ImageCLEF 2021 Best of Labs: The Curious Case of Caption Generation for Medical Images

Aaron Nicolson(✉), Jason Dowling, and Bevan Koopman

Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston 4006, QLD, Australia
{aaron.nicolson,jason.dowling,bevan.koopman}@csiro.au

Abstract. As part of Best of Labs, we have been invited to conduct further investigation on the ImageCLEFmed Caption task of 2021. The task required participants to automatically compose coherent captions for a set of medical images. The most popular means of doing this is with an encoder-to-decoder model. In this work, we investigate a set of choices with regards to aspects of an encoder-to-decoder model. Such choices include what pre-training data should be used, what architecture should be used for the encoder, whether a natural language understanding (e.g., BERT) or generation (e.g., GPT2) checkpoint should be used to initialise the parameters of the decoder, and what formatting should be applied to the ground truth captions during training. For each of these choices, we first made assumptions about what should be used for each choice and why. Our empirical evaluation then either proved or disproved these assumptions—with the aim to inform others in the field. Our most important finding was that the formatting applied to the ground truth captions of the training set had the greatest impact on the scores of the task's official metric. In addition, we discuss a number of inconsistencies in the results that others may experience when developing a medical image captioning system.

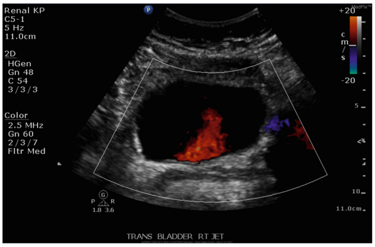
Keywords: Medical image captioning · Encoder-to-decoder · Multi-modal · Warm-starting

1 Introduction

ImageCLEFmed Caption 2021 is an international challenge where teams develop a system that automatically generates a coherent caption for a given medical image (for example, X-ray, computed tomography, magnetic resonance, or ultrasonography) [8,19]. To succeed, the system must not only identify medical concepts but also their interplay. As with most medical image analysis tasks, a deep learning model was the key component of the participants' systems. The model was trained using the provided dataset, containing medical images and their associated ground truth captions. Its training set was relatively small (2.8K

A. Barrón-Cedeño et al. (Eds.): CLEF 2022, LNCS 13390, pp. 190–203, 2022.
https://doi.org/10.1007/978-3-031-13643-6_15

examples), adding complexity to the task. A training example from the task is shown in Fig. 1. The most popular model for medical image captioning is the encoder-to-decoder model: the encoder produces features from a given image which are then used to condition the decoder when generating the caption [18].



(a) Medical image

"This image is a transverse evaluation of the bladder and right ureteral jet. Renal ultrasound studies also include evaluation of the ureterovesical junction through Color Flow Doppler study of fluid movement of the ureteral jet."

(b) Ground truth caption

Fig. 1. The task was to develop an automated system that, given a medical image, could predict the ground truth caption. Training example *synpic100306* from the ImageCLEFmed Caption 2021 dataset is shown, where **(a)** is the medical image and **(b)** is its ground truth caption.

Our approach to ImageCLEFMed Caption 2021 was to use a Vision Transformer (ViT) [4] as the encoder and PubMedBERT [6] as the decoder (both are detailed in Sect. 2) [14]. Neither a ViT nor a domain-specific natural language checkpoint such as PubMedBERT had previously been explored for medical image captioning. As such, we have been invited to conduct a further investigation on the previously mentioned task as part of Best of Labs.

For this work, we aim to investigate a set of important choices for an encoder-to-decoder model:

Choice 1: Pre-training data—The choice in question is what pre-training data to use for warm-starting. Warm-starting refers to the initialisation of a models parameters with those of a pre-trained checkpoint. A checkpoint includes the values of all the learned parameters of a trained model. The pre-training data could be from the general domain (e.g., Wikipedia articles used for BERT [3]); or domain specific (e.g., biomedical corpora used for PubMedBERT [6]). Moreover, does warm-starting with a checkpoint from a related task (e.g., Chest X-Ray (CXR) report generation) improve performance?

Choice 2: Encoder—The architecture of the encoder. Specifically, whether to add convolutional layers to the ViT or not.

Choice 3: Decoder—The type of pre-training task of the decoder checkpoint. The pre-training task could be a Natural Language Understanding (NLU) task (e.g., the self-supervised learning tasks used to form BERT), or a Natural Language Generation (NLG) task (e.g., the language modelling task used to form GPT2 [21]). NLU is the comprehension of natural language through grammar and context while NLG is the construction of natural language based on a given input.

