same time inferring any missing and relevant semantic words grounded in the image. To carry out the filtering, the comprehender makes use of grid features derived from a visual encoder based on CLIP. The semantic ranker then determines a linguistic ordering for the semantic words. The output of the semantic ranker is used together with the visual tokens of images to auto-regressively generate the output captions. The implementation is conducted using the x-modaler codebase [92]. The dataset used for training and testing is MS COCO.

5. Vision-Language Pre-Training in Captioning

5.1. Pre-Training

Recent research has shown pre-training to be a powerful approach in improving the effectiveness of transformer models. Similarly to other aspects of transformers, the transformer pre-training was initiated in the domain of natural language processing and then also found application in the domain of vision. This pre-training of transformer

models has impacted natural language processing in a similar way to how pre-training convolutional networks impacted vision applications, e.g., the pre-training on the large ImageNet dataset for applications such as image classification.

Pre-training is generally carried out on a large corpus and then further fine-tuned on a smaller dataset which is closely related to the target task. The nature and characteristics of natural language processing have meant that the corpus for the pre-training need not be fully labeled, so that the pre-training can be carried out in a self-supervised manner. In self-supervised learning [93,94], pseudo labels are automatically generated from a dataset of unlabeled data. The pseudo labels are then used to train a deep learning network on a predefined pretext task. As shown in Figure 9, the resulting learned parameters of the network are then transferred to other downstream target tasks via fine-tuning.

Self-supervised learning overcomes the inherent difficulty of acquiring labeled data for supervised learning, which is often time-consuming and expensive. Self-supervised learning enables the use of large amounts of readily available, non-annotated data.

In this section we briefly look at pre-training of transformers in the language domain (Section 5.1.1) followed by the visual domain (Section 5.1.2). We then detail a number of transformer implementations that employ pre-training for the task of image captioning (Section 5.2).

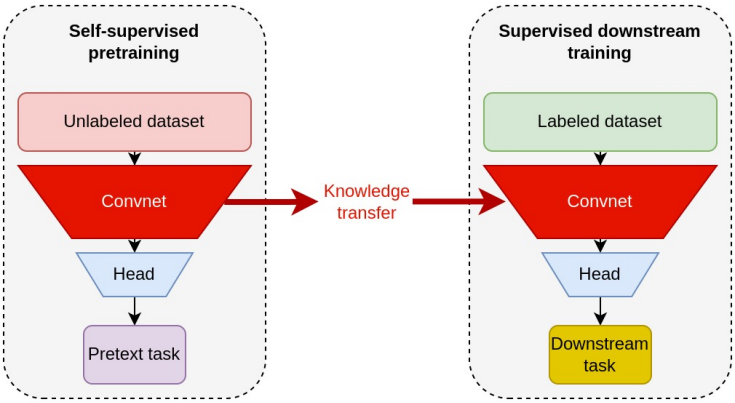

Figure 9. General pipeline for self-supervised learning.

5.1.1. Language Pre-Training

The pre-training for transformers was largely pioneered for language modeling tasks. The sections that follow point out some seminal models that have illustrated the potential inherent in pre-training. Devlin et al. [95] developed the BERT model (Bidirectional Encoder Representations from Transformers), which improved learned feature representations by encoding the left and right context of a word in a sentence, unlike previous methods which mainly attended to the context on the left of a given word in a sentence.

To enable the bidirectional architecture, they introduced a masked language model (MLM) training objective, inspired by the Cloze task [96]. In the masked model, one or more of the tokens in a sentence are masked and the task is to predict the masked tokens based on the context (the other tokens). In doing so, the model learns to incorporate the bidirectional context. Previous works had used unidirectional language models for pre-training. They also introduced the next sentence prediction (NSP) objective. In NSP, the model is given a pair of sentences and it learns to predict whether or not the second sentence follows the first. The labels for both MLM and NSP are generated automatically. The authors successfully fine-tuned their model for several downstream tasks such as question answering and language inference.

5.1.2. Vision-Language Pre-Training (VLP)

Following the success of BERT, a number of other closely related models were developed with the same idea of pre-training in mind. Vision-language pre-training brings

together the vision and language domains and shows that pre-training in the language domain can also improve performance from downstream tasks from the visual domain.

This approach was illustrated by Lu et al. [97], who developed a model they referred to as Vision-and-language BERT (ViLBERT). ViLBERT learns joint representations from the visual and language domains. It extends BERT to a multi-modal domain. The authors show that visual grounding is a pre-trainable and transferable task. They focus on the downstream tasks of visual question answering and visual commonsense reasoning, i.e., understanding-based tasks. LXMERT [98] is another cross-modality framework that learns connections between vision and language through vision-language pre-training.

5.1.3. Self-Supervised Pre-Training

A powerful approach that has emerged is that of pre-training models on vast amounts of image-text pairs collected from the internet. This is motivated by the fact that there are vast quantities of unstructured natural language data on the internet. This approach enables models to learn about images directly from raw text. The datasets used in these approaches are also interesting in that they do not need to be heavily curated and annotated using expensive and expert human agents. All this makes training at an unprecedented scale possible. After pre-training, natural language is used to refer to learned concepts from the visual domain are to describe new concepts not previously seen. Among other things, this facilitates zero-shot or few-shot transfer of the model to other downstream tasks. Representative models that employ this approach are CLIP [53] and ALIGN [54].

CLIP refers to Contrastive Language-Image Pre-training (Radfored et al. [53]). CLIP learns perception from the supervision based on natural language paired with images. They developed a dataset (WebImageText) consisting of 400 million image-text pairs taken from the internet. They experiment with two image encoders: one based on ResNet50 [12] and another based on the Vision Transformer (ViT) [99]. The text encoder is based on a transformer. They use a contrastive objective based on the cosine similarity function to learn a multimodal embedding space and to predict correct and incorrect text-image pairings. The optimization is via a cross entropy loss. The pre-trained model can then be transferred to various down-stream vision tasks, including in zero-shot settings, where it matches the performance of strong, fully supervised baseline models. For instance, in image classification on ImageNet, the CLIP model matches the original ResNet50 without using any of the crowd-labeled training examples that ResNet50 used.

ALIGN refers to A Large-scale ImaGe and Noisy-text embedding [54]. It is conceptually similar to CLIP but differs in that for pre-training, it uses noisy data taken from a large number of image alt-text pairs using an approach similar to that of Conceptual Captions [59], but without the expensive filtering or post-processing steps. The resulting corpus consists of 1.8 billion image-text pairs. The authors show that the scale of their resulting dataset makes up for the inherent noise and leads to good performance. During pre-training, they use a dual-encoder architecture to align the visual and language representations in a joint embedding space. The dual-encoder consists of an EfficientNet-based [100] image encoder and a BERT-based text encoder coupled with a cosine-similarity function. For their objective, they use a contrastive loss, similar to CLIP. The aligned vision-language representations enable zero-shot transfer learning and cross-modal retrieval (image-to-text and text-to-image).

5.2. VLP in Image Captioning

This subsection focuses on a number of models that employ vision-language pre-training and include image captioning as one of the downstream tasks.

5.2.1. Unified Vision-Language Pre-Training for Image Captioning and VQA

Zhou et al. [101] present a unified VLP model that can be fine-tuned for the downstream tasks of image captioning and visual question answering (vision-language generation and understanding). Unlike previous models, theirs uses a shared multi-layer

transformer network for encoding and decoding, for which reason they refer to it as *unified* (Figure 10). They point out that their approach is the first to present a single, unified model that is universally applicable to multiple downstream tasks. Their shared multi-layer transformer network is pre-trained on large amounts of image-caption pairs and is optimized for bidirectional and seq2seq masked language prediction.

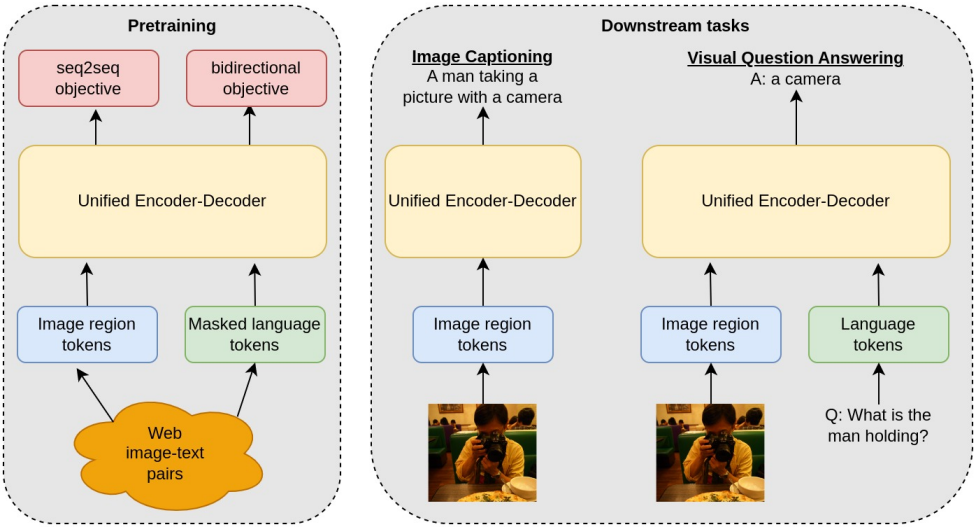

Figure 10. Unified VLP overview. The model is unified because the same multi-layer transformer network does the encoding and decoding; the same model is also fine-tuned for different downstream tasks, i.e., vision-language generation (image captioning) and understanding (VQA).

5.2.2. OSCAR: Object-Semantics Aligned Pre-Training for Vision-Language Tasks

Li et al. [38] introduce OSCAR, a VLP method to learn generic image-text representations for vision-language understanding and generation tasks. They employ anchor points (object tags detected in images) to ease the learning of semantic alignments between image regions and texts. The motivation is the observation that the salient objects in an image are often mentioned in the paired texts. These salient objects can accurately be detected using object detectors. The input of their method consists in the triplet of words, tags and image regions. The model is pre-trained with two losses: a masked token loss over words and tags, and a contrastive loss between tags and other polluted tokens. The model overview is illustrated in Figure 11. The model is pre-trained on a large dataset consisting of 6.5 m image-text pairs and then fine-tuned for five understanding and two generation downstream tasks, including visual question answering and image captioning.

