applied sciences

MDPI

Article

Convolutional Neural Network and Language Model-Based Sequential CT Image Captioning for Intracerebral Hemorrhage

Gi-Youn Kim [1], Byoung-Doo Oh [2], Chulho Kim [3] and Yu-Seop Kim [1,*]

[1] Department of Convergence Software, Hallym University, Chuncheon-si 24252, Republic of Korea; dekmj001@gmail.com
[2] Cerebrovascular Disease Research Center, Hallym University, Chuncheon-si 24252, Republic of Korea; iambd822@gmail.com
[3] Department of Neurology, Chuncheon Sacred Heart Hospital, Chuncheon-si 24252, Republic of Korea; gumdol52@hallym.or.kr
* Correspondence: yskim01@hallym.ac.kr

Abstract: Intracerebral hemorrhage is a severe problem where more than one-third of patients die within a month. In diagnosing intracranial hemorrhage, neuroimaging examinations are essential. As a result, the interpretation of neuroimaging becomes a crucial process in medical procedures. However, human-based image interpretation has inherent limitations, as it can only handle a restricted range of tasks. To address this, a study on medical image captioning has been conducted, but it primarily focused on single medical images. However, actual medical images often consist of continuous sequences, such as CT scans, making it challenging to directly apply existing studies. Therefore, this paper proposes a CT image captioning model that utilizes a 3D-CNN model and distilGPT-2. In this study, four combinations of 3D-CNN models and language models were compared and analyzed for their performance. Additionally, the impact of applying penalties to the loss function and adjusting penalty values during the training process was examined. The proposed CT image captioning model demonstrated a maximum BLEU score of 0.35 on the in-house dataset, and it was observed that the text generated by the model became more similar to human interpretations in medical image reports with the application of loss function penalties.

Keywords: medical image captioning; computed tomography; CNN; language model; natural language processing

Citation: Kim, G.-Y.; Oh, B.-D.; Kim, C.; Kim, Y.-S. Convolutional Neural Network and Language Model-Based Sequential CT Image Captioning for Intracerebral Hemorrhage. *Appl. Sci.* **2023**, *13*, 9665. https://doi.org/10.3390/app13179665

Academic Editors: Teen-Hang Meen, Chun-Yen Chang, Po-Lei Lee, Charles Tijus and Kuei-Shu Hsu

Received: 29 July 2023
Revised: 20 August 2023
Accepted: 24 August 2023
Published: 26 August 2023

1. Introduction

Intracerebral hemorrhage (ICH), an untreatable and severe form of brain hemorrhage, is a serious problem where one-third of patients die within a month, and even survivors may experience neurological complications [1,2]. Especially if rapid diagnosis and prompt treatment are not performed, the mortality rate of ICH patients can be increased [3]. Despite the seriousness of ICH, the incidence rate of ICH is steadily increasing. Bako et al. reported that the incidence rate of ICH increased by 11% over 15 years across the United States. They particularly highlighted the rising incidence of ICH among young economically-active and middle-aged populations in the United States, emphasizing the need for prevention strategies for ICH targeting [4]. In this situation, Rindler et al. investigated the impact and importance of neuroimaging examinations for ICH patient management [5]. This is because, for the prompt diagnosis and identification of the underlying cause of the ICH, neuroimaging examinations are essential. In particular, the results of neuroimaging examinations are used to prioritize patients and determine appropriate medical treatment. However, the individuals responsible for conducting these neuroimaging examinations and documenting the results are radiologists. In that situation, the Royal College of Radiologists stated that while the number of CT and MRI scans performed has increased by 7% annually,

there has only been a 4% annual increase in the number of radiologists [6]. This shortage of radiologists leads to delayed diagnosis and medical treatment.

In such circumstances, Rindler et al. highlighted that innovative technologies like automated ICH detection can be rapidly applied in the diagnostic stage of ICH [5]. In a similar context, Mohammed et al. stated that artificial intelligence techniques can analyze CT scans with high accuracy and speed, and it can be helpful for experts, radiologists, and patients [3]. This is why various medical image captioning studies have been conducted in the past. Using medical images as input, medical image captioning generates corresponding text. In the imageCLEFmedical Caption Task 2022, organized by Cross-Language Education and Function (CLEF), a medical image captioning challenge was held using datasets consisting of single CT images, MRI images, X-ray images, and their corresponding captions [7]. In this competition, Hajihosseini et al. utilized a ResNet-50-based multi-label classification model, treating words as individual labels, and achieved the highest performance with a BLEU score of 0.48 [8]. In the same challenge, Lebrat et al. utilized an encoder-to-decoder model for medical image captioning [9]. They used vision transformers and convolutional vision transformers as encoders and BERT, distilGPT2 models as decoders, with the method that sets the probability for certain tokens to 0 to address the n-gram repetition problem. Moreover, Selivanov et al. proposed the medical image captioning model, which combines the output vectors of the show-and-tell model and GPT-3 on the OPEN-I [10] and MIMIC-CXR [11] datasets [12].

Such studies have achieved success in single medical image captioning, but there is a limitation to applying such studies to diseases, which mainly use CT scans that represent 3D images in sequential 2D images. This is due to the characteristics of CT scans, which disperse 3D space information into sequential 2D images. Due to such limitations, it is difficult to find studies that have conducted captioning tasks for brain CT scans [13]. At the same time, unlike single medical image captioning, which considers the information from the single medical image, captioning for CT scans is a task that requires the model to generate text based on the information from all the images that make up the CT scan. However, CT scans are ready in most situations and can provide clues to determine the main cause of ICH even when there is little patient information [14,15]. Therefore, the need for a CT image captioning model that can deal with the characteristics of a CT scan is high.