Choice 4: Formatting—How should the captions be formatted for training? The official metric (described in Subsect. 3.2) employs a series of natural language formatting steps, such as removing punctuation and stopwords. These steps may seem innocuous and are rarely reported in other studies, but as part of our submissions we had a number of unexplained performance differences that we posit were a result of the differences between the caption formatting during training and that used for the official metric [14].

Anyone setting out to develop medical image caption generation systems are faced with the above choices, as we were before participating in ImageCLEFmed Caption 2021. From these choices and one's intuition, the following assumptions may be held:

Assumption 1: Pre-train data—Warm-starting with a domain-specific checkpoint, such as PubMedBERT, would outperform warm-starting with a general-domain checkpoint, such as BERT. Moreover, one would assume that an encoder-to-decoder model warm-started with a checkpoint from a related task (e.g., CXR report generation) would outperform a model with its encoder and decoder warm-started with general-domain checkpoints. This is based on our expectations that transferring knowledge learned on a related task to the final task typically results in an improvement in performance—especially when the training set of the final task is relatively small.

Assumption 2: Encoder—That a ViT with convolutional layers would outperform one without. This is based on the fact that convolutional layers (with small kernel sizes) have an inductive bias towards local spatial regions—an advantage for modelling the fine details present in medical images.

Assumption 3: Decoder—That an NLG checkpoint, such as GPT2, would outperform an NLU checkpoint, such as PubMedBERT. This based on the intuition that the pre-training task of an NLG checkpoint would be more transferable to the task of caption generation.

Assumption 4: Formatting—That formatting the ground truth captions of the training set does not have an impact on the performance of a model with regards to the official metric. This is based on the fact that the metric used for the challenge applies a series of formatting steps to both the predicted and ground truth captions—potentially rendering any formatting applied during training redundant.

Curiously, our experience was that many of these intuitive assumptions were not supported by our empirical evaluation.

The remainder of this paper is around an empirical evaluation of a set of models on the ImageCLEFmed Caption 2021 task, where the models were selected to prove or disprove the above assumptions. The results will help to inform others working on similar tasks who may share the same assumption. We test each assumption individually and discuss why they do or do not hold.

2 Background and Related Work

Prior to ImageCLEFmed Caption 2021, a Convolutional Neural Network (CNN) [7] and a decoder-only Transformer [27] were typically employed as the encoder and decoder, respectively. Convolutional layers have an inductive bias towards local spatial regions owing to their small kernel sizes, making them ideal for modelling the fine details present in medical images. Transformers, which leverage the attention mechanism, have the ability to model the relationship between all of its inputs simultaneously, lending themselves to modelling the free text of medical captions [27].

General-domain ImageNet checkpoints were also frequently employed to warm-start the encoder (where ImageNet is a large general-domain image classification task) [10,25]. The transfer of knowledge from the pre-training task to the final task can provide a significant performance boost, particularly when the pre-training dataset is significantly larger than that of the final task. Furthermore, warm-starting is particularly effective when the domain of the pre-training task is similar to that of the final task. However, there is a lack of medical image checkpoints outside of CXR tasks [22]. Furthermore, warm-starting the decoder was not common practice prior to the competition.

Prior to the challenge, ViTs were investigated for computer vision tasks and demonstrated the ability to model the relationship between patches of an image. However, ViTs lack the inductive biases that enable CNNs to perform well on such tasks. Surprisingly, ViTs are able to overcome this deficiency at larger dataset sizes (14M-300M images) [4]. Subsequently, it was demonstrated that a ViT encoder warm-started with an ImageNet checkpoint outperformed its CNN counterpart on general-domain image captioning [12]. Given this, and the aptitude of ImageNet checkpoints at warm-starting medical image tasks, we selected the ViT and its ImageNet checkpoint for the encoder of our original system.

While medical image checkpoints were scarce in the literature, many medical text checkpoints were available. Several large pre-trained NLU encoder-only Transformer checkpoints were formed via the self-supervised learning strategies of BERT [3] and large biomedical corpora. One instance is PubMedBERT [6]—an encoder-only Transformer pre-trained on biomedical articles from the PubMed corpus.[1] However, a decoder is typically warm-started with an NLG decoder-only Transformer checkpoint, such as GPT2 [21]. Despite this, Rothe *et al.* demonstrated that the decoder warm-started with an NLU checkpoint could outperform its NLG counterpart on several sequence-to-sequence NLG tasks [24]. This suggested that PubMedBERT would be ideal to warm-start the decoder.

