

AEHRC CSIRO at ImageCLEFmed Caption 2021

Aaron Nicolson¹, Jason Dowling¹ and Bevan Koopman¹

¹Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston 4006, Queensland, Australia

Abstract

We describe our participation in the ImageCLEFmed Caption task of 2021. The task required participants to automatically compose coherent captions for a set of medical images. To this end, we employed a sequence-to-sequence model for caption generation, where its encoder and decoder were initialised with pre-trained Transformer checkpoints. In addition, we investigated the use of Self-Critical Sequence Training (SCST) (which offered a marginal improvement) and pre-training on five external medical image datasets. Overall, our approach was kept intentionally general so that it might be applied to tasks other than medical image captioning. AEHRC CSIRO placed third amongst the participating teams in terms of BLEU score—with a score 0.078 worse than the first placed participant. Our best-performing submission had the simplest configuration—it did not use SCST or pre-training on any of the external datasets. An overview of ImageCLEFmed Caption 2021 is available at: <https://www.imageclef.org/2021/medical/caption>.

Keywords

Medical Image Captioning; Diagnostic Captioning; Medical Images; Image Captioning; Multi-modal; Sequence-to-sequence; Vision Transformer; PubMedBERT; Multi-label Classification;

1. Introduction

For ImageCLEFmed caption 2021, teams were tasked with developing systems that could automatically produce coherent captions for the entirety of a medical image [1, 2]. To succeed, a system must not only identify medical concepts but also their interplay. A system that can achieve this could improve the efficiency of radiologists' interpretation. An example of an image and its ground truth caption for ImageCLEFmed Caption 2021 is provided in Figure 1. Typical medical image captioning approaches make use of either a sequence-to-sequence (seq2seq) model to generate a caption for a medical image [3], or image retrieval, where it is assumed that similar images have similar captions [4]. For our submissions, a seq2seq model was considered only.

While medical data has many unique characteristics, general-purpose Natural Language Processing (NLP) and Computer Vision (CV) methods have proven effective in many domain-specific medical tasks. In NLP, for example, general-domain self-supervised pre-training strategies—such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) used to produce

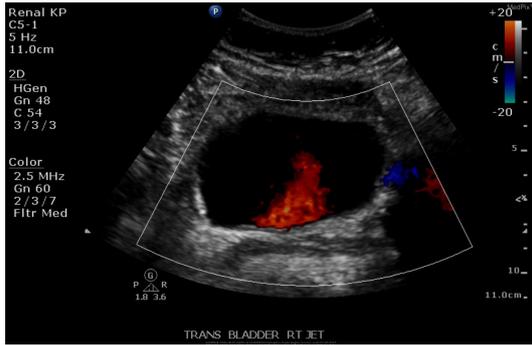
Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), September 21-24, Bucharest, Romania

✉ aaron.nicolson@csiro.au (A. Nicolson); jason.dowling@csiro.au (J. Dowling); bevan.koopman@csiro.au (B. Koopman)

ORCID 0000-0002-7163-1809 (A. Nicolson); 0000-0001-9349-2275 (J. Dowling); 0000-0001-5577-3391 (B. Koopman)

 © 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)



(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

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

Bidirectional Encoder Representations from Transformers (BERT) [5]—have been successfully adapted to medical text. One instance is PubMedBERT [6]—a Transformer encoder [7] pre-trained on PubMed articles using domain-specific self-supervised pre-training strategies. Another example is the use of Transfer Learning (TL) to significantly improve medical image classification accuracy on small datasets [8]. Here, a portion of a Convolutional Neural Network (CNN) trained on ImageNet 2012 [9] (a general-domain image classification dataset) is fine-tuned on the small amount of data for the medical image classification task.

A number of more recent NLP and CV machine learning techniques have not been investigated on medical data. One such approach for sequence generation is the use of pre-trained Transformer checkpoints to initialise both the encoder and the decoder of a seq2seq model [10]. Another method is the Vision Transformer (ViT)—a pre-trained Transformer checkpoint for image classification, which takes 16x16 patches of the image as input [11]. Building from this, a pre-trained ViT checkpoint was paired with a Transformer decoder to form a seq2seq model for image captioning [12].