At this time, 3D-CNNs (3D convolutional neural networks) are known for their ability to extract feature vectors from sequential images. As a result, they have been utilized in studies related to CT scans that involve artificial intelligence techniques. Perez et al. proposed an ICH prognosis prediction model using a custom 3D-CNN model and feed-forward network [16]. Neethi et al. proposed a stroke classification model for brain CT scans using a 3D-CNN model [17]. They achieved a 14.28% higher F1 score compared to the state-of-the-art stroke classification model at that time. Henderson et al. proposed a segmentation model for organs-at-risk (OARs) in the head and neck utilizing a 3D-CNN model [18]. They showed a performance that was on par with the state-of-the-art methods at that time, even with limited training data. Rani et al. proposed an automatic brain tumor detection model for CT and MRI images utilizing a 3D AlexNet model and a wireframe model [19]. This showed a high level of accuracy on the RSNA-MCCAI. These studies highlight the significant performance of 3D-CNN-based models in extracting meaningful information from CT scans.

Based on that, we propose an ICH-related CT scan captioning model that is based on the 3D-CNN model with a language model. Our goal is to generate corresponding reports of the given CT scan, which consists of normal and ICH CT scans. The proposed method utilizes the 3D-CNN model as an encoder and distilGPT-2, one of the language models, as a decoder. In addition, we present experimental results for combinations of models converted from ResNet-50 [20], EfficientNet-B5 [21], DenseNet-201 [22], and ConvNeXt-S [23] to 3D-CNN structures with distilGPT-2, as well as the result of utilizing penalty applied loss function to prevent the generation of specific sentences.

This paper is organized as follows. The proposed method, including utilized models and the loss functions with penalty for CT scan captioning, is outlined in Section 2. Section 3 analyzes the experimental results. This study is summarized in Section 4.

2. Methodology

The overall model structure is shown in Figure 1 and is composed of an encoder–decoder structure utilizing an end-to-end learning strategy known to be effective for caption generation tasks [24]. In the encoder–decoder model structure, the encoder transforms the input into a fixed-dimensional feature vector. The decoder receives this feature vector and generates the corresponding output. In this study, the input to the encoder is the CT scan, and the output is a fixed-dimensional feature vector corresponding to that CT scan. The decoder is trained to generate sentences corresponding to the feature vector from the encoder and employs the following strategies: cross-attention and teacher forcing. Cross-attention uses query, key, and value, where the values for the elements are from two different embeddings in the attention mechanism. In this case, the value of the query is the input text from Figure 1, while the key and value are the feature vectors from 3D-CNN in Figure 1. Teacher forcing is a training strategy in which the model is forced to learn the token that comes after the current time step's token. For example, if the input "<sos> I go to school" is provided, the model is trained to output "I go to school <eos>". This strategy addresses the issue in the training stage where a wrong prediction at an earlier time step leads to subsequent time steps making incorrect predictions during the training process. The end-to-end learning strategy utilizes the same loss function for both the encoder and decoder during the training process.

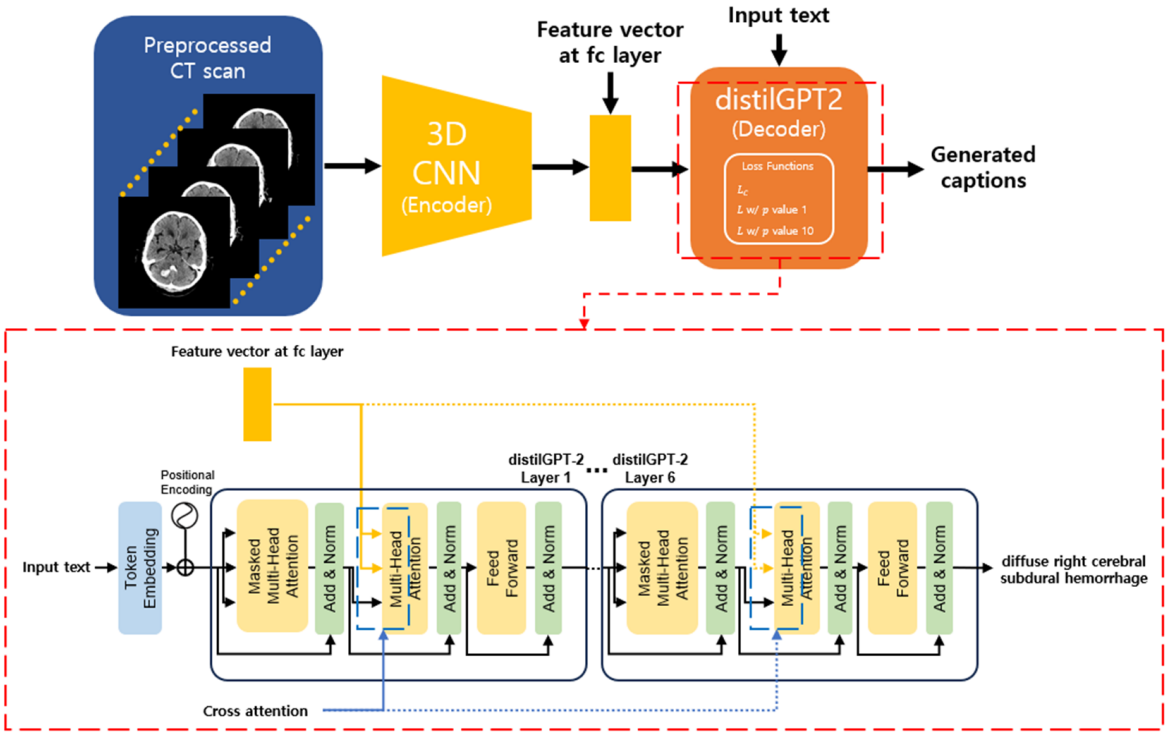
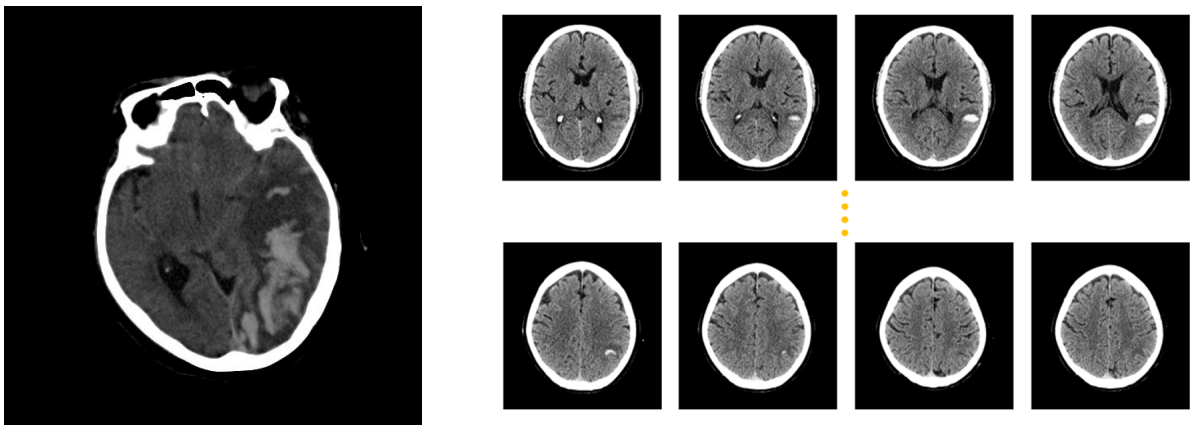