With our encoder-to-decoder model, ViT2PubMedBERT, we had nine submissions. Amongst the submissions we attempted additional pre-training of the encoder on medical images (from four X-ray datasets), pre-training of the encoder and decoder on a larger medical image captioning datset (ROCO [20]), and additional fine-tuning using reinforcement learning [23]. However, there was

[1] https://pubmed.ncbi.nlm.nih.gov/.

a discrepancy between the validation scores we were attaining on our metrics versus the test scores attained using the official metric: a validation score improvement did not correlate with a test score improvement [14]. One possible reason for this is that the formatting used on the ground truth captions of the training set was different to that used for the official metric.

Our subsequent work following ImageCLEFmed Caption 2021 investigated a range of architectures and checkpoints for warm-starting the encoder and decoder for a related task: CXR report generation. Other than the fact that this is a more specific task (i.e., only one modality is considered), the biggest difference is the size of the datset, with the MIMIC-CXR dataset including 270K examples in the training set [9]. We also found that CNNs such as ResNets outperformed ViT [7]. We also investigated improvements to the ViT and found that a Convolutional vision Transformer (CvT) encoder produced the highest performance. We also found that GPT2 and DistilGPT2 [26] outperformed domain-specific NLU checkpoints such as PubMedBERT—possibly due to the fact that GPT2 is an NLG checkpoint. This was different to the finding of Rothe *et al.*, likely due to one key difference: the task of the encoder for CXR report generation is to produce features from images rather than natural language. Another finding was that PubMedBERT—a domain-specific NLU checkpoint—was able to outperform BERT—a general-domain NLU checkpoint [16]. These findings have influenced some of the aforementioned assumptions—as our ancillary aim of this study is to determine if the findings on MIMIC-CXR are upheld on the ImageCLEFmed Caption 2021 task.

3 Methodology

In this section, we describe the dataset, metrics, models, fine-tuning strategy, image pre-processing, and text formatting.

3.1 Task Description and Dataset

For ImageCLEFmed Caption 2021, participants were tasked with developing a system that could generate a caption for a given medical image. The motivation behind this task is to help develop tools that can aid medical experts with interpreting and summarising medical images, a task that is often time-consuming and a bottleneck in clinical diagnosis pipelines. Each example from the dataset consisted of a medical image and its associated ground truth caption, as shown in Fig. 1. The data was divided into training ($n = 2\,756$), validation ($n = 500$), and test ($n = 444$) sets. Evaluation was performed by comparing the predicted captions to the annotations provided by medical doctors (i.e., the ground truth captions).

3.2 Metrics

We adopted the official metric of ImageCLEFmed Caption 2021 for the validation and test sets: CLEF-BLEU. It was computed as follows for each predicted and ground truth caption:

1. **Lowercased:** The caption was first converted to lower-case.
2. **Remove punctuation:** All punctuation was then removed and the caption was tokenized into its individual words.
3. **Remove stopwords:** Stopwords were then removed using NLTK's English stopword list (NLTK v3.2.2).
4. **Stemming:** Stemming was next applied using NLTK's Snowball stemmer (NLTK v3.2.2).
5. The score was then calculated as the average score of BLEU-1, BLEU-2, BLEU-3, and BLEU-4 between the predicted and ground truth captions [17].

Note that each caption was always considered as a single sentence, even if it contained several sentences.

Furthermore, the following metrics were used for evaluation on the validation set: BLEU-1, BLEU-2, BLEU-3, and BLEU-4 [17], ROUGE-L [11], and CIDEr [28]. This was to aid with understanding how formatting the ground truth captions impacted the performance each model. The formatting is detailed in Subsect. 3.5.

3.3 Models

The encoder-to-decoder models investigated in this work are listed below. An example of one is shown in Fig. 2. The input to the encoder is a medical image. The output of the encoder (CvT-21) is fed to the cross-attention module of the decoder (DistilGPT2), which then generates a caption in an autoregressive fashion—conditioned on the encoders output. It should be noted that each model employs a linear layer that projects the last hidden state of the encoder to the hidden size of the decoder.

ViT2BERT—ViT (86M parameters) is the encoder [4]. It was warm-started with a checkpoint pre-trained on ImageNet-22K (14M images, 21 843 classes) at a resolution of 224×224 and then additionally trained on ImageNet-1K (1M images, 1 000 classes) at resolution of 384×384. BERT (110M parameters) is the decoder, which is pre-trained on BookCorpus [31] and Wikipedia articles in an uncased manner using self-supervised learning [3]. Both ViT and BERT are 12 layers with a hidden size of 768.