Motivated by previous adaptations of general-domain NLP and CV machine learning techniques to medical data and the slew of recent techniques that have not been investigated on medical data, we investigate a seq2seq model for medical image captioning that employs a pre-trained ViT checkpoint as the encoder and the pre-trained PubMedBERT checkpoint as the decoder. We also experiment with various pre-training and fine-tuning strategies, such as additionally pre-training the encoder on a multi-label medical image classification task, as well as fine-tuning the seq2seq model with Self-Critical Sequence Training (SCST) [13].

2. Task Description

2.1. Dataset

The focus for ImageCLEFmed Caption 2021 was to use real medical images and have participants develop automated systems to predict natural language captions; evaluation was performed

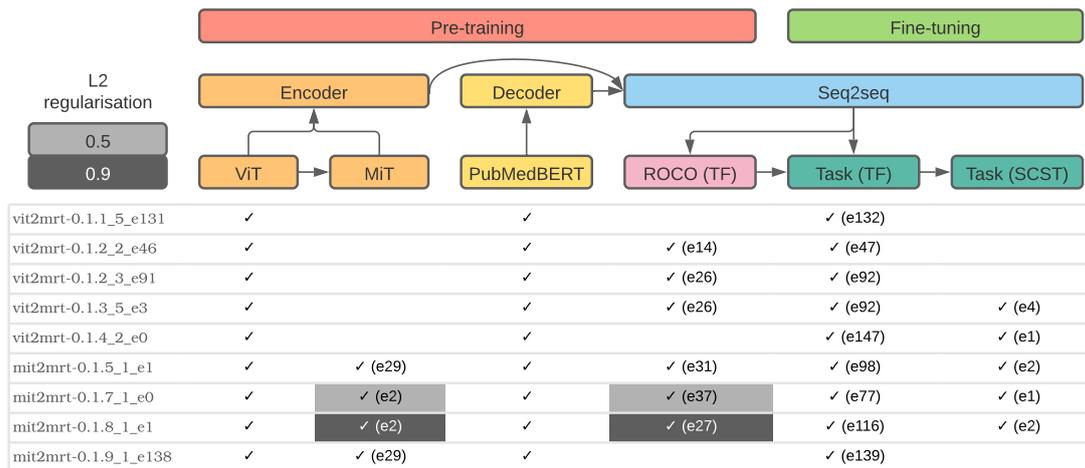


Figure 2: The stages of pre-training and fine tuning for each submission. The seq2seq model is formed by combining ViT with PubMedBERT or MIT with PubMedBERT. The order of training is from left-to-right; the MIT encoder is formed by additionally pre-training ViT on the medical image multi-label classification task described in Section 3.2.1; pre-training on the ROCO dataset (described in Section 3.2.1) with TF occurs before fine-tuning with TF; fine-tuning with SCST occurs after fine-tuning with TF. The epoch for each stage (signified by “e” followed by the epoch number) is selected using early stopping, as described in Section 3.2. Note that the epoch specified in each submission identifier for the final stage of training is offset by one. The selected epochs for ViT and PubMedBERT are not given, as these are publicly available pre-trained checkpoints. L2 regularisation was used during the MIT pre-training stage and for the ROCO pre-training stage with two of the submissions, where the used regularisation terms are indicated in the legend on the top-left corner.

by comparing the predicted captions to the annotations provided by medical doctors (i.e. the ground truth captions). Each example from the dataset consisted of a medical image and its associated ground truth caption, as shown in Figure 1. The training, validation, and test sets comprise of 2,756, 500, and 444 examples, respectively. We refer to the ImageCLEFmed Caption 2021 dataset as the task’s dataset henceforth.