Figure 1. Proposed method. CT scan preprocessing process is described in Section 3.2.2. The 3D-CNN is one of the versions of 3D augmented models in the list of ResNet-50, EfficientNet-B5, DenseNet-201, and ConvNeXt-S. Loss functions in the distilGPT2 block are described at the bottom of page 4.

2.1. Encoder

For the encoder, ResNet-50, EfficientNet-B5, DenseNet-201, and ConvNeXt-S models were selected. Then, 3D augmentations were used by converting the 2D-CNN structure of these models into 3D-CNN [25]. During this process, the number of layers and model

structure for each model were preserved as the same as the existing 2D-CNN models, thus preserving their characteristics. While 2D-CNN models only consider information from a single image when extracting feature vectors, 3D-CNN models have the characteristic of considering spatial information from an image sequence. This allows the model to include information related to location and size changes of the hemorrhagic region in the feature vector, as seen in Figure 2b. Each 3D-CNN model utilized the converted pretrained weights from the corresponding 2D-CNN model's pretrained weights on ImageNet [25]. This is based on the assumption that the images surrounding a specific image in a sequential image are composed of similar information. Accordingly, the pretrained weights of the 2D-CNN model were divided by the kernel depth of the 3D-CNN model and used as the weights for the 3D-CNN model.



(**a**) (**b**)

Figure 2. Example of medical images. (**a**) Single 2D image from brain CT scan in the in-house data; (**b**) Sequential 2D images from brain CT scan in the in-house data. Images order in (**b**) are left to right and top to bottom.

2.2. Decoder

For the decoder, distilGPT2 was selected. DistilGPT2 is a compressed version of GPT-2 with 6 layers, a hidden layer size of 768, 12 heads, and 82 million parameters. This considers the overall size of the model parameters, as large models can have longer inference times [26]. A longer inference time does not align well with the considerations in this study where rapid diagnosis is considered, and it may even potentially hinder the diagnostic process. DistilGPT2, in particular, is more than twice as fast as GPT-2 on average [27]. In this paper, the output layer of distilGPT2 was used as-is since it can represent most of the medical terms.

2.3. Penalty Applied Loss Function

In this paper, the loss function used is sparse categorical cross entropy (SCCE). However, despite using SCCE, the trained model sometimes generated sentences typically found in normal CT scans ("unremarkable finding of brain parenchyma and cerebrospinal fluid space") when dealing with ICH CT scans. Here, we will refer to this as None-ICH Text. To prevent this, a penalty-applied loss function is proposed, which applies a penalty when the model generates a None-ICH Text for ICH CT scans and is implemented in the model training. Equation (1) represents the categorical cross entropy (l) loss function.

$$l = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c} t_{ij}\log\left(y_{ij}\right). \quad (1)$$

In this case, n represents the number of training samples, c is the number of classes, y is the model prediction, and t is the ground truth. Equation (2) is the loss function L that applies a penalty when the input is an ICH CT scan, but the output is a None-ICH Text.

$$L = \begin{cases} l_i + p, & if\ r_i = 1 \\ l_i, & if\ r_i = 0 \end{cases}. \quad (2)$$

In this case, l_i is the i-th loss value calculated from Equation (1), and p is the penalty value we determined. r_i is 1 if the i-th input is ICH, and the output is a None-ICH Text and 0 otherwise.

2.4. Text Generation Strategy

For the text generation strategy, Greedy Search, Beam Search, and Top-k Sampling were used. These text generation strategies are used during the decoding process, which transforms the output vector from the decoder readable by humans. Greedy Search selects the word with the highest probability at each time step from the output vector of the language model. Beam Search maintains k sequences at each time step and finally selects the sequence with the highest probability. Top-k Sampling determines the k most probable next words, redistributes the probabilities among them, and then selects a word. The hyperparameter k was set to 3.

3. Experiment Results

3.1. Experimental Setup

The experiments for this study were conducted using 8 NVIDIA A100 GPUs (Nvidia Corporation, Santa Clara, CA, USA) provided by the HPC-AI infrastructure at the Supercomputing Center (https://cwww.gist.ac.kr/scent/, accessed on 21 July 2023) operated by the Gwangju Institute of Science and Technology (GIST). For the training hyperparameters, we used the Adam optimizer with a learning rate of 0.001, a batch size of 8, and early stopping with 15 patience.

3.2. Dataset

The data used in this study consist of the CT scan and the corresponding radiologist report from 35,511 people, which were collected from 2012 to 2020 at Hallym University Sacred Heart Hospital (https://hallym.hallym.or.kr/eng/, accessed on 22 July 2023) and Hallym University Chuncheon Sacred Heart Hospital (https://chuncheon.hallym.or.kr/eng/, accessed on 22 July 2023). During the collection, we used a 64-slice Sensation 64 or a 128-slice Somatotom Definition Flash, multidetector row CT scanner (Siemens Healthcare, Forchheim, Germany). The CT scanners were standardized as follows: slice thickness, 3 mm; tube voltage, 120 kVp; field of view, 250 × 250 mm; standardized window level and width, 80/35. The overall data have a 7:3 ratio of normal and ICH CT scans. The CT scan was saved in DICOM format, which was extracted as a sequential PNG format to use as input to the model. Data augmentation was not performed to prevent the mismatch of the spatial information inherent in CT scans with the corresponding report. For example, if the original CT scan had a hemorrhage on the left side, the location of the hemorrhage may move to the right side, which will cause a mismatch with the corresponding report, which possibly describes the location of the hemorrhage on the left side of the brain. This highlights the potential inconsistencies that can arise when spatial information is altered through data augmentation. Subsequently, the entire dataset was divided into train, validation, and test sets with an 8:1:1 ratio.

