algorithms

MDPI

Article

Nemesis: Neural Mean Teacher Learning-Based Emotion-Centric Speaker

Aryan Yousefi and Kalpdrum Passi *

School of Engineering and Computer Science, Laurentian University, Sudbury, ON P3E 2C6, Canada
* Correspondence: kpassi@laurentian.ca

Abstract: Image captioning is the multi-modal task of automatically describing a digital image based on its contents and their semantic relationship. This research area has gained increasing popularity over the past few years; however, most of the previous studies have been focused on purely objective content-based descriptions of the image scenes. In this study, efforts have been made to generate more engaging captions by leveraging human-like emotional responses. To achieve this task, a mean teacher learning-based method has been applied to the recently introduced ArtEmis dataset. ArtEmis is the first large-scale dataset for emotion-centric image captioning, containing 455K emotional descriptions of 80K artworks from WikiArt. This method includes a self-distillation relationship between memory-augmented language models with meshed connectivity. These language models are trained in a cross-entropy phase and then fine-tuned in a self-critical sequence training phase. According to various popular natural language processing metrics, such as BLEU, METEOR, ROUGE-L, and CIDEr, our proposed model has obtained a new state of the art on ArtEmis.

Keywords: image captioning; mean teacher learning; self-distillation; self-critical sequence training; natural language processing

Citation: Yousefi, A.; Passi, K. Nemesis: Neural Mean Teacher Learning-Based Emotion-Centric Speaker. *Algorithms* **2023**, *16*, 97. https://doi.org/10.3390/a16020097

Academic Editor: Ayan Biswas

Received: 12 December 2022
Revised: 26 January 2023
Accepted: 7 February 2023
Published: 9 February 2023

1. Introduction

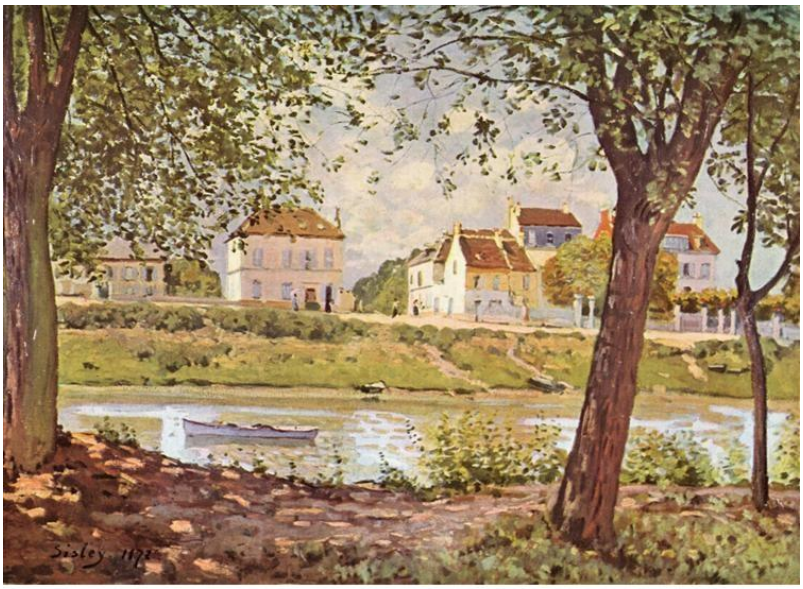
Image captioning is an important step towards developing scene-understanding ability in deep learning models for plenty of purposes. This multi-modal task deals with both textual and visual modalities with the goal of generating fluent natural language descriptions according to the contents of a digital image [1]. The applications of image captioning include usage in virtual assistants, helping visually impaired people to obtain a better perception of their surroundings, and industrial quality control. The early studies were based on retrieval-based [2,3], and template-based methods [4–6], where the captions are directly retrieved from an existing database causing the captions to be repetitive and not completely specific to the input image. The next step in the evolution of image captioning models was utilizing convolutional neural networks as visual feature extractors [7–10], and recurrent-neural-network-based modules as the caption generator of their model operating in an auto-regressive manner [7,10]. Recently, fully attentive transformer-based models have been one of the main trending methods to tackle this problem, following different variations of attentive encoder/decoder configurations utilizing self-attention [11].

However, most of the previous work is focused on generating purely objective content-based descriptions. Over time, these descriptions have gotten increasingly accurate, but they lack personality, emotion, and other human-like attributes. Stylized image captioning was the next step to overcome this limitation by generating captions following a given linguistic style. The absence of a large-scale dataset containing stylized (image, text) pairs has also led to adopting semi-supervised approaches, usually exploiting an unpaired stylized corpus to extract different styles [12–15]. However, some datasets have been introduced, such as Personality-Captions [12], containing 215 personality traits paired with images and their corresponding captions. FlickrStyle10k [16] also contains (image, text)

pairs with two different styles. In stylized image captioning, the core concept is generating objective descriptions, only with different wording according to the target linguistic style.

Various modalities can be utilized in the task of attention-based image captioning. Lu et al. [17] introduced a novel sound active attention framework to generate captions more specific to the interest of the observer. The process is similar to the active attention in the human mind generated by top-down signals. On the other hand, Wang et al. [18] proposed NUAN, a non-uniform attention network for multi-modal feature fusion. The attention mechanism in NUAN considers three modalities of text, sound, and vision non-uniformly, where the text is utilized as a determinative representation while visual and acoustic representations are leveraged to obtain a solid representation.