ViT2PubMedBERT—Identical to ViT2BERT, except that PubMedBERT (110M parameters) is the decoder. Its main difference to BERT is the pre-training data: abstracts from PubMed (4.5B words) and articles from PubMed Central (13.5B words).

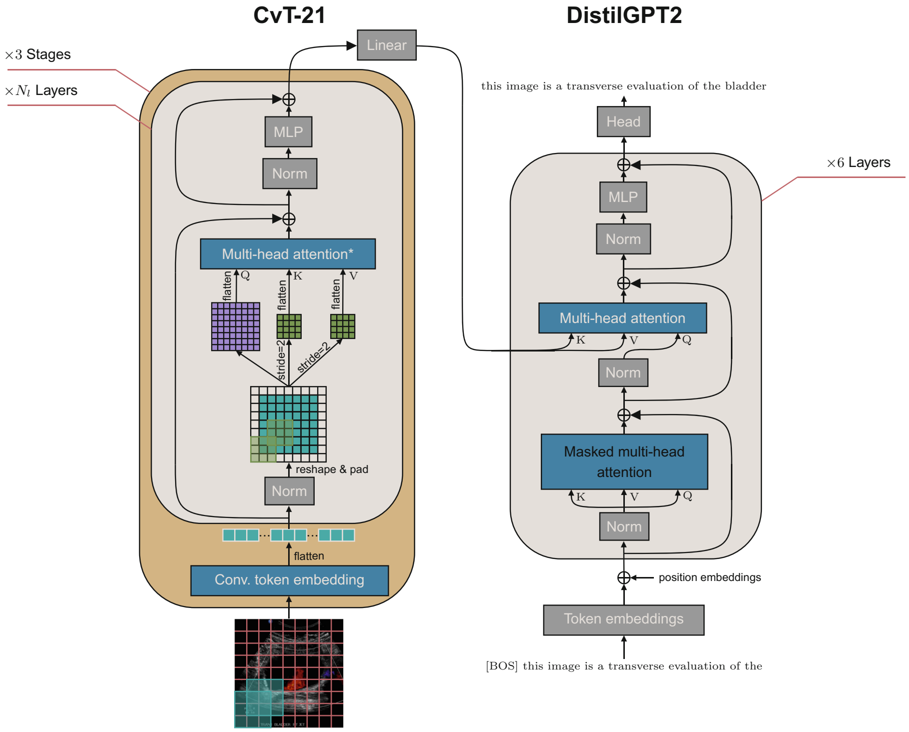


Fig. 2. CvT2DistilGPT2. Q, K, and V are the queries, keys, and values, respectively, for multi-head attention [27]. $*$ indicates that the linear layers for Q, K, and V are replaced with the convolutional layers depicted below the multi-head attention module. [BOS] is the beginning-of-sentence special token. N_l is the number of layers for each stage, where $N_l = 1$, $N_l = 4$, and $N_l = 16$ for the first, second, and third stage, respectively. The head for DistilGPT2 is the same used for language modelling.

ViT2DistilGPT2—Identical to ViT2BERT, except that DistilGPT2 (82M parameters) is the decoder. It is pre-trained using knowledge distillation where DistilGPT2 was the student and GPT2 was the teacher. OpenWebText, a reproduction of OpenAI's WebText corpus, was used as the pre-training data [5]. DistilGPT2 includes 6 layers with a hidden size of 768.

CvT2DistilGPT2—Identical to ViT2DistilGPT2, except that CvT-21 (32M parameters) is the encoder. CvT-21 was warm-started with an ImageNet-22K checkpoint with a resolution of 384×384 [30]. It has three stages, with a combined 21 layers.

CvT2DistilGPT2·MIMIC-CXR—This is CvT2DistilGPT2 warm-started with a MIMIC-CXR checkpoint [15,16]. The checkpoint was not additionally fine-tuned with reinforcement learning on MIMIC-CXR.

3.4 Medical Image Pre-processing

Each medical image $X \in \mathbb{R}^{C \times W \times H}$ (where C, W, and H denote the number of channels, the width, and height, respectively) had an 8-bit pixel depth and three channels ($C = 3$). The image was first resized using bilinear interpolation to a size of $\mathbb{R}^{3 \times 384 \times 384}$. During training, the image was also rotated at an angle sampled from $\mathcal{U}[-5°, 5°]$. Finally, the image was standardised using the mean and standard deviation of each channel provided with the encoder checkpoint.