The authors point out that existing VLP methods take visual region features and word embeddings of the paired text as input and then rely on self-attention to learn the image-text alignments and produce cross-modal contextual representations. A limitation with this is that it results in ambiguity owing to the oversampled and overlapping regions from the object detector. They also mention the lack of grounding due to VLP being a weakly-supervised learning problem: there are no explicitly labeled alignments between image regions and words or phrases in the text. Their method of using object tags as anchor points seeks to overcome these two limitations.

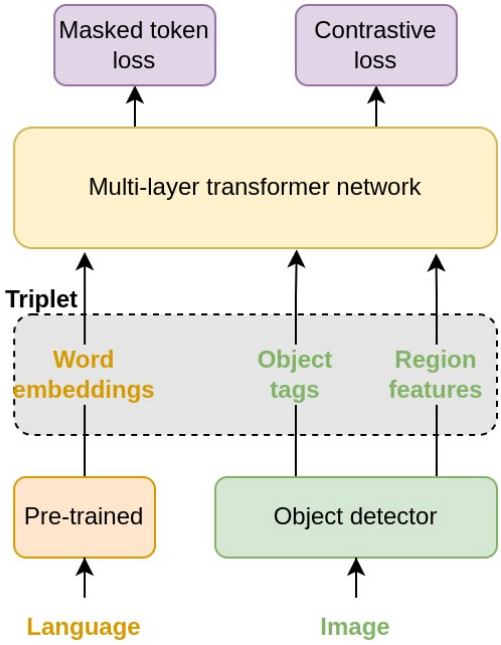

Figure 11. OSCAR overview. The object tags are used as anchor points to align image regions with word embeddings. The objective function is based on the masked token loss and a contrastive loss.

5.2.3. VinVL: Making Visual Representations Matter in Vision-Language Models

Zhang et al. [39] develop an object detection model that provides object-centric representations of images. Their model is bigger than the bottom-up top-down model [35] and is pre-trained on a larger corpus. Their aim is to generate representations of a richer collection of visual objects and concepts. They show that visual features matter a lot in VL models. They use the visual features from their model together with the OSCAR model [38] to yield improved performance in a number of downstream vision understanding and generation tasks.

They pre-train a large-scale object-attribute detection model based on the ResNeXt-152 C4 architecture [78]. Since the model is bigger, better designed for VL and trained on more and larger corpora, it yields richer semantics shown in the richer visual concepts and attribute information. The object detection model leverages feature pyramid networks [102]. The object detection model is pre-trained on a large-scale corpus and then fine-tuned with an additional attribute branch on Visual Genome. Thus, it can detect both objects and attributes.

5.2.4. VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning

Hu et al. [40] present a method of pre-training on image-tag pairs rather than image-caption pairs to build a visual vocabulary. The model is then later fine-tuned using image-caption pairs and then tested on a dataset with objects not seen during the image-caption training or fine-tuning. The visual vocabulary enables the model to reasonably generate captions on the objects not seen in the fine-tuning dataset. The visual vocabulary is a joint embedding space where image region features and tags of semantically similar objects are mapped into vectors that are close to each other. The cosine similarity is used to measure the distance between image region and tag.

The perceived benefit of this is the possibility of making use of many other datasets that only contain images and tags to pre-train for the downstream task of image captioning, given that image-caption datasets may be smaller or fewer and contain less diverse visual objects. Images paired with machine-generated tags can also be used as weak supervision signals. Since the tags for training are not ordered, they employ a Hungarian matching loss

with masked tag prediction to conduct the pre-training. The training objective is to predict the masked tags given a bag of image-level tags and image regions.

5.2.5. Scaling Up Vision-Language Pre-Training for Image Captioning

In this work, Hu et al. [52] study the scaling behavior of vision-language pre-training for image captioning. They develop a large-scale image captioner, called LEMON, based on the transformer architecture and scale the size from 13 million parameters all the way to 675 million parameters. To carry out the training, they develop ALT200M, a large-scale dataset that contains up to 200 million image-text pairs and that is based on the *alt* attribute of images from the internet. Pre-training is carried out on ALT200M followed by fine-tuning on MS COCO. They use subsets of ALT200M of various sizes to study the effects of scaling the dataset. The reference model they use is VinVL [39]; as such, they do not develop any new model. They evaluate the model using Nocaps [57], which assesses the effectiveness of the approach in regard to out-of-domain objects. They also test their model on the MS COCO Karpathy test split [27] as well as Conceptual Captions [59].

Their findings show that the performance of captioning models generally improve with an increase in the model size as well as the dataset size. More specifically, they show that larger models benefit more when the dataset size is beyond a certain threshold. In their experiments, they find that with only 3 million image-text pairs for pre-training, the performance plateaus early even as the model size increases. This plateauing effect is not observed (for the model sizes they look at) when there are more than 10 million image-text pairs for pre-training, i.e., with sufficient data, the performance keeps improving as the model size increases. This implies that the model capacity is the performance bottleneck, which suggests that training even larger models could push farther the limits of VLP for captioning tasks.

5.2.6. SIMVLM: Simple Visual Language Model Pre-Training with Weak Supervision

Wang et al. [103] present the Simple Visual Language Model (SimVLM), which uses large-scale weak supervision to simplify the vision-language pre-training process by reducing the reliance on expensive annotated and labeled datasets as well as object detection pre-training. Their model is also simpler in that it does not not require the use of multiple auxiliary loss functions (objectives).

They propose the Prefix Language Modeling (PrefixLM) objective. With PrefixLM, some initial tokens of an input sequence are designated as prefix tokens, which are generally taken from the image represenation. PrefixLM enables bi-directional attention on the prefix tokens whereas the remainder of the tokens in the input sequence, generally taken from textual representation, are processed via the autoregressive factorization (causal attention). The PrefixLM training objective is given by:

$$\mathcal{L}_{PrefixLM}(\theta) = -\mathbb{E}_{\mathbf{x}\sim D}[\log P_\theta(\mathbf{x}_{\geq T_p}|\mathbf{x}_{<T_p})] = -\mathbb{E}_{\mathbf{x}\sim D}\left[\sum_{t=T_p}^{T} \log P_\theta(\mathbf{x}_t|\mathbf{x}_{[T_p,t]}, \mathbf{x}_{<T_p})\right] \quad (5)$$

where θ are the trainable model parameters, D is the dataset and $\mathbf{x}$ is the input sequence. $\mathbf{x}_{<T_p}$ are the prefix tokens and $\mathbf{x}_{[T_p,t]}$ are the tokens for the autoregressive factorization. Therefore, PrefixLM takes advantage of bi-directional contextualized representation as well as autoregressive generation. It ends up being modality-agnostic since it can effectively deal with the visual and/or the textual domains.

The approach allows images to be considered as prefixes for their corresponding textual descriptions. The image section of the input sequence is acquired via a transformer backbone inspired by the Vision Transformer (ViT) [99] and CoAtNet [104]. This approach eliminates the need for an object detection module since they operate on raw image patches. The textual section of the input sequence is acquired via tokenization and embedding. The positional information of the tokens is also incorporated. The bi-directional attention and causal attention are then carried out on the image and textual tokens, respectively.

For training, they use the image and alt-text pairs taken from ALIGN [54] together with the Colossal Clean Crawled Corpus (C4) [105]. The modality-agnostic nature of their objective allows the model to train on these two datasets made up of image-text pairs and text-only data. Training on the text-only corpora effectively compensates for the inherently noisy nature of the alt-text data. They evaluate their model after fine-tuning on six downstream vision-language tasks, including visual question answering (VQA) and image captioning (using MS COCO and Nocaps). They also show strong generalization and transfer ability in zero-shot settings.

5.2.7. "One For All" (OFA) Framework

Wang et al. [106] present a "One For All" framework that aims to be task-agnostic and modality-agnostic. This work also pursues the objective of task comprehensiveness, in which a single model can generalize such that it performs well on a large variety of tasks, including vision-language, vision only and language only tasks. They also demonstrate their model's performance in zero-shot learning and task transfer.

Their approach unifies a variety of multimodal and unimodal vision and language tasks in a sequence-to-sequence learning framework. Their model has the advantage that it yields state-of-the-art performance despite pre-training on a significantly smaller dataset than other models pre-trained on large-scale datasets; they pre-train on a dataset of 20 million image-text pairs. For comparison, the pre-training dataset sizes for CLIP [53] and ALIGN [54] are 400 million and 1.8 billion, respectively.

The architecture is transformer-based and consists of an encoder-decoder model without task-specific layers added. The image encoder is similar to that of SimVLM [103] and is based on CoAtNet [104]; it extracts fixed-size patches from images. For the textual encoding, a given text sequence is transformed into a subword sequence which is then embedded into features. The inputs from the different modalities are represented in a unified embedding space. To do this, they discretize data from the visual and textual domains using vector quantization and byte-pair encoding, respectively, and then represent them using a unified vocabulary. The downstream tasks and modalities of interest are then represented in a sequence-to-sequence setting for training. They rely on handcrafted instructions (instruction-guided pre-training) to differentiate between the different tasks. To optimize the model, they use a cross-entropy loss that takes into account the input and an instruction. The training objective (loss function) to be minimized is given by:

$$\mathcal{L} = -\sum_{i=1}^{|y|} \log P_\theta(y_i | y_{<i}, x, s|) \tag{6}$$

where θ refers to the model parameters, x is the input, s is the handcrafted instruction and y is the output. This model achieved state-of-the-art performance and in the case of the downstream task of captioning, it topped the publicly available MS COCO captioning leaderboard [69] (as of 31 May 2022).