3.2.1. Image Caption

Table 1 shows examples of radiologist reports for normal and ICH CT scans. These are present for each one of the CT scans. At this time, some of the reports contain Korean. However, the included Korean mainly consists of content that does not affect the interpretation, such as conjunctions or adverbs. Therefore, preprocessing was performed by removing

Korean, converting it to lowercase, and removing special characters. The maximum token length of the report was limited to 129, and padding was added if the report was shorter. Subsequently, the report was tokenized using the tokenizer used by the distilGPT2 decoder to form the input text in Figure 1.

Table 1. Caption examples. Normal column shows one of the radiologist's reports on the normal CT scan. ICH column shows one of the radiologist's reports on the ICH CT scan.

Normal	ICH
Unremarkable finding of brain parenchyma and cerebrospinal fluid space	Left frontal subcortical intracerebral hemorrhage with surrounding edema

3.2.2. CT Scan

The CT scan data for each patient are shown in Figure 2b. Figure 2b presents a portion of the ICH CT scan used in this study. Unlike Figure 2a, which is represented as a single image, Figure 2b is represented as a sequence of multiple images. Specifically, Figure 2b shows the hemorrhagic area in the brain gradually increasing from the top and moving from left to right. In the bottom images, the opposite is observed. This characteristic results from representing a 3D space as a 2D image.

Preprocessing methods such as image normalization and size adjustments, used in existing image captioning studies, were used and supplemented by adding a step to compose the entire sequence of images into a single 3D image, as shown in Figure 3. In this case, the sequences of CT images were unique for each patient, and the number of images in each patient's sequence varied. According to this, the average number of images per CT scan was checked, and the CT scan's average image count was approximately 47, with a median value of 49. The maximum number of images in a single 3D image was set to 64. If the number of images was insufficient, post-padding with a black image was carried out, and if exceeded, images from the 65th onwards were excluded.

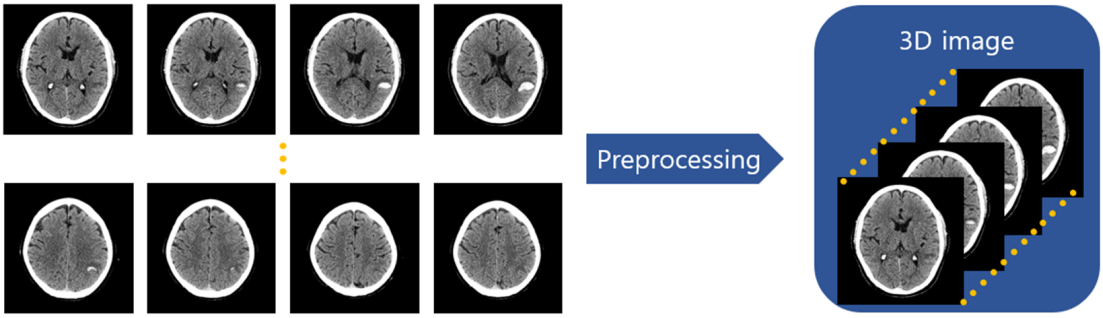


Figure 3. Preprocessing of sequential 2D CT scan images into 3D images. In the preprocessing, image normalizations and image resize are applied to the CT scan images and the images are stacked into one 3D image following the original order.

3.3. Evaluation Metric

For the evaluation of the experimental results, we used the nlgeval library [28] to calculate the BLEU (Bilingual Evaluation Understudy) score [29], METEOR (Metric for Evaluation of Translation with Explicit Ordering) score [30], and ROUGE-L (Recall Oriented Understudy of Gisting Evaluation) score [31]. Additionally, we used cosine similarity between sentence vectors utilizing embeddings from ClinicalBERT [32]. The BLEU score calculates how many n-grams overlap between a human-generated reference sentence and a model-generated sentence in machine translation. The METEOR score is similar to the BLEU score but considers recall in addition to precision, which the BLEU score only focuses on. Precision is the ratio of the number of words that overlap between the model-generated sentence and the reference sentence to the total number of words in the model-generated

sentence. Recall is the ratio of the number of words that overlap between the reference sentence and the number of words in the reference sentence. The ROUGE-L score is a type of ROUGE-N score that emphasizes n-gram recall between the model-generated and reference sentences, and it utilizes the longest common subsequence. The subsequence in this context does not have to be contiguous.

ClinicalBERT is a medical language model that pretrained the BERT model with a 1.2-billion-word corpus consisting of various disease-related terms and finetunes it with an EHR corpus of more than 3 million patients. In this paper, we embedded both the model-generated sentences and the reference sentences using ClinicalBERT and then calculated the cosine similarity between each resulting embedded sentence vector. Cosine similarity is calculated using the cosine angle between two vectors, where the value is 1 when the directions of the two vectors are identical, 0 when their angle is 90 degrees, and -1 when their direction is opposite at 180 degrees.

3.4. Experiment Result with All Test Data

Table 2 presents experimental results using all test data with a ratio of seven normal and three ICH. The experimental results using the loss function based on Equation (2) are shown in Table 2 as L w/p value 1 and L w/p value 10. Within the group using the l loss function, EfficientNet-B5+B scored the highest performance across all metrics, showing the overall highest score among all experimental results in Table 2. In the same model, the cosine similarity using ClinicalBERT was 0.78, confirming that the similarity between the generated sentences and the actual report was high. For the ResNet-50 model, except for the Top-k Sampling generation strategy, the scores were higher when using L w/p value 1 and L w/p value 10 compared to using l, and they showed the highest scores among the models using the same loss function group. This indicates that the penalty-applied loss functions had a positive impact. The reported METEOR and ROUGE-L scores from Hajihosseini et al., which placed first in the imageCLEFmedical Caption Task 2022 for single medical image captioning, were 0.09 and 0.14, respectively, while the reported BLEU score from Lebrat el al., which ranked third in the same competition, was 0.31 [8,9]. These results show that the proposed method in this study demonstrates a certain level of usability utilizing 3D-CNN models to consider the spatial information, even for a more challenging task, which should consider the spatial information compared to a single medical image captioning task.