Achlioptas et al. [19] introduced ArtEmis, the first large-scale dataset for affective image captioning accompanied by baseline neural speakers to generate emotion-based descriptions. In addition to being subjective, rich, and diverse, this dataset contains affective language. Figure 1 shows an example from the ArtEmis dataset containing multiple emotional responses for the same artwork. In this paper, we propose Nemesis, a Neural Mean Teacher Learning-based Emotion-centric Speaker trained on ArtEmis. Our model is capable of generating affective utterances describing the triggered human-like emotional responses stemming from visual stimuli. The generated utterances are according to both the visual features and emotion-based supervision signals.



Something Else: "serenity in the still, smooth waters and in the background a rustic village."

Contentment: "The feel of a beautiful sunny day by the lake and seeing my home across the way. I picture this within my mind and brings peace to my heart."

Awe: "The scenery painted is absolutely beautiful and makes you feel like you are there."

Figure 1. An example from the ArtEmis dataset containing multiple emotional responses for the same artwork. You can see the different descriptions along with their corresponding emotional class (in bold font).

The proposed pipeline consists of:

- Auxiliary text-to-emotion and image-to-emotion classification tasks;
- A visual encoder to extract visual features from the input image;

- Two interconnected language models following a transformer-based encoder/decoder architecture.

In Nemesis, we introduce two main contributions:

1. A novel approach for the image-to-emotion classification task by decreasing the texture bias of the classifier and encouraging the model towards a shape-based classification. This is because of the differing local textures in our input images (artworks) in comparison to the real world;
2. Achieving a state-of-the-art performance using the Nemesis on the ArtEmis dataset. We suggest that a self-critical mean teacher learning-based approach, supervised by extra emotional signals, is a promising path towards generating more human-like, emotionally rich captions.

1.1. Related Work

1.1.1. Mean Teacher Learning

Mean teacher learning is a semi-supervised paradigm based on the interaction between two models referred to as the teacher model and the student model. In the first place, Samuli et al. [20] proposed a novel architecture in which the temporal ensembling maintains an exponential moving average of the target predictions, while the inconsistent predictions are penalized by taking the mean squared error between the predictions of both models. However, temporal ensembling had a slow pace in utilizing the learned information in the training process since the targets are updated only once per epoch. Hence, it was not efficient when applied to large-scale datasets. Tarvainen et al. [21] proposed mean teacher learning, in which the teacher model maintains an average of the student model's weights consecutively during the training steps instead of the label predictions. We follow the latter approach in our proposed model.

1.1.2. Knowledge Distillation

The latent knowledge encapsulated within a larger network is often referred to as "dark knowledge" [22]. Knowledge distillation (KD) [23] is the self-supervised process of transferring dark knowledge from a bigger model to a smaller one, which has been shown to be utilized effectively in various vision-language tasks. The same process is called self-distillation when the models have equal sizes. In our case, the teacher model supplies an extra supervision signal to the student model to improve in the imitation of its behavior by providing predicted soft labels [24].

1.1.3. Self-Critical Sequence Training

Policy gradient-based reinforcement learning methods, such as REINFORCE [25], have been utilized in image captioning to overcome the mismatch and the exposure bias between the optimizing function and the non-differentiable evaluation metrics [26–28]. Self-critical sequence training (SCST) [29] is a special case of REINFORCE in which the model's own test-time inference is used to normalize the rewards it experiences rather than estimating the reward and the normalization method. Rennie et al. [29] proposed that directly optimizing the CIDEr metric [30] through the SCST process during the test-time can be a highly effective way to overcome the non-differentiability of such metrics and boost the performance significantly. Yang et al. [31] introduced Variational Transformer, in which the SCST process uses the range median of all samples to improve the diversity without sacrificing the accuracy.

1.1.4. Visual Encoding

As an early stage in an image captioning pipeline, the spatial information and structure are extracted from our input image to achieve a proper visual representation. Some works utilize non-attentive CNNs to extract the global features [7,10], while some other methods are based on grid-based and region-based feature extraction using additive attention [9,32].

In our case, we extract the visual features following the recent approach of employing large-scale vision-language architectures [33], such as CLIP [34] and BLIP [35].

1.1.5. Auxiliary Emotion Classification Tasks

Emotions in ArtEmis dataset are divided into 9 emotional classes; we have amusement, awe, contentment, and excitement as positive emotions, while we have anger, fear, disgust, and sadness as negative emotions. In addition, a ninth class named something else has been considered to express having no particular emotions or an additional feeling not listed. Following the work of [19], we employ two classifiers for our auxiliary emotion classification tasks, which will be utilized in both the captioning process and evaluation. We are basically facing a nine-way classification problem corresponding to each emotional class. The first module is a text-to-emotion classifier to predict the dominant emotional class of an utterance, and the utilization of this module has been discussed in Section 3.2.2. The text-to-emotion classification task is achieved by utilizing a fine-tuned pre-trained Bert model [36] to classify utterances to the emotional class to which they most likely belong.

The second module is an image-to-emotion classifier to predict the dominant emotional class of a visual input. Since the ArtEmis dataset consists of artworks and sketches, the local textures mostly differ from the ones in the real world. In addition, the human mind also tends to focus on shape-based information. On the other hand, the models pre-trained on the ImageNet [37] dataset are biased towards local textures. Therefore, we have utilized a ResNet-50 [38] encoder pre-trained on the Stylized-ImageNet [39] dataset, which is an augmentation of the standard ImageNet dataset. In Stylized-ImageNet, the local textures are heavily distorted, while the shapes have remained intact to increase the shape bias. Basically, we want to train our classifier to detect objects based on shapes rather than local textures.