3.5 Caption Formatting and Generation

We investigated five different formatting strategies for the ground truth captions of the training and validation sets, to determine their impact on CLEF-BLEU:

1. No formatting.
2. Lowercased.
3. Lowercased + no punctuation.
4. Lowercased + no punctuation + no stopwords.
5. Lowercased + no punctuation + no stopwords + stemming.

These formatting steps were described in Subsect. 3.2. When generating the captions during validation and testing, beam search with a beam size of four and a maximum number of 128 subwords was used.

3.6 Fine-Tuning

Teacher forcing was used for fine-tuning [29]. Each model was implemented in PyTorch version 1.10.1 and trained with 4×NVIDIA P100 16 GB GPUs. To reduce memory consumption, we employed PyTorch's automatic mixed precision (a combination of 16-bit and 32-bit floating point variables). For fine-tuning, the following configuration was used: categorical cross-entropy as the loss function; a mini-batch size of 32; early stopping with a patience of 20 epochs and a minimum delta of $1e{-}4$; *AdamW* optimiser for gradient descent optimisation [13]; an initial learning rate of $1e{-}5$ and $1e{-}4$ for the encoder and all other parameters, respectively, following [2]. All other hyperparameters for *AdamW* were set to their defaults. To select the best epoch for a model, the highest validation BLEU-4 score was used.

Table 1. Results on the validation and test sets of ImageCLEFmed Caption 2021. A higher colour saturation indicates a higher score. For CLEF-BLEU, the full formatting described in Subsect. 3.2 was applied to both the predicted and ground truth captions for every row. For the other metrics and for training, the indicated formatting was applied to the ground truth captions. ↪ MIMIC-CXR indicates CvT2DistilGPT2·MIMIC-CXR.

Model	**Validation Set**							**Test Set**
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**	**ROUGE-L**	**CIDEr**	**CLEF BLEU**	**CLEF BLEU**
Strategy 1: No formatting								
ViT2BERT	0.315	0.258	0.227	0.206	0.275	1.462	0.405	0.406
ViT2PubMedBERT	0.348	0.291	0.258	0.236	0.306	1.703	0.432	0.406
ViT2DistilGPT2	0.363	0.319	0.299	0.288	0.328	2.243	0.428	0.384
CvT2DistilGPT2	**0.370**	**0.326**	**0.305**	**0.293**	**0.332**	**2.297**	0.433	0.400
↪MIMIC-CXR	0.348	0.304	0.283	0.272	0.320	2.212	0.427	0.407
Strategy 2: Lowercased								
ViT2BERT	0.358	0.304	0.277	0.261	0.317	1.932	0.405	0.406
ViT2PubMedBERT	**0.395**	**0.342**	**0.314**	**0.297**	0.353	2.243	0.432	0.406
ViT2DistilGPT2	0.370	0.322	0.299	0.287	0.336	2.340	0.437	0.408
CvT2DistilGPT2	0.380	0.332	0.308	0.295	**0.356**	**2.466**	0.448	0.405
↪MIMIC-CXR	0.354	0.304	0.281	0.269	0.318	2.094	0.402	0.397
Strategy 3: Lowercased + no punctuation								
ViT2BERT	0.378	0.325	0.299	0.286	0.347	2.328	0.444	0.404
ViT2PubMedBERT	**0.417**	**0.364**	**0.337**	**0.323**	**0.379**	**2.591**	0.453	0.426
ViT2DistilGPT2	0.387	0.338	0.314	0.301	0.351	2.400	0.441	0.394
CvT2DistilGPT2	0.388	0.338	0.315	0.302	0.356	2.521	0.446	0.401
↪MIMIC-CXR	0.373	0.320	0.296	0.283	0.333	2.267	0.414	0.400
Strategy 4: Lowercased + no punctuation + no stopwords								
ViT2BERT	0.355	0.319	0.302	0.292	0.327	2.430	0.451	**0.430**
ViT2PubMedBERT	**0.374**	**0.337**	**0.319**	**0.308**	**0.347**	**2.601**	**0.458**	0.423
ViT2DistilGPT2	0.342	0.308	0.292	0.283	0.311	2.356	0.421	0.410
CvT2DistilGPT2	0.332	0.301	0.286	0.277	0.315	2.409	0.430	0.400
↪MIMIC-CXR	0.322	0.290	0.274	0.264	0.308	2.342	0.422	0.398
Strategy 5: Lowercased + no punctuation + no stopwords + stemming								
ViT2BERT	0.364	0.321	0.301	0.290	0.328	2.355	0.419	0.396
ViT2PubMedBERT	**0.393**	**0.346**	**0.323**	**0.310**	**0.366**	**2.593**	0.416	0.410
ViT2DistilGPT2	0.355	0.317	0.299	0.289	0.326	2.444	0.399	0.383
CvT2DistilGPT2	0.355	0.316	0.298	0.288	0.330	2.462	0.416	0.391
↪MIMIC-CXR	0.353	0.313	0.295	0.285	0.328	2.457	0.425	0.394