5.2.8. Florence: A New Foundation Model

Florence [107] aims at fostering development of foundation models [108], which are architectures that learn joint representations and generalize well to a wide range of downstream tasks with limited additional domain knowledge (zero-shot transfer or minimal task adaptation). Differing from models such as CLIP [53] and ALIGN [54] which focus on cross-modal joint representations for classification and retrieval, the Florence model broadens the representations to include object level, multiple modality and videos. Florence defines a computer vision foundation model that serves as a general purpose vision system for various vision tasks that can be mapped onto space-time-modality coordinates.

The different operations of Florence entail data curation, model pre-training, task adaptations and training infrastructure. Similarly to ALIGN, Florence is trained on large-scale, noisy web-data (900 million image-text pairs) with a unifying image-text contrastive

learning objective based on UniCL [109]. For pre-training, a transformer-based encoder-decoder model is used. For tasks adaptation, various adapters are employed which make it extensible and transferable.

Although the authors do not report results for image captioning, we look at their model since being a foundation model, it is applicable to image captioning as well.

5.2.9. Contrastive Captioners Are Image-Text Foundation Models

This work proposes a Contrastive Captioner (CoCa) [110] to pre-train a foundation model [108] jointly with a contrastive loss and a captioning loss. They employ a transformer-based encoder-decoder architecture. A distinguishing feature of their decoder is that the first half of the layers only attend to the unimodal textual representations, omitting the outputs of the image encoder. The second half of the decoder cross-attends to the encoder outputs as well to produce the multimodal image-text representations. The contrastive loss is applied to the image encoder and the unimodal section of the decoder whereas the captioning loss is applied to the multimodal section of the decoder, which enables it to generate textual tokens in an autoregressive manner. Their overall loss function is given by:

$$\mathcal{L}_{CoCa} = \lambda_{Con} \cdot \mathcal{L}_{Con} + \lambda_{Cap} \cdot \mathcal{L}_{Cap}, \tag{7}$$

where $\mathcal{L}_{Con}$ and $\mathcal{L}_{Cap}$ are loss weighting hyper-parameters. The pre-training is performed using alt-text data as well as annotated images; all labels are simply treated as text. CoCa unifies into a single model and single pre-training stage the three training paradigms of single-encoder (such as in image classification), dual-encoder (such as in contrastive learning for image-text alignment) and encoder-decoder (such as in image captioning and multimodal representation). CoCa gives state-of-the-art results on tasks pertaining to the categories of visual recognition, cross-modal alignment, image captioning and multimodal understanding. As an indication of the model performance in image captioning, they do not use CIDEr-optimization and yet their model performs competitively against other models that are CIDEr-optimized.

5.2.10. GIT for Vision and Language

The Generative Image-to-text Transformer (GIT) [111] focuses on simplicity. The architecture entails just one multi-modal encoder and one text decoder. The image encoder employed is a vision transformer [99], which therefore eliminates the need for an object detector. The encoder is pre-trained using a contrastive objective on a large-scale dataset consisting of 0.8 billion image-text pairs. The text decoder is a standard transformer; it is trained using the language modeling objective, which aligns the input image and the corresponding textual caption. It also generates text in an autoregressive manner. Their approach separates the pre-training based on the contrastive language modeling objectives, contrary to the approach used in CoCa [110], which unifies the training using these objectives. A key feature of their model is that they use a self-attention mechanism for the decoder that attends to both the image and text representations, i.e., they are concatenated. This is as opposed to using self attention separately on the text tokens and then incorporating the visual information via cross-attention in the transformer stack. They show that in a large-scale pre-training setting, their approach achieves superior performance. However, in a smaller-scale setting, the cross-attention approach outperforms the pure self-attention approach.

GIT is simple but achieves state-of-the-art performance. This could largely be attributed to the scaling up of the training data (0.8 B image-text pairs) and the model size (0.7 B parameters). They evaluate their model on challenging benchmarks in image and video captioning, visual question answering, image classification and scent text recognition. As of 22 August 2022, a variant of GIT topped the publicly available MS COCO leaderboard [69].

5.2.11. Universal Captioner

Cornia et al. [112] propose a model that can generate in- and out-of-domain captions characterized by the natural descriptive style and fluency of human captions. Their model trains well on datasets with different descriptive styles and semantic concepts, i.e., non-uniform or heterogeneous sources, and manages to separate the content from the descriptive style. The non-uniform sources are made up of curated datasets as well as weakly labeled or noisy web-scale datasets. A chief goal of the work is to foster the ability to describe more real-world concepts with a high level of caption quality.

The architecture entails a transformer-based encoder-decoder. The encoding of the images is based on a self-attentive visual encoder [99] and is, therefore, carried out without the need for an object detector. They use a large-scale multi-modal model based on CLIP [53] to extract *pivotal keywords*, which then help to train in a style-aware manner that enables the transfer of semantic concepts between sources. The pivotal keywords are similar in concept to the object tags used in OSCAR [38]; however, they take the style into account. The decoder then jointly takes in the CLIP-based keywords, style and other text to generate the caption in an autoregressive manner. In terms of the objective loss function for training, they only use unidirectional language modeling coupled with a prompting strategy, which highlights the simplicity of their approach. During inference, predictions are conditioned on a dataset indicator parameter, which is chosen according to the desired generation style. Their training is on a dataset of 35.7 million images.

5.2.12. PaLI: A Jointly-Scaled Multilingual Language-Image Model

The Pathways Language and Image (PaLI) model [55] combines inputs from the vision and language domains and generates outputs in the language domain. A key feature of this model is that is uses a very large-scale encoder-decoder language models and an enormous Vision Transformer (ViT-e) image model; the models in these two domains have been pre-trained separately. The reuse of the large unimodal backbones for language and vision modeling enables a significant reduction in training cost.

The resulting model is pre-trained on a very large-scale dataset, the Web Language Image (WebLI) dataset [55], which has the distinction of being very large-scale and multilingual. WebLI encompasses over a hundred languages and contains more than 10 B image-text pairs, which surpasses all the other datasets commonly used for vision-language pre-training. With this approach, the authors try to achieve a jointly scaled model and overcome the limitation which often results from the fact that language models tend to be much larger and trained on larger datasets compared to the vision models. The model performs well on several downstream language, vision and vision-language tasks, including image captioning and visual question answering.

The architecture of PaLI is simple and scalable. It consists of an enormous vision transformer that encodes an image into tokens which are then fed, together with encoded text tokens, into a large scale transformer-based encoder-decoder to produce the textual output. Tasks are therefore framed using an "image + query to answer" modeling interface. The high performance of the model comes in large part from the sheer size of the model as well as of the dataset on which it is trained; although the WebLI dataset has 10 B image-text pairs, they train their model on a subset of 1 B examples (so as to only use high-quality examples).

5.2.13. BLIP-2

Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (BLIP-2) [113] is a model that consists of a lightweight querying transformer (Q-former) that acts as a bridge between two off-the-shelf large-scale unimodal transformers that have been pre-trained and then frozen. Therefore, the training of the model is only conducted on the bridge resulting in a generic and compute-efficient vision-language pre-training method. The first large-scale pre-trained transformer is an image encoder which extracts features from the visual domain, which then acts as an input to the

Q-former. The Q-former output is fed to the second large-scale pre-trained transformer, which is a large language model (LLM). BLIP-2 achieves state-of-the-art performance on several vision-language tasks including image captioning, visual question answering and image-text retrieval.

6. Discussion and Analysis

In this Section we give an analysis and comparison of the various transformer-based image captioning models that have been assessed in Sections 4 and 5. We begin by looking at the major contributions of the various models, together with the datasets employed and some of the basic design choices. We then look at some quantitative aspects by comparing the performance of the different models based on the standard evaluation metrics.

6.1. Major Contributions

Table 3 gives a summary of the major contributions of the various transformer-based captioning methods that do not use vision-language pre-training. The Conceptual Captions transformer [59] is one of the earlier models that employed a transformer for captioning, shortly after the transformer was pioneered [36]. It had an associated dataset released (dubbed Conceptual Captions) which contains an order of magnitude more images thant the MS-COCO dataset.

Table 4 shows the major contributions of the methods that employ vision-language pre-training. It is interesting to note that the inherent design of the VLP transformers facilitated their application to multiple downstream vision-language tasks, resulting in very versatile models. VLP generally requires large amounts of data so a number of large-scale datasets associated with some of the models were created.

Table 3. Major contributions of the non-VLP transformer-based models.

Model	Major Contributions
Conceptual Captions transformer [59]	A novel dataset (Conceptual Captions) which contains an order of magnitude more images than the MS-COCO dataset.
Stacked attention transformer [77]	Multi-level supervision that utilizes the outputs of multiple transformer layers to generate output tokens and calculate the loss.
Object relation transformer [79]	Object relation transformer (using geometric attention) that incorporates the spatial relationships between various objects.
Attention on Attention [84]	Attention on Attention module, which doubly applies attention to determine object relationships and filter out irrelevant attention results.
Entangled Transformer [85]	Entangled attention, which exploits semantic and visual information; Gated Bilateral Controller to control visual and semantic interactions.
Meshed-Memory Transformer [86]	Mesh-like multi-level structure to capture high-level and low-level image region relations.
X-Linear Attention Networks [87]	Unified X-Linear attention block that models the second-order interactions with both spatial and channel-wise bilinear attention.
Image Transformer [90]	A modified attention module suited to the complex natural structure of image regions.
COS-Net [91]	A semantic comprehender that grasps and orders the rich semantics in images to generate linguistically coherent captions. They use the x-modaler codebase.

Table 4. Major contributions of the VLP transformer-based models.