2. Materials and Methods

Neural Mean Teacher Learning-based Emotion-centric Speaker or Nemesis is our proposed neural speaker capable of leveraging emotional supervision signals in the caption generation process. In this section, we will elaborate on the pipeline, architecture, and finally, the training strategy.

2.1. Pipeline

The pipeline of Nemesis, as shown in Figure 2, consisted of a visual encoder extracting visual features from the input image, then passing them to both the student model and the teacher model. Inspired by the work of Barraco et al. [40], both models had identical architectures linked based on two types of interactions: (1) the self-distillation process, where the teacher model provides regression targets via passing its predicted logits as soft labels to the equally sized student model [24]. This extra supervision signal enhances the ability of the student model to imitate the behavior of the teacher model. (2) The teacher model performs a form of model ensembling by updating its parameters θ_t according to the exponential moving average (EMA) [41] of the student model's parameters θ_s. This updating procedure can be formally defined by the equation below:

$$\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s, \tag{1}$$

where λ is a value between $[0,\ 1]$, indicating the intensity of this update. The exact role of these interactions in the training process is discussed in the training strategy section.

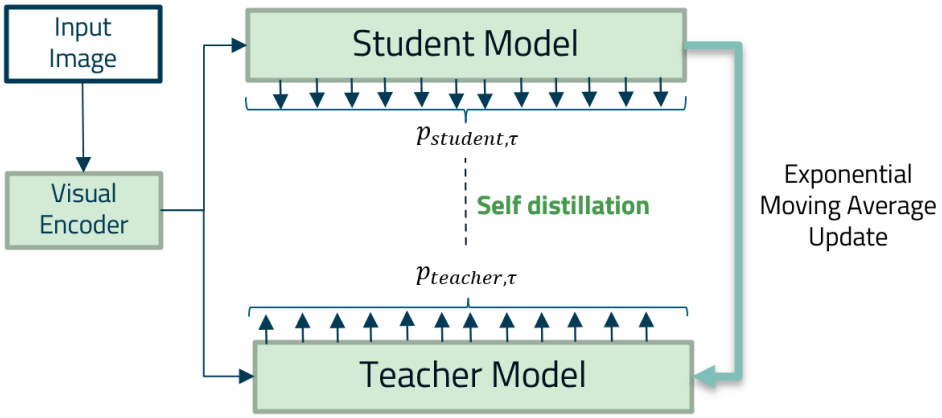


Figure 2. The interactions between our two language models: (1) the EMA update according to the student model's weights. (2) The self-distillation process using the teacher model's predicted logits passed to the student model, which will be treated as soft labels.

Emotional Grounding

In this process, at each time-step, an additional feature according to the dominant emotional class of the input image was also passed to the language models alongside the extracted visual features. This extra emotional signal enabled the model to decouple the emotion conveyed during the caption generation process from the input image's dominant emotional class. This was inspired by how the human mind works while describing an emotional response, where we decide how to feel about something first, and then we put it into words. The dominant emotional class was selected utilizing the image-to-emotion classifier during evaluation. The neural speaker leveraging this supervision signal was called Emotionally Grounded Nemesis or EGNemesis.

2.2. Architecture

Both language models followed the same architecture consisting of a stack of memory-augmented encoders and a stack of meshed decoders [42]. This architecture is illustrated in Figure 3, and different components are elaborated in the following.