2.2. Metrics

Each caption (predicted and ground truth) was pre-processed in the following way: The caption was first converted to lower-case. All punctuation was then removed and the caption was tokenized into its individual words. Stopwords were then removed using NLTK’s English stopword list (NLTK v3.2.2). Stemming was next applied using NLTK’s Snowball stemmer (NLTK v3.2.2). The score was then calculated as the average score of BLEU-1, BLEU-2, BLEU-3, and BLEU-4 [14]. Note that the caption was always considered as a single sentence, even if it contained several sentences. No smoothing function was used. All scores were summed and averaged over the number of captions, giving the final score. One downside of using BLEU for medical image caption evaluation is that it is a word overlap measure and may not capture clinical correctness, as noted in [4].

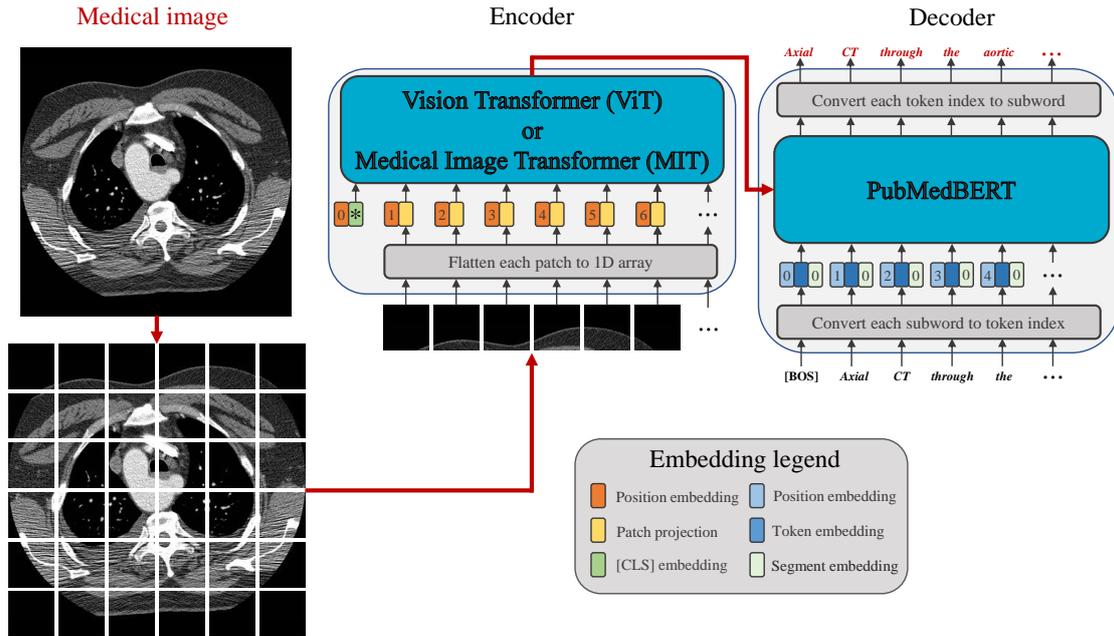


Figure 3: Transformer-based seq2seq model for medical image captioning. The red elements depict the general path from medical image to the predicted caption. The encoder for each submission was either ViT or MIT.

3. Methodology

The nine submissions for the AEHRC CSIRO team are described in Figure 2. In this section, we describe the model architecture, the pre-training and fine-tuning strategy, as well as the external pre-training datasets for each submission. Figure 2 helps the reader identify the stages of pre-training and fine tuning for the encoder, decoder, and the seq2seq model that were used for each submission, along with the epoch chosen for each stage.

3.1. Model

The same model architecture was used for each submission. An overview of the model is shown in Figure 3. In terms of architecture, the encoder is identical to ViT [11, ViT-Base, Table 1] and the decoder is identical to PubMedBERT [6]. Both ViT and PubMedBERT use 12 hidden layers, each with a size of 768, an intermediate size of 3,072 [7, see d_{ff} in Section 3.3], and 12 scaled dot-product attention heads. Next, we describe the medical image pre-processing, followed by the encoder and decoder.