4 Results and Discussion

Table 1 presents results from our empirical evaluation. We shall discuss the results as they relate to the four assumptions detailed in the introduction.

4.1 Assumption 1: Pre-training Data

The first assumption was that PubMedBERT as the decoder would outperform BERT—as it is a domain-specific checkpoint. In terms of the validation scores, this assumption stood, as ViT2PubMedBERT outperformed ViT2BERT (except for validation CLEF-BLEU on Strategy 5). However, this finding was not consistent with the test scores, with ViT2BERT producing the highest score of any model (0.430). This contradiction indicates that the results on the validation set do not generalise to the test set.

The next assumption was that warm-starting the encoder-to-decoder model with a CXR report generation checkpoint would improve performance, especially given the small size of the training set. The performance of CvT2DistilGPT2·MIMIC-CXR was not significantly different from CvT2DistilGPT2 in terms of the test scores. However, the validation scores refute the assumption, as CvT2DistilGPT2 consistently produced higher validation scores. One explanation is that X-rays are not the dominant modality in the ImageCLEFmed Caption 2021 training set, where computed tomography and magnetic resonance are more represented [1, Table 1].

4.2 Assumption 2: Encoder

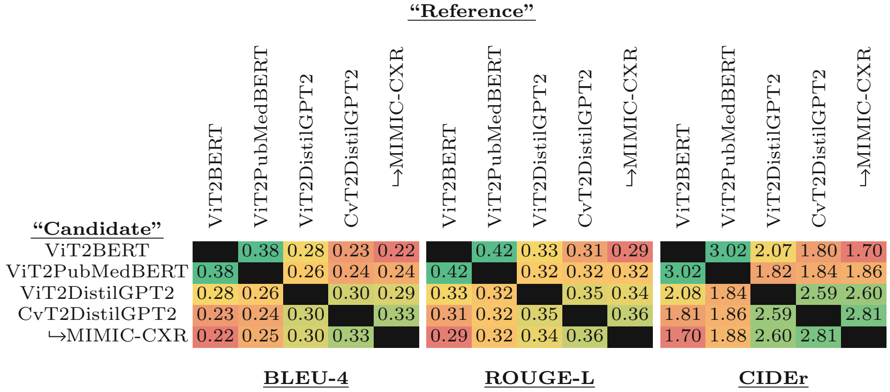
Here, we determine if including convolutional layers in the encoder, i.e., choosing CvT over ViT, improves performance. When no formatting is used during training, CvT2DistilGPT2 attains higher validation and test scores than ViT2DistilGPT2. However, when formatting is used during training, the picture becomes unclear. For example, CvT2DistilGPT2 attains higher validation and test scores for Strategies 3 and 5, while ViT2DistilGPT2 attains higher validation and test scores for Strategies 2 and 4. Therefore, it is unclear if adding convolutional layers to ViT (i.e., using CvT instead) is advantageous for this task, refuting the findings in [16]. However, it should be noted that CvT consumes drastically fewer parameters than ViT—demonstrating parameter efficiency.

4.3 Assumption 3: Decoder

The assumption made for the decoder was that an NLG checkpoint would outperform an NLU checkpoint. Comparing ViT2BERT to ViT2DistilGPT2 on the validation scores for `no formatting`, DistilGPT2 as the decoder outperforms BERT by a large margin. However, this margin decreases as the number of formatting steps increases—BERT as the decoder even outperforms DistilGPT2 on the validation set in certain cases. This indicates that BERT is less sensitive to the formatting steps applied to the ground truth captions of the training set. However, their scores on the test set tell a different story. BERT as the decoder attained a higher test score than DistilGPT2 for each strategy, except Strategy 2. Again, the results on the validation set are misleading, as they do not generalise to the test set [16].