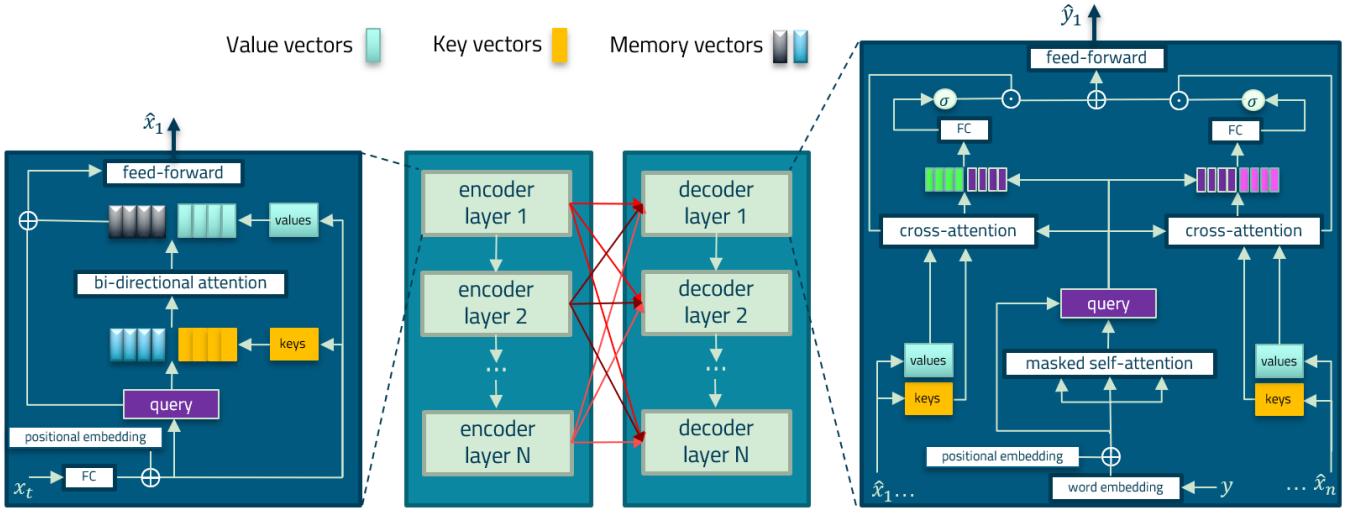


Figure 3. The architecture of both the teacher model and the student model. It consists of a stack of memory-augmented encoders and a stack of meshed decoders. The memory-augmented encoder encodes the multi-level visual relationships leveraging the priori knowledge provided by the memory vectors. The meshed decoder generates the textual tokens leveraging the meshed connectivity illustrated by the red arrows.

2.2.1. Memory-Augmented Encoder

The memory-augmented encoder utilized bi-directional attention to process the visual features received from the visual encoder. However, using only bi-directional attention would have deprived us of incorporating any priori knowledge in our encoding procedure. Hence, we utilized additional independent learnable memory vectors along with the key and value vectors to encode the additional priori knowledge. Finally, the encoder's output was the result of a feed-forward network applied to the memory-augmented bi-directional attention result. The outputs of all encoder layers were passed to each decoder layer via a meshed-like connectivity. The memory-augmented attention is formally defined by the equation below:

$$\begin{aligned} MemAug_{att}(X) &= Attention\big(W_q X,\ K_{MemAug},\ V_{MemAug}\big), \\ K_{MemAug} &= concat(W_k X,\ Mem_k), \\ V_{MemAug} &= concat(W_v X,\ Mem_v), \end{aligned} \tag{2}$$

where X is the input set, W_q, W_k, W_v are matrices of learnable weights, and Mem_k and Mem_v are learnable memory matrices.

2.2.2. Meshed Decoder

Our decoder predicted the next word in an auto-regressive manner according to both the previously generated words and the encoder outputs. It applied right-masked self-attention to process the input sequence and utilized cross-attention to process the encoder outputs received through the meshed-like connection. This meshed-like connectivity enabled our model to extract both low-level and high-level features through a meshed cross-attention process. The cross-attention module used queries based on the self-attention results and keys and values from the encoder outputs. Finally, the output of the position-wise feed-forward layer gave us the output logit at each time-step.

2.3. Training Strategy

2.3.1. Cross-Entropy (XE) Training

In this phase, the student model faces two objectives. The first objective is to optimize the cross-entropy loss according to the previously generated utterances, the input image, and the model parameters. This process is formally defined by the equation below:

$$\mathcal{L}(\theta) = \mathbb{E}_{x \sim D} \sum_{\tau} \log(u_\tau | u_{k<\tau},\ i,\ \theta), \tag{3}$$

where θ is the model parameters, u_τ is the generated utterance at time-step τ, and i is the input image.

The second objective for the student model was to optimize the self-distillation loss with respect to the student model's parameters. Self-distillation loss is defined as the mean squared error (MSE) between the output logits of both models. This process follows the expression below:

$$\min_{\theta_s} \sum_{\tau} (p_{t,\,\tau} - p_{s,\,\tau})^2, \tag{4}$$

where $p_{t,\,\tau}$ and $p_{s,\,\tau}$ are the output logits over the vocabulary of N words at time-step τ for the teacher model and the student model, respectively.

The next step was the EMA update of the teacher model's parameters θ_t based on the student model's parameters θ_s following the expression mentioned in Equation (1). This process was conducted after each stochastic gradient descent (SGD) update of the student model. It enabled the teacher model to keep up with the improvement in the student model in a stable and steady manner.

2.3.2. SCST Fine-Tuning

This phase is a special case of REINFORCE [25], where the idea is to weight the samples outperforming the current test-time model positively and, in contrast, weight the samples

inferior to the current test-time model negatively. This weighting process was completed via the reward function utilized to assign a proper score to the generated utterances at each time-step. For this purpose, the CIDEr-D [30] metric was used as the reward function, and we used the model's own test-time inference to normalize the rewards. Specifically, at each time-step, the top-1 utterance in each of the k returned beams was assigned with a proper reward, and the average of the rewards was used as a baseline to normalize them and reduce the variance. This phase enabled us to overcome the non-differentiability of such metrics and boost the performance significantly. The SCST-based fine-tuning process is formally defined by the expression below:

$$\nabla_\theta \mathcal{L}(\theta) = -\frac{1}{k}\sum_{j=1}^{k}((r\left(u^j\right) - ((\sum_j r\left(u^j\right))/k))\nabla_\theta \log p\left(u^j\right)), \quad (5)$$

where u^j is the j-th utterance in the beam, $r(.)$ is the reward function, and $(\sum_j r\left(u^j\right))/k$ is the average of the rewards to normalize the value. The utilization of other metrics, such as emotional alignment [19], was also experimented with. In this metric, the text-to-emotion classifier was utilized to predict the emotional class of the generated caption, then it was compared to the dominant emotional class of the ground-truth captions, and the percentage of matches is our score. However, the results were not acceptable due to the high variance of such metrics.