Model	Major Contributions
OSCAR [38]	The use of object tags as anchor points for cross-modal representation learning; a novel pre-training objective.
Unified VLP [109]	A single unified multilayer transformer network for encoding and decoding that is universally applicable to multiple downstream tasks.
VinVL [39]	An enhanced object detection model; they demonstrate that visual features matter significantly in VL models.
VIVO [40]	Creation of visual vocabulary by pre-training on image-tag pairs rather than image-caption pairs.
LEMON [52]	An empirical study of the effects of scaling up VLP; ALT200M, a large-scale pre-training dataset based on alt attribute of internet images.
SIMVLM [103]	A single Prefix Language Modeling objective which enables bi-directional contextualized representation and autoregressive generation.
CLIP [53]	An efficient and scalable way to learn image representations using a large dataset of image-text pairs collected from the internet.
ALIGN [54]	A large-scale dataset of 1.8 B image-text pairs from the internet without expensive processing; useful for image representations via pre-training.
OFA [106]	A "one-for-all" seq-to-seq framework that is task-agnostic, modality-agnostic and task comprehensive.
CoCa [110]	A Contrastive Captioner that pre-trains a foundation model jointly with contrastive and captioning losses combined into a single objective.
GIT [111]	A simplified architecture of an image encoder and a text decoder under one language modeling task. A new scheme of generation-based image classification.
Universal Captioner [112]	A model capable of generating in/out-of-domain captions characterized by the natural descriptive style and fluency of human captions.
PaLI [55]	Joint scaling of vision and language components resulting in a model that performs vision-language tasks in many languages; WebLI, a large-scale database with 10 B image-text pairs in more than 100 languages.
BLIP-2 [113]	A lightweight querying transformer that acts as a bridge between two off-the-shelf large-scale unimodal transformers pre-trained and then frozen; results in compute-efficient vision-language pre-training.

6.2. Datasets and Metrics

Table 5 shows the datasets and the metrics used by the non-VLP and VLP models. With regard to the datasets, it is seen that the MS COCO dataset has been used by all the models. Results reported based on the MS COCO dataset therefore offer a good basis of comparison of different models. The Visual Genome dataset is largely used for pre-training of the feature extractor, especially in cases where the attributes and object relationships are required to provide richer features.

The methods that use VLP usually require much larger datasets for the pre-training. Therefore, the some of the VLP methods resulted in the creation of large-scale datasets. Illustrative examples are the ALT200M [52], WebImageText those used by CLIP [53] and ALIGN [54] datasets. CoCa uses the JFT-3B dataset [114]. CC3M and CC12M are variants of Conceptual Captions; WIT is the Wikipedia-based image text dataset. Together with large datasets, the VLP methods entail increasingly larger model sizes.

The metrics used have been standard, which facilitates the comparison of the various models. The metrics are useful for offline as well as online evaluation on the official MS COCO test server. The COS-Net model [91] makes use of the CHAIR metric [115], which is used to measure object hallucination. However, since this is not used by the other models, it cannot be used here for comparison purposes.

Table 5. Datasets and metrics used by VLP and non-VLP models. CC is Conceptual Captions. B refers to BLEU-4, S to SPICE, M to METEOR, R to ROUGE-L, C to CIDEr and CH to CHAIR [115]. VG refers to Visual Genome and OI to OpenImages. Section 2 has more details on the datasets and metrics themselves.

Model	VLP	Datasets	Metrics
Conceptual Captions transformer [59]	✗	MS COCO, CC	B, M, R, C
Stacked attention transformer [77]	✗	MS COCO	B, R, C
Object relation transformer [79]	✗	MS COCO	B, S, M, R, C
Attention on Attention [84]	✗	MS COCO; ImageNet, Visual Genome	B, S, M, R, C
Entangled Transformer [85]	✗	MS COCO	B, S, M, R, C
Meshed-Memory Transformer [86]	✗	MS COCO, Nocaps, Visual Genome	B, S, M, R, C
X-Linear Attention Networks [87]	✗	MS COCO, ImageNet, Visual Genome	B, S, M, R, C
Image Transformer [90]	✗	MS COCO, ImageNet, Visual Genome	B, S, M, R, C
COS-Net [91]	✗	MS COCO	B, S, M, R, C, CH
OSCAR [38]	✓	MS COCO, CC, SBU, Flickr30k, VQA, GQA, Nocaps	B, S, M, C
Unified VLP [109]	✓	MS COCO, CC, Flickr30k, VQA v2.0; VG	B, S, M, C
VinVL [39]	✓	MS COCO, OI v5, Objects365, VG; CC, SBU, Flickr30k; VQA, GQA	B, S, M, C
VIVO [40]	✓	MS COCO, Nocaps; OI	B, S, M, C
LEMON [52]	✓	MS COCO, Nocaps, CC, ALT200M	B, S, M, C
SIMVLM [103]	✓	ALIGN, C4, MS COCO, Nocaps	B, S, M, C
OFA [106]	✓	MS COCO, CC, SBU, VG	B, S, M, C
CoCa [110]	✓	MS COCO, Nocaps, JFT-3B, ALIGN	B, S, M, C
GIT [111]	✓	MS COCO, CC, SBU, VG, ALT200M	B, S, M, C
Universal Captioner [112]	✓	MS COCO, Flicker30k, SBU, CC, WIT, YFCC100M, OI	B, S, M, R, C
PaLI [55]	✓	WebLI, MS COCO, Nocaps, TextCaps, VQA	C
BLIP-2 [113]	✓	MS COCO, VG, CC3M, CC12M, SBU	B, C

6.3. Training and Model Parameters

Table 6 shows the sizes of the non-VLP models. XE is the cross-entropy loss and SCST is the Self-Critical Sequence Training; N are the number of stacked layers, d the dimensionality of the layers and h the number of heads. Many of the models employ a transformer with a size similar to the original transformer used by Vaswani et al. [36], i.e., with the number of sub-layers, N = 6, the embedding dimensionality, d = 512 and the number of heads, h = 8. The models in Table 6 predominantly use Faster-RCNN [15] as the feature extractor. However, COS-Net [91] deviates from this pattern and uses a more recent feature extractor based on CLIP [53].

Table 6. Training and model details for non-VLP models. XE: cross-entropy loss, SCST: self-critical sequence training (CIDEr optimization).

Model	Feature Extractor	Training Objective	Optimizer	Transformer Details		
				N	d	h
Conceptual Captions [59]	Inception-ResNet-v2	XE	Adagrad	6	512	8
Stacked attention transformer [77]	ResNext	XE	Adam	6	512	8
Object relation transformer [79]	Faster RCNN	XE, SCST	Adam	6	512	8
Attention on Attention [84]	Faster RCNN	XE, SCST	Adam	6	1024	8
Entangled Transformer [85]	Faster RCNN	XE, SCST	Adam	6	512	8
Meshed-Memory Transformer [86]	Faster RCNN	XE, SCST	Adam	6	512	8
X-Linear Attention Networks [87]	Faster RCNN	XE, SCST	Adam	4	512	-
Image Transformer [90]	Faster RCNN	XE, SCST	Adam	3	1024	8
COS-Net [91]	CLIP-grid features	XE, SCST	Adam	6	512	-

For training and optimization, most of the models used cross-entropy optimization. This method is based on the use of the cross-entropy loss during optimization. A recent approach also employed by most of the models and which has become a standard optimization for image captioning is the Self-Critical Sequence Training (SCST) optimization, pioneered by Rennie et al. [116]. SCST optimization makes use of reinforcement learning to specifically optimize for the CIDEr evaluation metric, which is a fundamental metric used in the comparison of different models and for reporting results on the MS COCO captioning

leaderboard. SCST is used for fine-tuning after initial training using cross-entropy optimization. For the weight updates performed during back propagation, the approach that was predominantly used was the Adam optimizer [117]. However, Conceptual Captions [59] uses Adagrad.

Table 7 shows the details of the models that employ VLP. Compared to the non-VLP counterparts, these models tend to be much larger, especially when considering the number of layers and heads. The VLP models usually have several sizes for experimentation, which are typically labelled 'Base', 'Large' and 'Huge'. In Table 7, the best-performing size of each model is the one that has been taken into account.

Table 7. Training and model details for VLP models. MTL-b: bidirectional Masked Token Loss, MTL-s: seq2seq Masked Token Loss, XE: Cross-Entropy loss, CL: Contrastive Loss, SCST: Self-Critical Sequence Training, PxLM: Prefix Language Modeling.

Model	Feature Extractor	Objective (Loss)		Optimizer	Transformer Details		
		Pre-training	Fine-tuning		N	d	h
OSCAR [38]	Faster RCNN	MTL-b, CL2	MTL-s, SCST	AdamW	24	1024	16
Unified VLP [109]	Faster RCNN	MTL-bs	MTL-s, SCST	Adam	12	768	12
VinVL [39]	ResNeXt-152 C4	MTL-b, CL3	MTL-s, SCST	AdamW	24	1024	16
VIVO [40]	Faster RCNN	MTL-b, Hungarian	MTL-s, SCST	Adam	24	1024	16
LEMON [52]	Faster RCNN	MTL-s	MTL-s, SCST	AdamW	32	1280	16
SimVLM [103]	ViT; CoAtNet	PxLM	PxLM	AdamW	32	1280	16
OFA [106]	CoAtNet	XE	XE, SCST	AdamW	24	1280	16
CoCa [110]	ViT	CL, XE	XE	Adafactor	36	1408	16
GIT [111]	CoSwin [107]	CL, XE	CL, XE, SCST	AdamW	6	768	12
GIT2 [111]	DaViT [118]	UniCL [109], XE	CL, XE, SCST	AdamW	24	1024	16
Universal [112]	CLIP-ViT [53]	Unidirectional LM	Prompt LM, SCST	LAMB [119], Adam	24	1024	16
PaLI [55]	ViT-e	XE	XE	Adafactor	56	1792	16
BLIP-2 [113]	CLIP-ViT	CL, MTL-s, PxLM	CL, MTL-s, PxLM	AdamW	12	768	12

It is worth noting that many of the VLP models deviate from the use of Faster-RCNN as a feature extractor and instead use extractors based on vision transformers [99]. Faster-RCNN has served as a powerful feature extractor for a number of years but the effectiveness of transformer-based feature extractors has resulted in a shift in this area. This has yielded significant improvements in performance and has reduced the reliance on an explicit object detector. It has further shown that the vision part of vision-language models really does matter [39].