4.4 Assumption 4: Formatting

Originally, we assumed that formatting the ground truth captions of the training set would have no impact on performance. However, the results indicate that, in fact, it does have an impact. ViT2BERT experienced an absolute test CLEF-BLEU improvement of 2.4% when Strategy 3 was used instead of no formatting. This is opposite to the original assumption made—applying formatting to the predicted and ground truth captions before evaluation does not mean that there is no benefit to using formatted ground truth captions as the training target.



	"Reference"														
	BLEU-4					**ROUGE-L**					**CIDEr**				
"Candidate"	ViT2BERT	ViT2PubMedBERT	ViT2DistilGPT2	CvT2DistilGPT2	↪MIMIC-CXR	ViT2BERT	ViT2PubMedBERT	ViT2DistilGPT2	CvT2DistilGPT2	↪MIMIC-CXR	ViT2BERT	ViT2PubMedBERT	ViT2DistilGPT2	CvT2DistilGPT2	↪MIMIC-CXR
ViT2BERT	■	0.38	0.28	0.23	0.22	■	0.42	0.33	0.31	0.29	■	3.02	2.07	1.80	1.70
ViT2PubMedBERT	0.38	■	0.26	0.24	0.24	0.42	■	0.32	0.32	0.32	3.02	■	1.82	1.84	1.86
ViT2DistilGPT2	0.28	0.26	■	0.30	0.29	0.33	0.32	■	0.35	0.34	2.08	1.84	■	2.59	2.60
CvT2DistilGPT2	0.23	0.24	0.30	■	0.33	0.31	0.32	0.35	■	0.36	1.81	1.86	2.59	■	2.81
↪MIMIC-CXR	0.22	0.25	0.30	0.33	■	0.29	0.32	0.34	0.36	■	1.70	1.88	2.60	2.81	■

Fig. 3. The *similarity* between the predicted captions of the models on the validation set. Each metric requires *reference* and *candidate* captions. Here, the predicted captions of one model are used as the reference captions (instead of the ground truth captions) and the predicted captions of the other model as the candidate captions. No formatting was applied to the ground truth or predicted captions during training or evaluation. The presented matrices are not symmetric as each metric treats the candidate and reference captions differently. ↪ MIMIC-CXR indicates CvT2DistilGPT2·MIMIC-CXR.

On another note, BERT and PubMedBERT appear to be either less sensitive to formatting, or benefit from formatting, especially with the third and fourth formatting strategies. This could be caused by multiple factors; an NLU checkpoint may be more robust than an NLG checkpoint to formatting. Moreover, DistilGPT2 may be disadvantaged by the fact that it is cased, rather than uncased like BERT and PubMedBERT.

4.5 Similarity Between Predicted Captions

The results in Table 1 are solely focus on model differences according to their effectiveness on the ImageCLEFmed Caption 2021 task. While this provides some insight, we also want to understand how similar the captions generated by the models (i.e., the predicted captions) are to one another. Specifically, two models may have a similar effectiveness on the ImageCLEFmed Caption task, but they may generate significantly different captions. To compute the similarity between a pair of models, we give their generated captions to a metric. Each metric consumes *candidate* and *reference* captions and treats each differently. Hence, we conduct a pair-wise comparison between the generated captions of a pair of models. The results of this are shown in Fig. 3.

It can be seen that the generated captions of ViT2PubMedBERT and ViT2BERT were the most similar to one another; CvT2DistilGPT2 and CvT2DistilGPT2·MIMIC-CXR also attained a high similarity. This is somewhat surprising given that the pre-training data of the checkpoints in each comparison are different. However, the high similarity makes sense from an architectural point of view as the models in each comparison employ the same (or very similar) encoder and decoder architectures. The most dissimilar models are ViT2BERT and CvT2DistilGPT2·MIMIC-CXR. This makes sense as they are the most dissimilar in terms of their pre-training data, encoder, and decoder. Finally, ViT2DistilGPT2 versus CvTDistilGPT2 had a higher similarity than ViT2DistilGPT2 and ViT2BERT, indicating that the decoder has a larger impact on dissimilarity than the encoder.

5 Conclusion

For our Best of Labs contribution, we posed a set of assumptions regarding choices pertaining to an encoder-to-decoder model for medical image captioning, and then set out to prove or disprove them through an empirical evaluation. Our key finding was that the type of formatting applied to the ground truth captions of the training set had the greatest impact on the scores obtained on the official metric of the task. The results also indicate that BERT and PubMedBERT as the decoder are less sensitive to additional formatting steps than DistilGPT2. Unfortunately, assumptions made about the pre-training data, encoder, and decoder could not be proved or disproved, as the results were inconclusive. A key problem was that the hierarchy of performance amongst the models on the validation set did not generalise to the test set. This could be due to the limited size of the dataset or significant differences between the validation and test sets.