2.4. Dataset

The ArtEmis dataset [19] was utilized to train and evaluate our proposed model. It contains 454,684 emotion-centric utterances related to 80,031 artworks publicly available in the WikiArt (https://www.wikiart.org, accessed on 21 February 2022) dataset. The corpus contains 37,250 distinct words, which took 11,138 h to gather by 6788 annotators via Amazon's Mechanical Turk (AMT) services. The reason for using artworks is that they are the best tools to trigger emotional responses. Each utterance belonged to one of these 9 emotional classes: amusement, awe, contentment, and excitement as positive emotions, while we have anger, fear, disgust, and sadness as negative emotions. In addition, a ninth class named something else was considered to express having no particular emotions or an additional feeling not listed. Efforts were made to include at least one negative and one positive emotional response for each artwork. Partitions of 85%, 5%, and 10% were considered for train, validation, and test splits, respectively.

3. Results

3.1. Metrics and Implementation Details

We employ the following captioning metrics: BLEU [43], METEOR [44], ROUGE [45], and CIDEr [30]. For the cross-entropy training and SCST fine-tuning stages, batch sizes of 50 and 30 were considered, respectively. In the captioning model's training phase, the byte pair encoding (BPE) [46] method was utilized to represent the words. Sinusoidal positional encodings [11] were employed to represent word positions. Three layers of both encoders and decoders were utilized, each with a dimensionality of 512. The emotional grounding module was a single layer linear feed-forward network where the input was a one-hot vector of size 9 corresponding to our emotional classes, and the output was an embedding vector of size 10. The feed-forward dimensionality was 2058 in EGNemesis and 2048 in Nemesis with a head-number of 8. The memory size was set to 40. In addition, a dropout of 0.1 was applied to each sub-layer.

Adam [47] optimizer has been employed in all experiments along with a beam size of 5. In the cross-entropy training, the typical transformer learning rate scheduling strategy [11] has been utilized with a 10,000 iteration warmup. While in the SCST fine-tuning phase, a fixed learning rate of 5×10^{-6} has been considered, with a momentum λ of 0.999 for the teacher model.

3.2. Ablation Study

3.2.1. Visual Encoder

First, we evaluated the role of employing different models as the visual encoder to extract visual features. We experimented with detecting the object bounding boxes using a Faster R-CNN [48] model using a ResNet-101 [38] module pre-trained on the Visual Genome dataset [49]. Additionally, extracting grid-based features via CLIP [34] and region-based features via BLIP [35] was examined. In particular, the CLIP-RN50 $\times$ 16 variant was utilized, which is based on an EfficientNet-style [50] scaling. On the other hand, we utilized the $BLIP_{ViT-L/16}$ variant, which is based on the vision transformer [51] approach.

As can be observed in Tables 1 and 2, the best performance was achieved by the teacher model of Nemesis utilizing CLIP-RN50 $\times$ 16 as the visual encoder; hence, we will be referring to this exact configuration when mentioning the Nemesis. For the EGNemesis, the best performance was achieved by the student model utilizing $BLIP_{ViT-L/16}$; therefore, this particular configuration will be referred to as the EGNemesis in the following sections. As mentioned previously, the student and teacher models had equal sizes. In addition, the student model experiencing a higher number of parameter updates provided the opportunity to outperform the teacher model.

Table 1. Performance of the teacher model utilizing different visual encoders for both the Nemesis and EGNemesis models. (B: BLEU, M: METEOR, R: ROUGE, C: CIDEr).

Model	**Visual Encoder**	**Teacher Model**						
		B-1	**B-2**	**B-3**	**B-4**	**M**	**R**	**C**
Nemesis	Faster R-CNN	0.503	0.277	0.154	0.089	0.141	0.278	0.093
	CLIP-RN50 $\times$ 16	0.539	0.311	0.178	0.106	0.141	0.294	0.130
	$BLIP_{ViT-L/16}$	0.526	0.304	0.175	0.105	0.138	0.291	0.127
EGNemesis	Faster R-CNN	0.458	0.233	0.121	0.066	0.118	0.242	0.070
	CLIP-RN50 $\times$ 16	0.475	0.252	0.136	0.076	0.124	0.254	0.095
	$BLIP_{ViT-L/16}$	0.470	0.252	0.137	0.077	0.123	0.255	0.099

Table 2. Performance of the student model utilizing different visual encoders for both the Nemesis and EGNemesis models. (B: BLEU, M: METEOR, R: ROUGE, C: CIDEr).

Model	**Visual Encoder**	**Student Model**						
		B-1	**B-2**	**B-3**	**B-4**	**M**	**R**	**C**
Nemesis	Faster R-CNN	0.498	0.273	0.151	0.086	0.130	0.276	0.087
	CLIP-RN50 $\times$ 16	0.532	0.304	0.172	0.102	0.137	0.290	0.120
	$BLIP_{ViT-L/16}$	0.509	0.290	0.165	0.097	0.137	0.281	0.116
EGNemesis	Faster R-CNN	0.455	0.233	0.122	0.066	0.114	0.243	0.066
	CLIP-RN50 $\times$ 16	0.472	0.251	0.134	0.076	0.124	0.254	0.095
	$BLIP_{ViT-L/16}$	0.479	0.260	0.141	0.080	0.129	0.262	0.099

On the other hand, to experiment with handling the big data in various paradigms, these visual encoders have been utilized in two visual encoding modes. The CLIP and BLIP visual encoders have been incorporated in online visual encoding mode, where the visual feature extraction is carried out in real-time while training. On the other hand, the Faster R-CNN visual encoder has been utilized in offline visual encoding mode, where the features have been extracted before training the language models. These extracted features are stored on the disk and accessed during the training process. Data parallelism has been leveraged with the BLIP visual encoder in the XE training phase to reduce the training

time. Table 3 contains training times corresponding to each visual encoding mode for both training phases. However, the SCST fine-tuning was not carried out for the model trained on the Faster R-CNN encoder due to the incompetent results.