Table 2. Experimental result with all test data. G denotes Greedy Search, B denotes Beam Search, and T denotes Top-k Sampling. Highest score for each metric within the loss function is highlighted with bold text.

Loss Function	Models	BLEU *	METEOR	ROUGE-L	Cosine Similarity
l	ResNet-50 + G	0.26	0.19	0.51	0.75
	ResNet-50 + B	0.31	0.21	0.54	0.77
	ResNet-50 + T	0.08	0.06	0.13	0.58
	EfficientNet-B5 + G	0.28	0.20	0.52	0.76
	EfficientNet-B5 + B	**0.35**	**0.23**	**0.56**	**0.78**
	EfficientNet-B5 + T	0.09	0.08	0.17	0.62
	DenseNet-201 + G	0.28	0.20	0.53	0.76
	DenseNet-201 + B	0.23	0.16	0.44	0.73
	DenseNet-201 + T	0.07	0.06	0.12	0.57
	ConvNeXt-S + G	0.26	0.19	0.51	0.75
	ConvNeXt-S + B	0.25	0.19	0.51	0.75
	ConvNeXt-S + T	0.06	0.05	0.12	0.57

Table 2. *Cont.*

Loss Function	Models	BLEU *	METEOR	ROUGE-L	Cosine Similarity
L w/p value 1	ResNet-50 + G	0.29	0.19	0.51	0.76
	ResNet-50 + B	**0.33**	**0.22**	**0.54**	**0.77**
	ResNet-50 + T	0.07	0.06	0.12	0.58
	EfficientNet-B5 + G	0.24	0.17	0.44	0.73
	EfficientNet-B5 + B	0.21	0.15	0.34	0.68
	EfficientNet-B5 + T	0.09	0.08	0.15	0.60
	DenseNet-201 + G	0.27	0.20	0.52	0.76
	DenseNet-201 + B	0.19	0.12	0.33	0.66
	DenseNet-201 + T	0.07	0.06	0.13	0.59
	ConvNeXt-S + G	0.26	0.19	0.51	0.75
	ConvNeXt-S + B	0.25	0.19	0.51	0.75
	ConvNeXt-S + T	0.06	0.05	0.11	0.56
L w/p value 10	ResNet-50 + G	0.28	0.2	0.52	0.76
	ResNet-50 + B	**0.31**	**0.22**	**0.54**	**0.77**
	ResNet-50 + T	0.08	0.07	0.14	0.59
	EfficientNet-B5 + G	0.27	0.17	0.45	0.73
	EfficientNet-B5 + B	0.20	0.15	0.34	0.67
	EfficientNet-B5 + T	0.11	0.09	0.18	0.63
	DenseNet-201 + G	0.28	0.20	0.53	**0.77**
	DenseNet-201 + B	0.24	0.16	0.44	0.75
	DenseNet-201 + T	0.09	0.07	0.16	0.6
	ConvNeXt-S + G	0.06	0.1	0.16	0.56
	ConvNeXt-S + B	0.06	0.11	0.14	0.53
	ConvNeXt-S + T	0.03	0.01	0.11	0.56

* Mean score of BLEU-1, 2, 3, and 4 score.

However, as can be seen in Figure 4a,b, there are significant differences in the report that makes up the radiology reports of the images within the test dataset. Figure 4a shows a radiology report of a normal CT scan, which is mainly composed of similar sentences and includes many sentences that are the same or similar to None-ICH texts. In contrast, Figure 4b shows a radiology report of an ICH CT scan, which contains various information. The experimental results presented in Table 2 utilized test data where the proportion of normal CT scans was around 70%, similar to the training data. This suggests that the model's performance could be inaccurately evaluated if it primarily generates sentences similar to those found in Figure 4a or None-ICH texts. Furthermore, since one of the goals of this study is to generate reports for the patient's CT scans, it is essential to evaluate the performance of the model, specifically for the ICH CT scans. Therefore, we conducted experiments using only the test data of ICH CT scans.

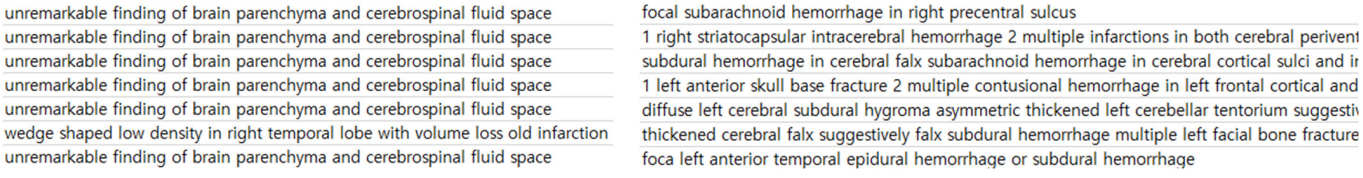


Figure 4. Example of radiologist reports within the whole test dataset. (**a**) Radiologist's reports on normal CT scan; (**b**) radiologist's reports on ICH CT scan. Unlike radiologist reports in (**b**), radiologist reports in (**a**) consist of mostly similar texts.

3.5. Experiment Result with the Test Data Consisted of ICH CT Scan

Table 3 presents the experimental results using test data composed solely of ICH CT scans. In the imageCLEFmedical Caption Task 2022, the highest reported METEOR score was 0.09, while the ROUGE-L score was 0.20 [7]. Some of the results in Table 3

exhibit similar or higher scores. This demonstrates that even when evaluating the model's performance using only the ICH CT scans, the proposed method in this paper provides a certain level of usability.

Table 3. Experimental results with test data consisted of an ICH CT scan. G denotes Greedy Search, B denotes Beam Search, T denotes Top-k Sampling. Highest score for each metric within the loss function is highlighted with bold text.