As seen from Table 7, the loss functions of the VLP methods are more varied. Apart from the objective loss function used during fine-tuning, they also have an objective loss function for the pre-training. For the pre-training, a common approach is to use the Masked Token Loss (MTL) [38]. The MTL is similar to the Masked Language Modeling (MLM) used in BERT [95]. Whereas in BERT the tokens only pertain to the language domain, in VLP the tokens can belong to the language or visual (image) domain. In Table 7, MTL-b refers to a Masked Token Loss that is bidirectional, whereas MTL-s is sequence-to-sequence [120]. CL2 and CL3 are two-way and three-way [38] contrastive loss functions, respectively. MTL and CL are particularly useful during the pre-training stages, where large-scale text-image-pair datasets are employed. Cross-entropy loss (XE) and CIDEr optimization (SCST) are then used in a similar manner to the non-VLP methods. The effectiveness of the VLP methods is made manifest in the fact that some of the models ([103,110]) do not use SCST optimization and yet they achieve comparable performance to the non-VLP methods that use SCST. VIVO [40] uses the Hungarian matching loss [121] to address the unordered nature of image tags. Some of the models ([103,106,111,111]) stand out for their simplicity in that they use the same objective for pre-training and fine-tuning.

6.4. Results and Performance Metrics

When discussing the results for image captioning, the offline performances based on the Karpathy splits [27] as well as the online performance on the MS COCO evaluation server [68] are often considered. The performance on the Karpathy splits is very useful for purposes of comparison since not all methods will appear on the online server.

Table 8 shows the offline performance metrics of the VLP and non-VLP models. The results for Conceptual Captions were only reported for the online evaluation server and have therefore been left out of Table 8. The CIDEr results are all based on cross-entropy optimization with CIDEr optimization (SCST). A gradual improvement of the results is noted when moving towards the more recent models. COS-Net [91] achieves the best performance among the non-VLP models assessed. The performance of COS-Net is remarkable since, being a non-VLP model, it outperforms a number of other models that use VLP. This goes to show the effectiveness of performing grid feature extraction using an encoder based on CLIP as well the approach of comprehending and ordering the rich semantics in images.

Table 8. Offline performance (Karpathy test split) of the VLP and non-VLP models. B refers to BLEU-4, S to SPICE, M to METEOR, R to ROUGE-L and C to CIDEr. $C^{\dagger}$ gives values for CIDEr optimization.

Model	VLP	B	S	M	R	C	$C^{\dagger}$
Stacked attention [77]	✗	33.3	-	-	54.8	-	108.1
Object relation [79]	✗	38.6	22.6	28.7	58.4	-	128.3
AoA [84]	✗	40.2	22.8	29.3	59.4	122.7	132.0
Entangled [85]	✗	39.9	22.6	28.9	59.0	119.3	127.6
Meshed-Memory [86]	✗	40.5	23.5	29.7	59.5	-	134.5
X-Linear [87]	✗	40.7	23.8	29.9	59.7	122.1	135.3
Image Transformer [90]	✗	39.5	22.8	29.1	59.0	-	130.8
COS-Net [91]	✗	42.9	24.7	30.8	61.0	129.5	143.0
OSCAR [38]	✓	41.7	24.5	30.6	-	127.8	140.0
Unified VLP [109]	✓	39.5	23.2	29.3	-	-	129.3
VinVL [39]	✓	41.0	25.2	31.1	-	130.8	140.9
VIVO [40]	✓	34.9	21.7	28.4	-	119.8	-
LEMON [52]	✓	42.6	25.5	31.4	-	139.1	145.5
SimVLM [103]	✓	40.6	25.4	33.7	-	143.3	-
OFA [106]	✓	44.9	26.6	32.5	-	145.3	154.9
CoCa [110]	✓	40.9	24.7	33.9	-	143.6	-
GIT [111]	✓	44.1	26.3	32.2	-	144.8	151.1
GIT2 [111]	✓	44.1	26.4	32.2	-	145.0	152.7
Universal [112]	✓	42.9	25.2	31.5	-	-	150.2
PaLI [55]	✓	-	-	-	-	149.1	-
BLIP-2 [113]	✓	43.7	-	-	-	145.8	-

For the VLP models, the Rouge Metric has been omitted since many of the authors of these VLP models do not report it. The CIDEr performance is indicated in the case of cross-entropy loss and SCST CIDEr optimization. In all instances where there is CIDEr optimization, there is an improvement over plain cross-entropy loss, which shows the effectiveness of CIDEr optimization. The improvements gained from vision-language pre-training are evident when comparing the non-VLP and VLP results in Table 8. The OFA [106] and GIT2 [111] models have the best performance, achieving CIDEr scores of 154.9 and 152.7, respectively. Note that GIT2 is a bigger version of GIT. The Universal Captioner [112] also achieves a high CIDEr score of 150.2, which supports the effectiveness of their approach of taking the natural descriptive style of captions into account to generate more fluent and human-like captions.

Table 9 shows the online performance of both the non-VLP and VLP models on the MS COCO evaluation server (for the cases in which those results are available). The

VLP methods continue to yield better performance than the non-VLP methods. The best-performing models are GIT and OFA, which corroborates their good performance as reported in the offline Karpathy splits. As of 17th August 2023, GIT2, OFA and GIT were at the top three positions on the *public* leaderboard [69].

Table 9. Online performance based on the evaluation test server. CIDEr values are based on evaluation with 40 reference captions per image.

Model	B	M	R	C
Conceptual Captions [59]	-	0.4	0.7	1.0
AoA [84]	71.2	38.5	74.5	129.6
Entangled Transformer [85]	70.2	38.0	73.9	124.4
Meshed-Memory [86]	72.8	39.0	74.8	132.1
X-Linear [87]	72.4	39.2	75.0	133.5
Image Transformer [90]	71.5	38.4	74.5	129.6
VinVL [39]	74.9	40.8	76.8	138.7
COS-Net [91]	74.7	40.1	76.4	138.3
OFA [106]	78.7	42.7	79.0	149.6
GIT [111]	78.3	42.0	78.4	148.8
GIT2 [111]	78.3	42.1	78.4	149.8

6.5. Out-of-Domain Captioning

A key challenge for captioning systems is that they largely perform well on test sets closely related to the training datasets whereas when gauged against other out-of-domain datasets, their performance is wanting.

In trying to meet this challenge, Hendricks et al. [122] proposed the Deep Compositional Captioner (DCC) to enable generation of captions relating to objects not present in paired image-sentence datasets used during training. Their image encoder is a CNN and the decoder for the language model is an LSTM. They developed their out-of-domain test dataset by holding out a subset of the MS COCO dataset, ensuring that the out-of-domain dataset has objects not seen during training. Venugopalan et al. [123] developed the Novel Object Captioner (NOC), which was an extension of DCC [122].

Yao et al. [124] developed LSTM-C, which is a Long Short-Term Memory with a *Copying* Mechanism. The copying mechanism gives the model the ability to select novel objects learned from external sources and insert them at the appropriate places in the generated captions. Li et al. [125] developed a similar model (Long Short-Term Memory with Pointing, LSTM-P), which enabled vocabulary expansion and incorporation of novel object categories via a *pointing* mechanism.

Lu et al. [126] came up with Neural Baby Talk, an attention-based method for captioning that is based on using visual concepts from an image to fill in slots in a sentence template. The visual concepts are derived from object detectors. This allows the use of different object detectors to help fill in the template slots. This lends itself to generating sentences based on out-of-domain images entailing novel objects not seen during training. The Decoupled Novel Object Captioner (DNOC) [127] follows a similar approach of filling in slots, referred to as placeholders. The Switchable Novel Object Captioner [128] was an extension of the DNOC model in which the placeholder during sentence generation was replaced by a proxy visual word, which is meant to enable generation of better sentences by taking advantage of visual similarities between novel objects and seen objects. They used a *switchable* LSTM which switches between standard LSTM sentence generation and retrieving object classes from a key-value object memory. Anderson et al. [129] developed an architecture-agnostic model, which uses a constrained beam search to enforce the inclusion of selected words (novel object classes) in the generated captions.

Many of the works follow the approach of [122] of using a held-out subset of MS COCO for the out-of-domain dataset. The interest in this research area has brought about the creation of the Novel Object Captioning at Scale (Nocaps) dataset [57]. This dataset offers

a large-scale benchmark which is geared towards improving out-of-domain captioning. Nocaps draws from MS COCO and the Open Images dataset, which contains many object categories not present in MS COCO. It has been seen that models that perform well on the standard MS COCO do not necessarily perform very well on Nocaps or when exposed to test sets containing object classes not seen during training.

Table 10 shows the CIDEr scores of the VLP models on the Nocaps dataset [57]. The 'in', 'near' and 'out' refer to the metrics based on the in-domain, near-domain and out-of-domain subsets, respectively. The in-domain subset contains images with object classes seen during training, the out-of-domain subset has images with object classes not seen during training and the near-domain subset has images with both in-domain and out-of-domain object classes. There is generally a decrease in the performance when it comes to the out-of-domain scores, showing the increased difficulty in carrying out the out-of-domain captioning. The higher CIDEr scores in Tables 8 and 9 compared to those in Table 10 also illustrates the same challenge of out-of-domain captioning.

Table 10. CIDEr scores for VLP models on the Nocaps test split dataset.

Model	In	Near	Out	Overall
OSCAR [38]	84.8	82.1	73.8	80.9
VIVO [40]	89.0	87.8	80.1	86.6
VinVL [39]	98.0	95.2	78.0	92.5
LEMON [52]	112.8	115.5	110.1	114.3
SimVLM [103]	113.7	110.9	115.2	112.2
CoCa [110]	-	-	-	120.6
GIT [111]	122.4	123.9	122.0	123.4
GIT2 [111]	124.2	125.5	122.3	124.8
UniversalCap [112]	118.9	120.6	114.3	119.3
PaLI [55]	-	-	-	124.4
BLIP-2 [113]	123.7	120.2	124.8	121.6

7. Challenges and Future Directions

The models assessed in the previous sections have shown the potential and power of transformers in the specific computer vision task of image captioning. The transformer models have yielded better performance than the previous mechanisms. Despite the fact that there have been clear advancements, there remain a number of open challenges, which present a wide field for research in the area. Image captioning continues to be an active area of research as evidenced by the frequent and current submissions of improved models to the MS COCO leaderboard [69]. In this section, we look at some of the open challenges and discuss some of the emerging trends and future directions related to transformers in image captioning.