Table 3. Training time based on utilizing different visual encoding modes.

Training Phase	Visual Encoder	Encoding Mode	Data Parallelism	GPU Type	Time Per Epoch
XE	Faster R-CNN	Offline	-	NVIDIA P100	1 h
	CLIP-RN50 × 16	Online	-	NVIDIA V100	4 h
	$BLIP_{ViT-L/16}$	Online	✓	NVIDIA V100	1 h
SCST	Faster R-CNN	-	-	-	-
	CLIP-RN50 × 16	Online	-	NVIDIA V100	7 h
	$BLIP_{ViT-L/16}$	Online	-	NVIDIA V100	7 h

3.2.2. SCST Fine-Tuning

Table 4 shows the effect of the SCST fine-tuning stage, where the CIDEr-D metric was used as the reward function to encourage the model's generations that outperform the current test-time model following Equation (5). As is observable, the model's performance is boosted with respect to all utilized metrics in comparison with the same model after the cross-entropy training phase. The most significant improvements are related to the CIDEr and BLEU-1 scores. The BLEU-1 metric increased from 0.539 to 0.711 for the Nemesis, and from 0.479 to 0.700 for the EGNemesis. On the other hand, the CIDEr score was boosted from 0.130 to 0.219 for the Nemesis, and from 0.099 to 0.224 for the EGNemesis.

Table 4. The comparison of the results before and after applying the SCST fine-tuning.

Metric	*Nemesis*	$Nemesis_{SCST}$	*EGNemesis*	$EGNemesis_{SCST}$
BLEU-1	0.539	0.711	0.479	0.700
BLEU-2	0.311	0.406	0.260	0.403
BLEU-3	0.178	0.211	0.141	0.214
BLEU-4	0.106	0.113	0.080	0.115
METEOR	0.141	0.166	0.129	0.165
ROUGE-L	0.294	0.341	0.262	0.336
CIDEr	0.130	0.219	0.099	0.224

3.2.3. Image-to-Emotion Classifier

A comparison of the model's performance according to the utilization of different image-to-emotion classifiers can be found in Table 5. These classifiers are: (1) the ResNet-32 classifier pre-trained on the ImageNet dataset (IN), which gave the best performance in the previous work by Achlioptas et al. [19]. (2) The ResNet-50 classifier pre-trained on the Stylized-ImageNet dataset (SIN), which is our proposed classifier.

Table 5. Results with respect to the utilized image-to-emotion classifiers.

Metric	$EGNemesis_{IN}$	$EGNemesis_{SIN}$
BLEU-1	0.466	0.479
BLEU-2	0.251	0.260
BLEU-3	0.137	0.141
BLEU-4	0.077	0.080
METEOR	0.128	0.129
ROUGE-L	0.253	0.262
CIDEr	0.093	0.099

As is observable, the performance has improved by using our proposed classifier module in all the utilized metrics. For instance, the BLEU-1 score has improved from 0.466 to 0.479, and the CIDEr metric has improved from 0.093 to 0.099. This shows that the decrease in texture bias while increasing shape bias achieved a better performance in our auxiliary image-to-emotion classification task.

3.2.4. Emotional Grounding

The results with and without incorporating the extra emotional supervision signal are shown in Table 4. This signal is provided based on the emotional class indicated by the image-to-emotion classifier during the training time to keep the assessment fair.

As can be observed, the evaluation scores experience a decrease after emotional grounding; however, this degradation in evaluation metrics does not necessarily indicate a decrease in the quality of generated captions in our case. Most evaluation metrics return a higher score if the generated caption includes more words from the ground-truth captions or their synonyms, which is not the best way to assess generated captions in our subjective emotion-centric utterances. In fact, an increase in the diversity of the captions can result in a degradation of evaluation metrics. As shown in Figure 4, the generated caption of Nemesis for the first image is "it looks like a cold winter day". While the EGNemesis generated "this painting makes me feel nostalgic. it reminds me of my childhood" grounded in the Contentment emotional class, which is more emotionally rich according to human judgment. However, this emotionally grounded utterance will achieve a lower evaluation score since it does not contain the frequent words in the ground-truth captions, which are "cold" and "winter".

Figure 4. Examples of generated captions for unseen artworks. These samples include utterances from Nemesis model, and EGNemesis model along with the emotional class extracted by the image-to-emotion classifier, which has been utilized in the emotional grounding process. The descriptions contain various human-like emotional expressions, such as "reminds me of my childhood", "makes me feel nostalgic".

3.3. Comparison with the State-of-the-Art

3.3.1. Auxiliary Classification

The nine-way emotional classification problem is an extremely challenging task because of the subjectivity of emotions and diversity of emotional utterances. In a previous work, a user study was conducted to measure the accuracy of this classification task by humans [19]. This study consisted of three human experts attempting to guess the dominant emotional class based on a ground-truth utterance of ArtEmis, where they achieved an accuracy of only 61.2%. This depicts how challenging this task is, even based on human judgment. However, the BERT-based text-to-emotion classifier utilized in our model achieved a 64.8% accuracy, which is a surprising performance compared to the human results.