Loss Function	Models	BLEU *	METEOR	ROUGE-L	Cosine Similarity
l	ResNet-50 + G	0.01	0.03	0.08	0.51
	ResNet-50 + B	0.09	0.08	0.20	0.61
	ResNet-50 + T	0.04	0.05	0.09	0.59
	EfficientNet-B5 + G	0.06	0.09	0.16	0.58
	EfficientNet-B5 + B	**0.16**	0.13	**0.28**	**0.63**
	EfficientNet-B5 + T	0.09	0.10	0.17	**0.63**
	DenseNet-201 + G	0.06	0.09	0.09	0.56
	DenseNet-201 + B	0.12	**0.16**	0.10	0.62
	DenseNet-201 + T	0.08	0.10	0.07	0.60
	ConvNeXt-S + G	0.01	0.02	0.06	0.53
	ConvNeXt-S + B	0.01	0.02	0.06	0.53
	ConvNeXt-S + T	0.03	0.03	0.07	0.55
L w/p value 1	ResNet-50 + G	0.15	0.05	0.11	0.58
	ResNet-50 + B	0.14	0.10	0.22	0.62
	ResNet-50 + T	0.04	0.06	0.09	0.60
	EfficientNet-B5 + G	0.07	0.10	0.17	0.62
	EfficientNet-B5 + B	**0.17**	**0.17**	**0.30**	**0.68**
	EfficientNet-B5 + T	0.10	0.11	0.18	0.65
	DenseNet-201 + G	0.04	0.08	0.13	0.54
	DenseNet-201 + B	0.13	0.15	0.27	0.62
	DenseNet-201 + T	0.07	0.10	0.16	0.60
	ConvNeXt-S + G	0.01	0.02	0.06	0.53
	ConvNeXt-S + B	0.01	0.02	0.06	0.53
	ConvNeXt-S + T	0.04	0.05	0.09	0.58
L w/p value 10	ResNet-50 + G	0.03	0.05	0.11	0.57
	ResNet-50 + B	0.12	0.10	0.24	0.63
	ResNet-50 + T	0.04	0.06	0.09	0.60
	EfficientNet-B5 + G	0.07	0.08	0.16	0.61
	EfficientNet-B5 + B	**0.18**	**0.14**	**0.29**	**0.66**
	EfficientNet-B5 + T	0.11	0.10	0.18	0.65
	DenseNet-201 + G	0.06	0.09	0.16	0.58
	DenseNet-201 + B	0.15	**0.14**	0.28	0.63
	DenseNet-201 + T	0.08	0.10	0.16	0.63
	ConvNeXt-S + G	0.01	0.05	0.05	0.52
	ConvNeXt-S + B	0.02	0.07	0.07	0.51
	ConvNeXt-S + T	0.03	0.08	0.08	0.57

* Mean score of BLEU-1, 2, 3, and 4 score.

Additionally, an important observation from the results in Table 3 is the impact of the penalty-applied loss functions on the model's performance. The goal of the penalty-applied loss functions was to prevent the model from generating None-ICH text to the ICH CT scan. As seen in Table 2, when EfficientNet-B5 utilized L w/p value 1 and L w/p value 10 for model training, the scores were lower compared to when using l. However, Table 3 shows an improvement in the scores when EfficientNet-B5 utilized L w/p value 1 instead of l, and achieved the highest scores among the loss functions on the ICH CT scans. Notably, in Table 3, the EfficientNet-B5 model, which recorded the highest score among the models using l in Table 2, showed the highest scores across all three loss functions (l, L w/p value 1 and L w/p value 10) when evaluated on the only ICH CT scan. This suggests that utilizing penalty-applied loss functions showed a certain degree of effectiveness for

the EfficientNet-B5 model. Furthermore, the effect of penalty-applied loss functions is more shown in the ResNet-50 model. Comparing the scores of the ResNet-50 model between l and L w/p value 1 in Table 3, we observed a large improvement in BLEU scores for Greedy Search and Beam Search strategy, as well as slight score improvements for other metrics when l was utilized. The effect of penalty-applied loss functions can also be observed in the model-generated text shown in Table 4.

Table 4. EfficientNet-B5 + B generated texts with all test data. Top 2 rows are from the ICH CT scan, and bottom 2 rows are from the normal CT scan. Columns except reference text show a change in the generated text with the change in loss function and the penalty value. Effect of the loss function with penalty is shown in each row as the additional information in the generated text increases while it brings the generated text closer to the reference text. Overlapping tokens are highlighted with the colors.

Reference Text	l	L w/p Value 1	L w/p Value 10
diffuse left cerebral subdural hemorrhage midline shifting focal lacunar infarction in left internal capsule	diffuse right cerebral subdural hemorrhage	diffuse left cerebral subdural hemorrhage midline shifting subdural hemorrhage along the cerebral subarachnoid hemorrhage in right sylvian cistern	diffuse right cerebral subdural hemorrhage midline shifting multiple old lacunar infarctions in both basal ganglia and right thalamus
sdh, right cerebral hemisphere small vessel disease with lacunar infarctions otherwise no demonstrable abnormal finding	sdh in right fronto parietal otherwise no demonstrable abnormal finding	sdh, left fronto temporo parietal otherwise no demonstrable abnormal finding	sdh, right cerebral hemisphere small vessel disease with lacunar infarctions otherwise no demonstrable abnormal finding
unremarkable finding of brain parenchyma and cerebrospinal fluid space	unremarkable finding of brain parenchyma and cerebrospinal fluid space	unremarkable finding of brain parenchyma and cerebrospinal fluid space	unremarkable finding of brain parenchyma and cerebrospinal fluid space
unremarkable finding of brain parenchyma and cerebrospinal fluid space	unremarkable finding of brain parenchyma and cerebrospinal fluid space	diffuse brain atrophy	diffuse brain atrophy

There are also some noteworthy points in the results of the DenseNet-201 model. In Table 2, the scores of DenseNet-201 + G and DenseNet-210 + B models did not show the highest performance compared to other models, but they recorded scores close to the models with the highest scores within their loss function groups. Furthermore, in Table 3, the DenseNet-210 + B model demonstrated scores that were either close to, equal to, or higher than the models with the highest scores within their loss function groups in some metrics.