7.1. Better Multimodal Alignment

Image captioning is one of various tasks that have shown the power of models incorporating different modalities. In this case the models entail the vision and the language domains. A key challenge in this regard is coming up with better ways to improve the cross-modal representations of the different domains viz. vision and language domains. Li et al. [38] pointed out the lack of grounding, i.e., the lack of explicitly labeled alignments between image regions and words in corresponding texts. Closely related to this is the ambiguity of the visual representations in the semantic embedding space. The ambiguous and overlapping region features in the embedding space are illustrated in Figure 12. Li et al. [38] proposed the use of object tags for a better alignment. In [112] *pivotal keywords*, which take the style into account, are proposed for better alignment. The use of large-scale text-image-pair datasets for vision-language pre-training using contrastive and masked language modeling losses also considerably assists in improving the multimodal alignment. There is, however, a need for more research on better ways of achieving the alignment

between region features and words as this will result in clearer and more precise generated captions for images.

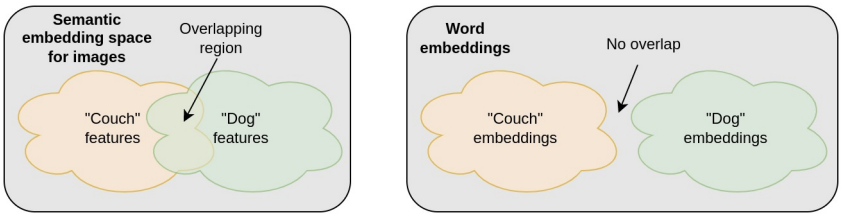

Figure 12. Ambiguous and overlapping region features from images with *couches* and *dogs*. In the semantic embedding space for images,there is a lot of overlap. The word embeddings on the other hand are more distinct. Multimodal alignment can help disambiguate the image semantic embeddings.

7.2. Visual Feature Representation

The bottom-up model of Anderson et al. [35] relied significantly on the well-annotated and attribute-rich Visual Genome dataset [58]. Their object detector was based on Faster RCNN [15]. Their model showed the usefulness of enriching the visual features by employing a variety of other object attributes of image regions instead of only using the object names. Since its invention, the bottom-up model has became the de facto image extractor for many of the subsequent models.

Zhang et al. [39] further showed that the visual features really do matter in image captioning systems. Many previous models had focused on the encoder-decoder aspect of captioning while leaving the feature extractor virtually untouched since the invention of the bottom-up attention model. The VinVL model [39] focused on improving the feature extractor by improving on the underlying object detection through building a bigger and better model as well as employing much larger datasets for training.

We expect that there will be renewed interest and further research in improving the feature extractor used in captioning models. Subsequent to Faster RCNN [15] and the bottom-up approach [35], which have come to constitute a standard approach in object detection, there have been a number of recent developments in object detectors which have yielded improved results.

The Faster RCNN and its family of object detectors are two-stage object detectors, which enjoy relatively high accuracy. However, they tend to be slower than the one-stage object detectors. Single Shot Detector (SSD) [130] and You Only Look Once (YOLO) [131] are classic one-stage object detectors and when they were introduced, they represented a significant step towards real-time object detection. There have been subsequent versions of YOLO released [132–135], which have consistently improved on speed and accuracy. YOLOv3 was already as accurate as SSD but three times faster. The most recent version in the YOLO family is YOLOv7 [136] and there is already ongoing work to release YOLOv8 [137].

Other recent object detectors include [102,138–140]. The recent transformer-based detectors, such as DETR [121], have demonstrated improved performance and facilitate the design of end-to-end transformer-based vision models. More research and advancements in these areas will yield greater benefits for the vision part of captioning systems and other computer vision tasks.

7.3. Cross-Fertilization across Domains

There has been a marked interplay of computer vision and natural language processing (NLP). A number of the advancements that have arisen in computer vision models have been inspired by advancements in the natural language processing domain, and vice versa. Machine translation in particular has had a great impact on image captioning systems. Out of machine translation was born the encoder-decoder framework that has come to dominate

the field of captioning. The attention mechanisms and transformer models that are used in the current state-of-the-art models also had their origins in the language domain.

The language domain has also benefited from some of the approaches used in the vision domain. An example of this is the pre-training of models on large amounts of data to create better-performing natural language processing models. Through this approach, the NLP models end up benefiting from transfer learning and domain adaptation [141]. This kind of transfer learning is a concept that was previously used in computer vision with great success. Many vision models commonly use the large ImageNet dataset [49] pre-training and initialization of the parameters prior to subsequent fine-tuning on downstream tasks such as image classification, object detection and segmentation. ImageNet greatly facilitates the learning of general low-level visual features. ImageNet has had a tremendous impact in the field of computer vision. Together with ImageNet, the Visual Genome dataset [58] and OpenImages [51] have also proven useful for such pre-training. The large-scale use of pre-trianing in natural language tasks resulted in the *ImageNet moment* for NLP [142].

The adoption of mechanisms and models from one domain to another is an approach that will continue to happen as each of the domains becomes enriched. With the most recent approach in image captioning relying on transformers, it will be important to keep looking out for new developments in the language domain that can further improve performance of the captioning systems and computer vision in general.

7.4. Vision-Language Pre-Training (VLP)

It was shown in Section 5 that vision-language pre-training is a powerful approach to improve the performance of captioning systems. VLP makes it possible to perform training using unsupervised or weakly supervised techniques. VLP is specifically pre-training that takes place across domains. It is expected that this will continue to be the trend given that it is worthwhile to research on different ways of improving on the pre-training. For instance, whereas OSCAR [38], Unified VLP [101], VinVL [39] and VIVO [40] all employed pre-training, VIVO focused on a different kind of pre-training which leveraged on building a visual vocabulary and pre-training without the use of complete image captions. This had the consequent result of its better performance in out-of-domain captioning as measured by results drawn from the Nocaps dataset [57]. Recent approaches have entailed pre-training on image-text pairs.

7.5. Bigger Datasets and Models

In machine learning, bigger datasets and models have often resulted in better performance. In image captioning, this has been shown to a great extent by models such as OSCAR [38] and VinVL [39], which combined a number of big datasets to create an even bigger dataset for pre-training and training. Following those works, models trained on even bigger datasets have been presented. The ALT200M [52], WebImageText used by CLIP [53] and the ALIGN [54] datasets have sizes to the tune of 200 million, 400 million and 1.8 billion image-text pairs, respectively. As [52] showed that with sufficient data, the performance keeps improving as the model size increases, we can expect to see increasingly larger datasets.

The larger datasets are accompanied by increasingly larger models. It was also shown by [52] that the model capacity is the performance bottleneck. The big model of the original transformer [36] had 213 million parameters, whereas models such as GIT [111], OFA [106] and CoCa [110] have 545 M, 930 M and 2.1 B parameters, respectively, for their large or giant models. Although the original transformer was purely in the NLP domain whereas GIT, OFA and CoCa combine vision and language, the difference in sizes is, however, illustrative of the fact that the models are getting bigger and bigger.

7.6. More Out-of-Domain Captioning

The results of the works reviewed show that there is still plenty of room for research and improvement, especially when it comes to out-of-domain captioning. Many of the

high-performing models were initially geared towards the MS COCO dataset, which largely involved in-domain captioning. Moving forward, we will see more works based on this challenging Nocaps benchmark [57] and similar datasets. Nocaps also improves on the previous approach of a held-out MS COCO subset [122]. Further research in out-of-domain captioning is useful since the captioning systems are eventually intended to be used in all environments including those that are markedly different from the training environment.

7.7. Few-Shot and Zero-Shot Learning

Many of the current state-of-the-art vision-language models involve weakly supervised pre-training and subsequent task adaptation through fine-tuning on the downstream tasks of interest. It often happens that the fine-tuning still necessitates the use of large datasets to achieve good performance. A current research trend is to develop models that require less data for the downstream fine-tuning.

In few-shot learning, the effort is to develop high-performing models which only require a few training samples in the downstream fine-tuning. A number of works ([103,143–145]) have focused on this kind of model. Flamingo [143] takes inspiration from large language models, which perform quite well on few-shot language-only tasks, and applies similar methods to achieve good performance in vision-language tasks. Flamingo achieves state-of-the-art few-shot learning performance in 16 multimodal language and image/video understanding tasks with as few as 32 examples per task.

When taken to the limit, a situation arises in which a model that is pre-trained can perform competitively without the need for any downstream fine-tuning. This constitutes zero-shot learning and is a more challenging research area. Multimodal vision-language models ([53,54]) have used large-scale contrastive learning to enable zero-shot generalization with regard to novel downstream tasks. However, current zero-shot models are still mainly effective in limited use cases such as classification [143]. LEMON [52] demonstrated a capability for zero-shot caption generation but they noted that their zero-shot captions were very short due to the nature of the pre-training datasets. As more research is carried out these models will improve and will be capable of carrying out more open-ended tasks such as captioning or visual question answering. This will be useful as models will be able to rapidly adapt to diverse and challenging tasks.

7.8. Single Models Performing Multiple Tasks

Some of the recent high-performing models have been effective for multiple tasks [38,39,101]. A trend that is emerging, thanks in large part to the effectiveness of vision-language pre-training, is the design and training of models that can perform well on several downstream tasks, including image-text retrieval, image captioning, novel object captioning, visual question answering and visual reasoning. These models are generally pre-trained on a large common corpus and then fine-tuned on different datasets depending on the downstream task in question. The transformer architecture lends itself to this and we will, therefore, continue to see many models which have good performance on multiple tasks.