For the image-to-emotion classification task, which is arguably more difficult than the text-to-emotion classification, the ResNet-32 module pre-trained on ImageNet (IN) achieved a 60.2% accuracy, while the ResNet-50 module pre-trained on Stylized-ImageNet (SIN) achieved a 59.4% accuracy. However, the accuracy of this auxiliary task is not directly important to us. As mentioned earlier, Table 5 shows that our model performs better when utilizing the classifier trained on SIN because of the more diverse and shape-driven label predictions.

3.3.2. Emotion-Centric Image Captioning Task

We compared our proposed neural speaker to the best-performing emotion-centric image captioning models introduced by [19] on the ArtEmis dataset. These neural speakers include a captioning model inspired by meshed-memory transformer (M^2) [42] architecture, and an LSTM-based model inspired by "Show, Attend and Tell" (SAT) [9]. In addition, this comparison included the emotionally grounded variations of these models (i.e., M^2-EG and SATEG).

As can be observed in Table 6, our proposed model outperforms both the emotionally grounded and standard variations of M^2 and SAT speakers with respect to all incorporated metrics. This improvement is more notable in models after the SCST fine-tuning stage. The only exception is the EGNemesis, and the reason for this degradation has been elaborated on previously. As an example, Figure 5 shows some generated utterances from SATEG and EGNemesis models, where EGNemesis appears to generate more abstract, diverse, emotionally rich, and human-like captions.

Table 6. Comparison of state-of-the-art results and Nemesis after both cross-entropy training and SCST fine-tuning.

Metric	*SAT*	*SATEG*	M^2	M^2-*EG*	*Nemesis*	$Nemesis_{SCST}$	*EGNemesis*	$EGNemesis_{SCST}$
BLEU-1	0.536	0.520	0.507	0.511	0.539	0.711	0.479	0.700
BLEU-2	0.290	0.280	0.282	0.282	0.311	0.406	0.260	0.403
BLEU-3	0.155	0.146	0.159	0.154	0.178	0.211	0.141	0.241
BLEU-4	0.087	0.079	0.095	0.090	0.106	0.113	0.080	0.115
METEOR	0.142	0.134	0.140	0.137	0.141	0.166	0.129	0.165
ROUGE-L	0.297	0.294	0.280	0.286	0.294	0.341	0.262	0.336

EGNemesis: this painting makes me feel tired. the women look like they are working together.
[Something Else]

SATEG: the women are enjoying their time together.
[Contentment]

EGNemesis: the fish look like they are swimming in a storm.
[Excitement]

SATEG: I feel confused because i do not know what this is.
[Something Else]

EGNemesis: this painting makes me feel hungry for some fruit .
[Contentment]

SATEG: the colors are bright and the scene is very peaceful.
[Contentment]

EGNemesis: this man looks like he is up to no good.
[Fear]

SATEG: the man looks like he is about to cry.
[Sadness]

Figure 5. A comparison between the examples of generated captions by EGNemesis and SATEG models along with the emotional class extracted by the image-to-emotion classifier, which has been utilized in the emotional grounding process. It can be observed that the generated utterances by EGNemesis appear to be more abstract, human-like, and emotionally rich.

3.3.3. Limitations

The main challenge in the task of emotion-centric image captioning is the lack of a proper evaluation metric that aligns well with human judgment. Most of the evaluation metrics focus on comparing words in the generated captions with the words or synonyms in the reference captions, which is not the best approach for our diverse and subjective task. On the other hand, the generated captions are still far from including the ideal human-like properties to describe both emotionally and linguistically accurate emotional responses.

4. Conclusions

Neural speakers capable of producing affective utterances are an important step toward generating more engaging captions by provoking human emotions. As humans, emotions are a crucial part of expressing ourselves when we aim to describe different phenomena. Therefore, it is logical to expect the automatic image captioning process to consider this essential aspect of our perceptions. In this paper, we introduced Nemesis, a Neural Mean Teacher Learning-based Emotion-centric Speaker, an image captioning model capable of describing emotional responses to visual stimuli. We used the ArtEmis dataset to train our proposed neural speaker, the first large-scale dataset for affective image captioning containing 455K emotional descriptions of 80K artworks from WikiArt. We showed that incorporating a mean teacher learning-based approach followed by SCST-based fine-tuning, which utilizes extra emotional supervision signals, is a promising path toward generating more human-like emotion-centric descriptions. This was achieved by both experimenting with the utilization of different modules in the proposed pipeline and comparing it with the latest state-of-the-art methods.

Author Contributions: Conceptualization, A.Y. and K.P.; methodology, A.Y. and K.P.; software, A.Y.; validation, A.Y. and K.P.; formal analysis, A.Y.; investigation, A.Y.; resources, A.Y.; data curation, A.Y.; writing—original draft preparation, A.Y.; writing—review and editing, K.P.; visualization, A.Y.; supervision, K.P.; project administration, K.P.; funding acquisition, K.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets are publicly available as follows: ArtEmis dataset: https://www.artemisdataset.org/ (accessed on 21 February 2022); WikiArt dataset: https://github.com/cs-chan/ArtGAN/tree/master/WikiArt%20Dataset (accessed on 22 February 2022).