3.6. Examples of Generated Text

Table 4 displays the text generated by the EfficientNet-B5 + B model, which achieved the highest scores in Tables 2 and 3. For both the first and second rows of the ICH CT scan, we can see that applying L w/p value 1 and L w/p value 10 to the model helps generate sentences closer to the reference texts compared to applying l. Simultaneously, the texts generated by the models using L w/p value 1 and L w/p value 10 provide more detailed information than the model that applied only l for both ICH CT scans. For example, in the case of the first row, the model applying l does not provide information on "midline shifting focal lacunar infarction in the left internal capsule." However, L w/p value 1 includes the information of "midline shifting," and L w/p value 10 incorporates "lacunar

infarctions". Additionally, looking at the text in the second row, we can see that the text from the model with L w/p value 10 applied is identical to the reference text, confirming that the application of L w/p value 1 and L w/p value 10 had a positive effect.

The spatial information consideration of 3D-CNN is indeed evident in Table 4. In the first row of the reference text in Table 4, we can observe information regarding "diffuse subdural hemorrhage." Subdural hemorrhage refers to bleeding between the dura mater and the arachnoid membrane, and "diffuse subdural hemorrhage" signifies hemorrhage that spreads widely between these layers. To determine whether hemorrhage is diffuse or not, it is necessary to examine the spatial information in the CT scan. In this regard, when we look at the model-generated text in the first row of Table 4, we can see that the information "diffuse" is included regardless of the penalty value applied to the loss function. At the same time, midline shifting indicates that the brain's central line has moved. This can occur when there is insufficient space for the brain due to bleeding. However, for the model to confirm that the brain's central line has shifted, it would need to determine where the brain's central line should be. However, the brain central line may not be consistent across patients in the resulting CT scan due to factors such as the patient's position or head shape during the CT imaging process. Therefore, the model needs to determine the normal position of the brain's central line and when the shifting has occurred based on spatial information about where the brain is located. In models trained with a penalty-based loss function, you can observe information about midline shifting. This suggests that the proposed method handled this information effectively.

Indeed, while applying L w/p value 1 and L w/p value 10 did not affect the generated text of the model for the normal CT scan in the third row, it did impact the fourth row's normal CT scan, resulting in the generation of incorrect text. This may be related to the observation that as the p-value increases, additional ICH-related information is included in the model-generated text for the ICH CT scan. Therefore, finding an appropriate p-value to apply to the loss function L might help address this issue in the future. Additionally, in Table 4, it can be observed that the model sometimes fails to correctly identify the location of ICH-related information. This is likely due to the difficulty of accurately pinpointing the specific location in CT scans since they are composed of black and white images. In the future, providing additional information regarding the relevant locations might help improve the model's performance.

4. Conclusions

This study proposed a CT scan captioning model for ICH and introduced penalty applied loss functions L w/p value 1 and L w/p value 10 to control the bias during the model training process. The captioning model for ICH CT scans was structured as an encoder–decoder using a 3D-CNN model and a distilGPT2, demonstrating a certain level of usability compared to previous studies. Furthermore, the impact of L w/p value 1 and L w/p value 10 was examined during the model training process, and it was observed that they assisted the model in including more ICH-related information in the generated texts.

However, this study involves a captioning task on the results of 2D images representing 3D spaces captured in a dispersed manner, making the captioning task more challenging than conventional single medical image captioning. Simultaneously, it is difficult to find publicly-available ICH CT scan captioning task data and related prior studies. As a result, this study was conducted using an in-house dataset. Therefore, there are limitations to directly comparing the results of this study to those of other studies. In the future, when related gold standard datasets become publicly available, these limitations can be addressed through additional experiments. Furthermore, to the best of our knowledge, studies on 3D medical image captioning, such as the one presented in this paper, are challenging to find. Therefore, we referenced the scores reported in the ImageCLEFmedical Caption task 2022, which focused on the 2D medical image captioning task, to determine the usability of our proposed method. Because of this, when interpreting the performance of the proposed method in this paper, it is essential to take this into account.

Additionally, we proposed applying a penalty value to the SCCE loss function and observed its effects. However, there may be a more effective way to incorporate penalties into the loss function. Experiments with changing the loss function require the model's training and evaluation process, which is a time-consuming task. Therefore, we conducted this experiment with p-values of 1 and 10, which can be remedied through various additional experiments in the future.

In future studies, we will conduct more and various experiments to determine an appropriate p-value to guide the model, not to generate ICH-related text for the normal CT scan with the additional experiments concerning the loss functions.

Author Contributions: Conceptualization, Y.-S.K. and C.K.; methodology, G.-Y.K. and B.-D.O.; formal analysis, G.-Y.K. and B.-D.O.; resources, Y.-S.K.; data curation, C.K.; writing—original draft preparation, G.-Y.K.; writing—review and editing, Y.-S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1A5A8019303), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO. 2021-0-02068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], and Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) grant funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR21C0198).

Institutional Review Board Statement: This study was performed in accordance with the Declaration of Helsinki, and it was approved by the Institutional Review Board at Chuncheon Sacred Heart Hospital (IRB No. 2021-10-012).

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing is not applicable.