7.9. All-Transformer Design

Many of the current captioning systems mainly employ the transformer for the encoding and decoding functions. For a number of models, the visual feature extraction is still carried out based on a CNN of one form or another. A recent line of research is in the use of transformers as well for the feature extraction, thus showing that the reliance on CNNs and explicit object detectors may not be necessary. Dosovitskiy et al. [99] experimented with applying a standard transformer directly to images instead of using a feature extractor, e.g., Faster-RCNN. They came up with the Vision Transformer (ViT). In their work, an image is split into patches and then passed through a linear embedding whose output is then used as input to a transformer. The image patches are treated as tokens and the transformer is then trained on an image classification task in a supervised fashion.

Their vision transformer matches or exceeds the state-of-the-art on a number of image classification datasets while requiring less computational resources. Since ViT, a number of other transformer-based feature extractors have been developed, e.g., CoAtNet [104], CoSwin [107], CLIP-ViT [53] and DaViT [118]. These have become the basis of the current state-of-the-art models ([103,106,110–112]).

Touvron et al. [146] improve on ViT and achieve competitive results in image classification on ImageNet while requiring even less computational resources than [99] and without the use of external datasets. Carion et al. [121] and Zhu et al. [140] applied transformers to the task of object detection and achieved competitive results. He et al. [90] also explored the idea of an image transformer.

A continuing research area still remains in effectively applying the vision transformer to other computer vision tasks. Another active research area is in improving self-supervised pre-training methods for the vision transformer so as to achieve comparable performance with large-scale supervised pre-training. The transformer architecture was born in the language domain and although it has been used in computer vision tasks, it is by its nature very suited to NLP tasks. There is a need for further research in developing a more refined vision transformer which is well adapted to the nature of images and other computer vision tasks. Improvements in these areas would make possible the design of image captioning systems that are entirely built on transformers.

8. Conclusions

The application of transformers has yielded many performance improvements, first in natural language processing and then in computer vision. In this paper we have looked at the application of transformers to the specific computer vision task of image captioning. We focused on transformer-based approaches because they are the backbone of the current state-of-the-art models and also because other traditional models have been adequately assessed in other reviews. We first reviewed the various datasets and metrics used in training and evaluation of image captioning systems. We also reviewed the concept of pre-training and saw how the kind of pre-training used in natural language processing has found application in computer vision in the form of vision-language pre-training. We then studied a number of captioning models that are based on the vanilla transformer architecture as well as those that employ vision-language pre-training. We made an in-depth analysis of these models, which entailed a study of a number of the design choices for each model. We also compared the performance of the different models and looked at the chief contributions of each model. In general, the current state-of-the-art models are those that use vision-language pre-training. We concluded by looking at some of the challenges faced, the emerging trends and future directions. Transformer-based systems continue to be dominant in vision and language models and their application to image captioning in particular is a vibrant area of research.

Author Contributions: Conceptualization, O.O.; writing—original draft preparation, O.O.; writing—review and editing, O.O., H.O. and P.A.; supervision, H.O. and P.A.; funding acquisition, H.O. and P.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the African Development Bank (AfDB) through the Kenyan Ministry of Education in conjunction with the University of Nairobi (Contract No. MOE/HEST/02/2017-2018) and the National Research Fund (NRF) Kenya-South Africa No. 5-2017.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Farhadi, A.; Hejrati, M.; Sadeghi, M.A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; Forsyth, D. Every Picture Tells a Story: Generating Sentences from Images. In Proceedings of the Computer Vision—ECCV 2010, Crete, Greece, 5–11 September 2010; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 15–29.
2. Kulkarni, G.; Premraj, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. Baby Talk: Understanding and Generating Simple Image Descriptions. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011 ; pp. 1601–1608. [CrossRef]
3. Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Goyal, A.; Berg, A.; Yamaguchi, K.; Berg, T.; Stratos, K.; Daumé, H. Midge: Generating Image Descriptions from Computer Vision Detections. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France, 23–27 April 2012; pp. 747–756.
4. Kuznetsova, P.; Ordonez, V.; Berg, A.; Berg, T.; Choi, Y. Collective Generation of Natural Image Descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jeju Island, Republic of Korea, 8–14 July, 2012; pp. 359–368.
5. Kuznetsova, P.; Ordonez, V.; Berg, T.L.; Choi, Y. TreeTalk: Composition and Compression of Trees for Image Descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 351–362. [CrossRef]
6. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
7. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *arXiv* **2014**, arXiv:1409.0575.
8. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet Classification with Deep Convolutional Neural Networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems , Lake Tahoe, NV, USA, 3–6 December 2012 ; pp. 1097–1105.
9. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. *arXiv* **2013**, arXiv:1311.2901.
10. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015 .
11. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *arXiv* **2014**, arXiv:1409.4842.
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
13. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 580–587. [CrossRef]
14. Girshick, R. Fast R-CNN. *arXiv* **2015**, arXiv:1504.08083.
15. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc. : New York, NY, USA, 2015; pp. 91–99.
16. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *arXiv* **2017**, arXiv:1703.06870.
17. Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal Neural Language Models. In Proceedings of the 31st International Conference on Machine Learning—Volume 32, Beijing, China, 21–26 June 2014; pp. 595–603.
18. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). *arXiv* **2015**, arXiv:1412.6632.
19. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
20. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3156–3164. [CrossRef]
21. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [CrossRef] [PubMed]
22. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2016**, arXiv:1409.0473.
23. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; ; pp. 1724–1734. [CrossRef]
24. Kalchbrenner, N.; Blunsom, P. Recurrent Continuous Translation Models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013; pp. 1700–1709.
25. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA , 2014.
26. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv* **2014**, arXiv:1411.2539.
27. Karpathy, A.; Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3128–3137. [CrossRef]

28. Ba, J.; Mnih, V.; Kavukcuoglu, K. Multiple Object Recognition with Visual Attention. *arXiv* **2015**, arXiv:1412.7755.
29. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent Models of Visual Attention. In Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2, Montreal, QC, Canada, 8–13 December 2014; pp. 2204–2212.
30. Fu, K.; Jin, J.; Cui, R.; Sha, F.; Zhang, C. Aligning Where to See and What to Tell: Image Captioning with Region-Based Attention and Scene-Specific Contexts. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2321–2334. [CrossRef] [PubMed]
31. Cho, K.; Courville, A.; Bengio, Y. Describing Multimedia Content Using Attention-based Encoder–Decoder Networks. *IEEE Trans. Multimed.* **2015**, *17*, 1875–1886.
32. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image Captioning with Semantic Attention. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4651–4659. [CrossRef]
33. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–25 July 2017; pp. 3242–3250. [CrossRef]
34. Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on Machine Learning—Volume 37, Lille, France, 6–11 July 2015; pp. 2048–2057.
35. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 8–22 June 2018; pp. 6077–6086.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc. : New York, NY, USA, 2017.
37. Bloem, P. Transformers from Scratch | Peterbloem.Nl. 2019. Available online: http://peterbloem.nl/blog/transformers (accessed on 12 May 2022).
38. Li, X.; Yin, X.; Li, C.; Hu, X.; Zhang, P.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
39. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. VinVL: Revisiting Visual Representations in Vision-Language Models. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021.
40. Hu, X.; Yin, X.; Lin, K.; Wang, L.; Zhang, L.; Gao, J.; Liu, Z. VIVO: Surpassing Human Performance in Novel Object Captioning with Visual Vocabulary Pre-Training. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 1575–1583.
41. Wang, H.; Zhang, Y.; Yu, X. Hindawi—An Overview of Image Caption Generation Methods. *Comput. Intell. Neurosci.* **2020**, *2020*, e3062706. [CrossRef]
42. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.* **2019**, *51*, 1–36. [CrossRef]
43. Chaudhari, S.; Mithal, V.; Polatkan, G.; Ramanath, R. An Attentive Survey of Attention Models. *arXiv* **2021**, arXiv:1904.02874.
44. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in Vision: A Survey. *arXiv* **2021**, arXiv:2101.01169.
45. Stefanini, M.; Cornia, M.; Baraldi, L.; Cascianelli, S.; Fiameni, G.; Cucchiara, R. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 539–559. . [CrossRef] [PubMed]
46. Services, A.W. Amazon Mechanical Turk. 2023. Available online: https://www.mturk.com/ (accessed on 15 August 2023).
47. University, P. WordNet. 2023. Available online: https://wordnet.princeton.edu/ (accessed on 15 August 2023).
48. Miller, G.A. WordNet: A Lexical Database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
49. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
50. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision—ECCV, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.
51. Kuznetsova, A.; Rom, H.; Alldrin, N.G.; Uijlings, J.R.R.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. The Open Images Dataset V4. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981. [CrossRef]
52. Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; Wang, L. Scaling up Vision-Language Pretraining for Image Captioning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 17959–17968. [CrossRef]
53. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Volume 139 , pp. 8748–8763.

54. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up Visual and Vision-Language Representation Learning with Noisy Text Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Volume 139, , pp. 4904–4916.
55. Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.J.; Padlewski, P.; Salz, D.; Goodman, S.A.; Grycner, A.; Mustafa, B.; Beyer, L.; et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
56. Lab, S.V. ImageNet. 2016. Available online: http://image-net.org/ (accessed on 2 March 2018).
57. Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; Anderson, P. Nocaps: Novel Object Captioning at Scale. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8947–8956. [CrossRef]
58. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. [CrossRef]
59. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 2556–2565.
60. Changpinyo, S.; Sharma, P.; Ding, N.; Soricut, R. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training to Recognize Long-Tail Visual Concepts. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Event, 19–25 June 2021; IEEE Computer Society: Los Alamitos, CA, USA, 2021; pp. 3557–3567. [CrossRef]
61. Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *J. Artif. Int. Res.* **2013**, *47*, 853–899. [CrossRef]
62. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
63. Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C.L.; Parikh, D.; Batra, D. VQA: Visual Question Answering. *Int. J. Comput. Vision* **2017**, *123*, 4–31. [CrossRef]
64. Visual Question Answering. Available online: http://visualqa.org (accessed on 12 August 2023).
65. Hudson, D.A.; Manning, C.D. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 6693–6702. [CrossRef]
66. Ordonez, V.; Kulkarni, G.; Berg, T.L. Im2Text: Describing Images Using 1 Million Captioned Photographs. In Proceedings of the 24th International Conference on Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; Curran Associates Inc.: Red Hook, NY, USA, 2011; pp. 1143–1151.
67. Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; Sun, J. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8429–8438. [CrossRef]
68. Pavao, A.; Guyon, I.; Letournel, A.C.; Baró, X.; Escalante, H.; Escalera, S.; Thomas, T.; Xu, Z. *CodaLab Competitions: An Open Source Platform to Organize Scientific Challenges*; Technical Report; Université Paris-Saclay: Paris, France, 2022.
69. CodaLab-Microsoft COCO Image Captioning Challenge. 2015. Available online: https://competitions.codalab.org/competitions/3221#results (accessed on 31 January 2023).
70. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; Association for Computational Linguistics: Stroudsburg, PA, USA , 2002; pp. 311–318. [CrossRef]
71. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the ACL-04 Workshop, Barcelona, Spain, 25–26 July 2004; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
72. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, 29 June 2005; Association for Computational Linguistics: Ann Arbor, Michigan, 2005; pp. 65–72.
73. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4566–4575. [CrossRef]
74. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: Semantic Propositional Image Caption Evaluation. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016 ; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 382–398.
75. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollar, P.; Zitnick, C.L. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv* **2015**, arXiv:1504.00325.
76. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**, arXiv:1602.07261.
77. Zhu, X.; Li, L.; Liu, J.; Peng, H.; Niu, X. Captioning Transformer with Stacked Attention Modules. *Appl. Sci.* **2018**, *8*, 739. [CrossRef]

78. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 1492–1500.
79. Herdade, S.; Kappeler, A.; Boakye, K.; Soares, J. Image Captioning: Transforming Objects into Words. In Proceedings of the Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 11135–11145.
80. Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; Wei, Y. Relation Networks for Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3588–3597.
81. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-Encoding Scene Graphs for Image Captioning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 10677–10686.
82. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring Visual Relationship for Image Captioning. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 711–727.
83. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Hierarchy Parsing for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2621–2629. [CrossRef]
84. Huang, L.; Wang, W.; Chen, J.; Wei, X.Y. Attention on Attention for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4633–4642.
85. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled Transformer for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8927–8936. [CrossRef]
86. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. *arXiv* **2020**, arXiv:1912.08226.
87. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-Linear Attention Networks for Image Captioning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10968–10977.
88. Fukui, A.; Park, D.H.; Yang, D.; Rohrbach, A.; Darrell, T.; Rohrbach, M. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 457–468. [CrossRef]
89. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
90. He, S.; Liao, W.; Tavakoli, H.R.; Yang, M.; Rosenhahn, B.; Pugeault, N. Image Captioning Through Image Transformer. In Proceedings of the Asian Conference on Computer Vision (ACCV), Kyoto, Japan, 30 November–4 December, 2020; Ishikawa, H., Liu, C.L., Pajdla, T., Shi, J., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 153–169.
91. Li, Y.; Pan, Y.; Yao, T.; Mei, T. Comprehending and Ordering Semantics for Image Captioning. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 17969–17978. [CrossRef]
92. Li, Y.; Pan, Y.; Chen, J.; Yao, T.; Mei, T. X-Modaler: A Versatile and High-Performance Codebase for Cross-Modal Analytics. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 3799–3802. [CrossRef]
93. Jing, L.; Tian, Y. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 4037–4058. [CrossRef]
94. Liu, X.; Zhang, F.; Hou, Z.; Wang, Z.; Mian, L.; Zhang, J.; Tang, J. Self-Supervised Learning: Generative or Contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 857–876.
95. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
96. Taylor, W.L. "Cloze Procedure": A New Tool for Measuring Readability. *J. Mass Commun. Q.* **1953**, *30*, 415–433. [CrossRef]
97. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In Proceedings of the Annual Conference on Neural and Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 13–23.
98. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA , 2019; pp. 5099–5110.
99. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, Austria, 3–7 May 2021.
100. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; Chaudhuri, K., Salakhutdinov, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA , 2019; Volume 97, pp. 6105–6114.

101. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.J.; Gao, J. Unified Vision-Language Pre-Training for Image Captioning and VQA. In Proceedings of the Association for the Advancement of Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 13041–13049.
102. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]
103. Wang, Z.; Yu, J.; Yu, A.W.; Dai, Z.; Tsvetkov, Y.; Cao, Y. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. *arXiv* **2022**, arXiv:2108.10904.
104. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M. CoAtNet: Marrying Convolution and Attention for All Data Sizes. In Proceedings of the Advances in Neural Information Processing Systems, Virtual Event, 6–14 December 2021; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 3965–3977.
105. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2022**, *21*, 1–67.
106. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. OFA: Unifying Architectures, Tasks, and Modalities through a Simple Sequence-to-Sequence Learning Framework. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2002; Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., Eds.; Curran Associates, Inc. : Red Hook, NY, USA, 2022; Volume 162, pp. 23318–23340.
107. Yuan, L.; Chen, D.; Chen, Y.L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. Florence: A New Foundation Model for Computer Vision. *arXiv* **2021**, arXiv:2111.11432.
108. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2022**, arXiv:2108.07258
109. Yang, J.; Li, C.; Zhang, P.; Xiao, B.; Liu, C.; Yuan, L.; Gao, J. Unified Contrastive Learning in Image-Text-Label Space. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022; pp. 19141–19151. [CrossRef]
110. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive Captioners Are Image-Text Foundation Models. *arXiv* **2022**, arXiv:2205.01917.
111. Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; Wang, L. GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv* **2022**, arXiv:2205.14100.
112. Cornia, M.; Baraldi, L.; Fiameni, G.; Cucchiara, R. Universal Captioner: Inducing Content-Style Separation in Vision-and-Language Model Training. *arXiv* **2022**, arXiv:2111.12727.
113. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv* **2023**, arXiv:2301.12597.
114. Zhai, X.; Kolesnikov, A.; Houlsby, N.; Beyer, L. Scaling Vision Transformers. *arXiv* **2022**, arXiv:2106.04560.
115. Rohrbach, A.; Hendricks, L.A.; Burns, K.; Darrell, T.; Saenko, K. Object Hallucination in Image Captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 4035–4045. [CrossRef]
116. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1179–1195. [CrossRef]
117. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.
118. Ding, M.; Xiao, B.; Codella, N.; Luo, P.; Wang, J.; Yuan, L. DaViT: Dual Attention Vision Transformers. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022, Proceedings, Part XXIV; Springer: Berlin/Heidelberg, Germany, 2022; pp. 74–92. [CrossRef]
119. You, Y.; Li, J.; Reddi, S.; Hseu, J.; Kumar, S.; Bhojanapalli, S.; Song, X.; Demmel, J.; Keutzer, K.; Hsieh, C.J. Large Batch Optimization for Deep Learning: Training BERT in 76 Minutes. *arXiv* **2020**, arXiv.1904.00962.
120. Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; Hon, H.W. Unified Language Model Pre-Training for Natural Language Understanding and Generation. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019.
121. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Virtual Event, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 213–229.
122. Hendricks, L.A.; Venugopalan, S.; Rohrbach, M.; Mooney, R.; Saenko, K.; Darrell, T. Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1–10. [CrossRef]
123. Venugopalan, S.; Hendricks, L.A.; Rohrbach, M.; Mooney, R.; Darrell, T.; Saenko, K. Captioning Images with Diverse Objects. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1170–1178. [CrossRef]

124. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5263–5271. [CrossRef]
125. Li, Y.; Yao, T.; Pan, Y.; Chao, H.; Mei, T. Pointing Novel Objects in Image Captioning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12489–12498. [CrossRef]
126. Lu, J.; Yang, J.; Batra, D.; Parikh, D. Neural Baby Talk. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7219–7228. [CrossRef]
127. Wu, Y.; Zhu, L.; Jiang, L.; Yang, Y. Decoupled Novel Object Captioner. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1029–1037. [CrossRef]
128. Wu, Y.; Jiang, L.; Yang, Y. Switchable Novel Object Captioner. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1162–1173. [CrossRef]
129. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; Association for Computational Linguistics: Stourdburg, PA, USA, 2017; pp. 936–945. [CrossRef]
130. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
131. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788. [CrossRef]
132. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525. [CrossRef]
133. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
134. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
135. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
136. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *arXiv* **2022**, arXiv:2207.02696.
137. Ultralytics. Ultralytics YOLOv8 Docs. 2023. Available online: https://docs.ultralytics.com/ (accessed on 10 August 2023).
138. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
139. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A Simple and Strong Anchor-free Object Detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1922–1933. [CrossRef]
140. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
141. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; IT Press: Cambridge, MA, USA, 2016. Available online: http://www.deeplearningbook.org (accessed on 10 August 2023).
142. Ruder, S. NLP's ImageNet Moment Has Arrived. Available online: https://www.ruder.io/nlp-imagenet/ (accessed on 7 August 2023).
143. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv* **2022**, arXiv.2204.14198.
144. Tsimpoukelli, M.; Menick, J.L.; Cabi, S.; Eslami, S.M.A.; Vinyals, O.; Hill, F. Multimodal Few-Shot Learning with Frozen Language Models. In Proceedings of the Advances in Neural Information Processing Systems, Virtual Event, 6–14 December 2021; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 200–212.
145. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Virtual Event, 6–12 December 2020; Curran Associates, Inc.: Red Hook, NY, USA, 2020.
146. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jegou, H. Training Data-Efficient Image Transformers & Distillation through Attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Meila, M., Zhang, T., Eds.; Curran Associates, Inc.: Red Hook, NY, USA , 2021; Volume 139, pp. 10347–10357.