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

References

1. 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]
2. Jia-Yu, P.; Yang, H.-J.; Duygulu, P.; Faloutsos, C. Automatic image captioning. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No. 04TH8763), Taipei, Taiwan, 27–30 June 2004; Volume 3.
3. Ordonez, V.; Kulkarni, G.; Berg, T. Im2text: Describing images using 1 million captioned photographs. *Adv. Neural Inf. Process. Syst.* **2011**, *24*.
4. Yang, Y.; Teo, C.; Daume, H., III; Aloimonos, Y. Corpus-guided sentence generation of natural images. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, UK, 27–31 July 2011; pp. 444–454.
5. Gupta, A.; Verma, Y.; Jawahar, C. Choosing linguistics over vision to describe images. In Proceedings of the AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012; Volume 26, pp. 606–612.
6. Yao, B.Z.; Yang, X.; Lin, L.; Lee, M.W.; Zhu, S.-C. I2t: Image parsing to text description. *Proc. IEEE* **2010**, *98*, 1485–1508. [CrossRef]
7. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
8. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 652–663. [CrossRef] [PubMed]
9. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 2048–2057.

10. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3128–3137.
11. 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 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., 4–9 December 2017; 2017; 30.
12. Gan, C.; Gan, Z.; He, X.; Gao, J.; Deng, L. Stylenet: Generating attractive visual captions with styles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3137–3146.
13. Mathews, A.; Xie, L.; He, X. Semstyle: Learning to generate stylised image captions using unaligned text. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8591–8600.
14. Guo, L.; Liu, J.; Yao, P.; Li, J.; Lu, H. Mscap: Multi-style image captioning with unpaired stylized text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4204–4213.
15. Zhao, W.; Wu, X.; Zhang, X. Memcap: Memorizing style knowledge for image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12984–12992.
16. Shuster, K.; Humeau, S.; Hu, H.; Bordes, A.; Weston, J. Engaging image captioning via personality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12516–12526.
17. Lu, X.; Wang, B.; Zheng, X. Sound active attention framework for remote sensing image captioning. *IEEE* **2019**, *58*, 1985–2000. [CrossRef]
18. Wang, B.; Dong, G.; Zhao, Y.; Li, R.; Cao, Q.; Chao, Y. Non-uniform attention network for multi-modal sentiment analysis. In *International Conference on Multimedia Modeling*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 612–623.
19. Achlioptas, P.; Ovsjanikov, M.; Haydarov, K.; Elhoseiny, M.; Guibas, L.J. Artemis: Affective language for visual art. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11569–11579.
20. Laine, S.; Aila, T. Temporal ensembling for semi-supervised learning. *arXiv* **2016**, arXiv:1610.02242.
21. Tarvainen, A.; Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Adv. Neural Inf. Process.Syst.* **2017**, *30*.
22. Xu, G.; Liu, Z.; Li, X.; Loy, C.C. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 588–604.
23. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.025312.
24. Ba, J.; Caruana, R. Do deep nets really need to be deep? In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014.
25. Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **1992**, *8*, 229–256. [CrossRef]
26. Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; Murphy, K. Improved image captioning via policy gradient optimization of spider. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 873–881.
27. Sutton, R.S.; McAllester, D.; Singh, S.; Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In Proceedings of the 12th International Conference on Neural Information Processing Systems, Denver, CO, USA, 29 November–4 December 1999; 12.
28. Ranzato, M.; Chopra, S.; Auli, M.; Zaremba, W. Sequence level training with recurrent neural networks. *arXiv* **2015**, arXiv:1511.06732.
29. 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. 7008–7024.
30. 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.
31. Yang, L.; Shang, S.; Liu, Y.; Peng, Y.; He, L. Variational transformer: A framework beyond the trade-off between accuracy and diversity for image captioning. *arXiv* **2022**, arXiv:2205.14458.
32. Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; Mei, T. Boosting image captioning with attributes. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4894–4902.
33. Shen, S.; Li, L.H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; Keutzer, K. How much can CLIP benefit vision-and-language tasks? In Proceedings of the International Conference on Learning Representations. Virtual, 25–29 April 2022.
34. 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 International Conference on Machine Learning, PMLR, Virtual, 13–14 August 2021; pp. 8748–8763.
35. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv* **2022**, arXiv:2201.12086.
36. 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, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
37. 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.

38. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
39. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
40. Barraco, M.; Stefanini, M.; Cornia, M.; Cascianelli, S.; Baraldi, L.; Cucchiara, R. CaMEL: Mean Teacher Learning for Image Captioning. In Proceedings of the International Conference on Pattern Recognition, Montréal, QC, Canada, 21–25 August 2022; pp. 4087–4094.
41. Wang, Y.; Albrecht, C.M.; Zhu, X.X. Self-Supervised Vision Transformers for Joint SAR-Optical Representation Learning. In Proceedings of the IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 139–142.
42. MCornia; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer 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. 10578–10587.
43. 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, Stroudsburg, PA, USA, 7–12 July 2002; pp. 311–318.
44. Lavie, A.; Agarwal, A. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007; pp. 228–231.
45. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
46. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1715–1725.
47. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.
48. 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, 18–23 June 2018; pp. 6077–6086.
49. 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. Vision* **2017**, *123*, 32–73. [CrossRef]
50. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 6105–6114.
51. 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 16 $\times$ 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.