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

References

1. Cordonnier, C.; Demchuk, A.; Ziai, W.; Anderson, C.S. Intracerebral haemorrhage: Current approaches to acute management. *Lancet* **2018**, *392*, 1257–1268. [CrossRef] [PubMed]
2. Krishnamurthi, R.V.; Feigin, V.L.; Forouzanfar, M.H.; Mensah, G.A.; Connor, M.; Bennett, D.A.; Moran, A.E.; Sacco, R.L.; Anderson, L.M.; Truelsen, T.; et al. Global and regional burden of first-ever ischaemic and haemorrhagic stroke during 1990–2010: Findings from the Global Burden of Disease Study 2010. *Lancet Glob. Health* **2013**, *1*, e259–e281. [CrossRef] [PubMed]
3. Mohammed, B.A.; Senan, E.M.; Al-Mekhlafi, Z.G.; Rassem, T.H.; Makbol, N.M.; Alanazi, A.A.; Almurayziq, T.S.; Ghaleb, F.A.; Sallam, A.A. Multi-Method Diagnosis of CT Images for Rapid Detection of Intracranial Hemorrhages Based on Deep and Hybrid Learning. *Electronics* **2022**, *11*, 2460. [CrossRef]
4. Bako, A.T.; Pan, A.; Potter, T.; Tannous, J.; Johnson, C.; Baig, E.; Meeks, J.; Woo, D.; Vahidy, F.S. Contemporary trends in the nationwide incidence of primary intracerebral hemorrhage. *Stroke* **2022**, *53*, e70–e74. [CrossRef]
5. Rindler, R.S.; Allen, J.W.; Barrow, J.W.; Pradilla, G.; Barrow, D.L. Neuroimaging of Intracerebral Hemorrhage. *Neurosurgery* **2020**, *86*, E414–E423. [CrossRef] [PubMed]
6. London, T.R.C.O.R. Clinical Radiology UK Workforce Census 2020 Report. Available online: https://www.rcr.ac.uk/system/files/publication/field_publication_files/clinical-radiology-uk-workforce-census-2020-report.pdf (accessed on 21 July 2023).
7. Ionescu, B.; Müller, H.; Péteri, R.; Rückert, J.; Abacha, A.B.; de Herrera, A.G.S.; Friedrich, C.M.; Bloch, L.; Brüngel, R.; Idrissi-Yaghir, A.; et al. Overview of the ImageCLEF 2022: Multimedia Retrieval in Medical, Social Media and Nature Applications. In *International Conference of the Cross-Language Evaluation Forum for European Languages*; Springer International Publishing: Cham, Switzerland, 2022; pp. 541–564.
8. Hajihosseini, M.; Lotfollahi, Y.; Nobakhtian, M.; Javid, M.M.; Omidi, F.; Eetemadi, S. IUST_NLPLAB at ImageCLEFmedical Caption Tasks. In Proceedings of the Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022.
9. Lebrat, L.; Nicolson, A.; Santa Cruz, R.; Belous, G.; Koopman, B.; Dowling, J. CSIRO at ImageCLEFmedical Caption 2022. In Proceedings of the CLEF 2022: Conference and Labs of the Evaluation Forum, Bologna, Italy, 5–8 September 2022; Volume 3180, pp. 1455–1473.
10. Demner-Fushman, D.; Kohli, M.D.; Rosenman, M.B.; Shooshan, S.E.; Rodriguez, L.; Antani, S.; Thoma, G.R.; McDonald, C.J. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Inform. Assoc.* **2016**, *23*, 304–310. [CrossRef] [PubMed]

11. Johnson, A.E.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.-Y.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef] [PubMed]
12. Selivanov, A.; Rogov, O.Y.; Chesakov, D.; Shelmanov, A.; Fedulova, I.; Dylov, D.V. Medical image captioning via generative pretrained transformers. *Sci. Rep.* **2023**, *13*, 4171. [CrossRef]
13. Yang, S.; Ji, J.; Zhang, X.; Liu, Y.; Wang, Z. Weakly Guided Hierarchical Encoder-Decoder Network for Brain CT Report Generation. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, 9–12 December 2021; pp. 568–573.
14. Caceres, J.A.; Goldstein, J.N. Intracranial Hemorrhage. *Emerg. Med. Clin. N. Am.* **2012**, *30*, 771–794. [CrossRef] [PubMed]
15. Freeman, W.D.; Aguilar, M.I. Intracranial hemorrhage: Diagnosis and management. *Neurol. Clin.* **2012**, *30*, 211–240. [CrossRef]
16. Perez del Barrio, A.; Esteve Domínguez, A.S.; Menéndez Fernández-Miranda, P.; Sanz Bellón, P.; Rodríguez González, D.; Lloret Iglesias, L.; Marques Fraguela, E.; González Mandly, A.A.; Vega, J.A. A deep learning model for prognosis prediction after intracranial hemorrhage. *J. Neuroimaging* **2023**, *33*, 218–226. [CrossRef]
17. Neethi, A.S.; Niyas, S.; Kannath, S.K.; Mathew, J.; Anzar, A.M.; Rajan, J. Stroke classification from computed tomography scans using 3D convolutional neural network. *Biomed. Signal Process. Control* **2022**, *76*, 103720. [CrossRef]
18. Henderson, E.G.A.; Vasquez Osorio, E.M.; van Herk, M.; Green, A.F. Optimising a 3D convolutional neural network for head and neck computed tomography segmentation with limited training data. *Phys. Imaging Radiat. Oncol.* **2022**, *22*, 44–50. [CrossRef] [PubMed]
19. Rani, S.; Kumar, S.; Ghai, D.; Prasad, K. Automatic Detection of Brain Tumor from CT and MRI Images using Wireframe model and 3D Alex-Net. In Proceedings of the 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, 23–25 March 2022; pp. 1132–1138.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June 2016–1 July 2016; pp. 770–778.
21. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
22. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
23. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
24. Olivastri, S.; Singh, G.; Cuzzolin, F. End-to-end video captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Long Beach, CA, USA, 16–17 June 2019.
25. Solovyev, R.; Kalinin, A.A.; Gabruseva, T. 3D convolutional neural networks for stalled brain capillary detection. *Comput. Biol. Med.* **2022**, *141*, 105089. [CrossRef] [PubMed]
26. Li, Z.; Wallace, E.; Shen, S.; Lin, K.; Keutzer, K.; Klein, D.; Gonzalez, J. Train Big, Then Compress: Rethinking Model Size for Efficient Training and Inference of Transformers. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 5958–5968.
27. Mars, M. From Word Embeddings to Pre-Trained Language Models: A State-of-the-Art Walkthrough. *Appl. Sci.* **2022**, *12*, 8805. [CrossRef]
28. Sharma, S.; Asri, L.E.; Schulz, H.; Zumer, J. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv* **2017**, arXiv:1706.09799.
29. 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 of the Association for Computational Linguistics, Stroudsburg, PA, USA, 7–12 July 2002; pp. 311–318.
30. 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, MI, USA, 29 June 2005; pp. 65–72.
31. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.
32. MedicalAI. ClinicalBERT. Available online: https://huggingface.co/medicalai/ClinicalBERT (accessed on 21 July 2023).