Acknowledgement. This work was partially funded by CSIRO's Machine Learning and Artificial Intelligence Future Science Platform (MLAI FSP).

References

1. Charalampakos, F., Karatzas, V., Kougia, V., Pavlopoulos, J., Androutsopoulos, I.: AUEB NLP group at ImageCLEFmed caption tasks 2021. In: Proceedings of the 12th International Conference of the CLEF Association, Bucharest, Romania, pp. 1–17, September 2021
2. Chen, Z., Song, Y., Chang, T., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1439–1449. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.112

3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long and Short Papers), Minneapolis, Minnesota, vol. 1, pp. 4171–4186. Association for Computational Linguistics, June 2019. https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
4. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929 [cs.CV], October 2020
5. Gokaslan, A., Cohen, V.: OpenWebText Corpus (2019). http://Skylion007.github.io/OpenWebTextCorpus
6. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. arXiv:2007.15779 [cs.CL], July 2020
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2016. https://doi.org/10.1109/cvpr.2016.90
8. Ionescu, B., et al.: Overview of the ImageCLEF 2021: multimedia retrieval in medical, nature, internet and social media applications. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 345–370. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_23
9. Johnson, A.E.W., et al.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv:1901.07042 [cs.CV], January 2019
10. Ke, A., Ellsworth, W., Banerjee, O., Ng, A.Y., Rajpurkar, P.: CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation. In: Proceedings of the Conference on Health, Inference, and Learning, pp. 116–124. ACM, April 2021. https://doi.org/10.1145/3450439.3451867
11. Lin, C., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004), Barcelona, Spain, pp. 605–612, July 2004. https://doi.org/10.3115/1218955.1219032. https://aclanthology.org/P04-1077
12. Liu, W., Chen, S., Guo, L., Zhu, X., Liu, J.: CPTR: full transformer network for image captioning. arXiv:2101.10804 [cs.CV], January 2021
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=Bkg6RiCqY7
14. Nicolson, A., Dowling, J., Koopman, B.: AEHRC CSIRO at ImageCLEFmed caption 2021. In: Proceedings of the 12th International Conference of the CLEF Association, Bucharest, Romania, pp. 1–12, September 2021
15. Nicolson, A., Dowling, J., Koopman, B.: Chest X-Ray report generation checkpoints for CvT2DistilGPT2 (2022). https://doi.org/10.25919/64WX-0950
16. Nicolson, A., Dowling, J., Koopman, B.: Improving chest X-Ray report generation by leveraging warm-starting, January 2022
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics, July 2002. https://doi.org/10.3115/1073083.1073135. https://www.aclweb.org/anthology/P02-1040
18. Pavlopoulos, J., Kougia, V., Androutsopoulos, I., Papamichail, D.: Diagnostic captioning: a survey, January 2021. arXiv:2101.07299 [cs.CV]

19. Pelka, O., Ben Abacha, A., García Seco de Herrera, A., Jacutprakart, J., Friedrich, C.M., Müller, H.: Overview of the ImageCLEFmed 2021 concept & caption prediction task. In: CLEF2021 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 21–24 September 2021
20. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): a multimodal image dataset. In: Stoyanov, D., et al. (eds.) LABELS/CVII/STENT -2018. LNCS, vol. 11043, pp. 180–189. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01364-6_20
21. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
22. Rajpurkar, P., et al.: CheXNet: radiologist-level pneumonia detection on chest X-Rays with deep learning. arXiv:1711.05225 [cs.CV], November 2017
23. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, July 2017. https://doi.org/10.1109/cvpr.2017.131
24. Rothe, S., Narayan, S., Severyn, A.: Leveraging pre-trained checkpoints for sequence generation tasks. Trans. Assoc. Comput. Linguist. **8**, 264–280 (2020). https://doi.org/10.1162/tacl_a_00313
25. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
26. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL], October 2019
27. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS 2017, pp. 6000–6010. Curran Associates Inc., Red Hook (2017)
28. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015
29. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural Comput. **1**(2), 270–280 (1989). https://doi.org/10.1162/neco.1989.1.2.270
30. Wu, H., et al.: CvT: introducing convolutions to vision transformers. arXiv:2103.15808 [cs.CV], March 2021
31. Zhu, Y., et al.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 2015