Medical image pre-processing: A given medical image $X \in \mathbb{R}^{C \times W \times H}$ (where C , W , and H denote the number of channels, the width, and the height of the medical image, respectively) is first resized using bilinear interpolation so that its smallest side has 416 pixels. Next, the resized image is cropped to a size of $\mathbb{R}^{3 \times 384 \times 384}$ (the size required for ViT), with the crop location random during training and centered during testing. The cropped image is then split into a

set of non-overlapping patches—each of size $\mathbb{R}^{3 \times 16 \times 16}$ (i.e., 576 non-overlapping patches). Each patch is then flattened into a one-dimensional array of size \mathbb{R}^{768} . A colour depth of 8-bits was used for the images (where images with a higher colour depth were downsampled to 8-bits).

Encoder: The set of inputs given to the encoder consist of the projection of each patch and the [CLS] embedding. The patch projections are formed by passing each flattened patch through a learnable projection matrix: $W_{patch}^{encoder} \in \mathbb{R}^{768 \times 768}$. The [CLS] embedding is learnt using matrix $W_{class}^{encoder} \in \mathbb{R}^{1 \times 768}$ (where [CLS] is the classification token whose corresponding output is fed to a classification head during pre-training). The corresponding output for the [CLS] embedding forms an aggregate representation over all patches. Before the set of inputs are given to the first ViT hidden layer, a position embedding is added to each element of the set. There are 577 position embeddings for the encoder, with position “0” reserved for the [CLS] embedding and positions “1” to “576” reserved for the patch projections (which provide information about the location of each patch within the medical image). Each position embedding is stored in a learnable matrix: $W_{position}^{encoder} \in \mathbb{R}^{577 \times 768}$.

The weights for the encoder (including its embeddings and the patch projection) are initialised using one of the pre-trained ViT checkpoints from [11]. This checkpoint has been pre-trained on ImageNet21k [15] and then subsequently on ImageNet 2012 [9]. We also investigate additionally pre-training this checkpoint on a multi-label medical image classification task; we denote this *Medical Image Transformer* (MIT). The multi-label medical image classification task is comprised of four datasets, as described in Section 3.2.1. The pre-trained checkpoint of either ViT or MIT is used to initialise the encoder for the submissions in Figure 2, where a tick in the “MIT” column indicates that MIT was used over ViT. Moreover, submission identifiers using the pre-trained ViT checkpoint are labelled in Figure 2 starting with “vit”, while those using a pre-trained MIT checkpoint have are labelled starting with “mit”.

Decoder: The weights of the decoder (along with its embeddings) are initialised using the pre-trained PubMedBERT checkpoint [6]. We classify PubMedBERT as a *Medical Report Transformer* (MRT)—a Transformer that has been pre-trained on medical text (in this case medical literature) and is suitable for generating medical reports. The output of the last hidden layer of the encoder is fed to each decoder hidden layer via a randomly initialised multi-head cross-attention module, which is inserted between the masked multi-head self-attention module and the Feedforward Neural Network (FNN) module of each layer [7, Section 3.1, Decoder]. PubMedBERT has a vocabulary size of 30,522, comprising subword units. When feeding a subword unit to the encoder, it is first converted to its corresponding token index, and then subsequently into a token embedding. Each token embedding is stored in learnable matrix $W_{token}^{decoder} \in \mathbb{R}^{30,522 \times 768}$.

Next, a position and a segment embedding are added to the token embedding. The position embedding indicates the location of the subword within the caption. A maximum of 512 positions are used for PubMedBERT, with each position embedding stored in learnable matrix $W_{position}^{decoder} \in \mathbb{R}^{512 \times 768}$. As only one caption is generated per medical image (even though PubMedBERT is pre-trained using two segments), the embedding for segment “0” will only be used. Each segment embedding is stored in learnable matrix $W_{segment}^{decoder} \in \mathbb{R}^{2 \times 768}$.

When generating a caption, the token [BOS] (beginning of sentence) is first fed to output the

first subword of the caption. Caption generation finishes once the decoder generates the [EOS] token. Each submission used PubMedBERT, as shown by the submission identifiers in Figure 2 (i.e. “vit2mrt” and “mit2mrt”), where “mrt” indicates that the decoder is an MRT, where, in this case, PubMedBERT is the MRT. During testing, the maximum amount of subwords that the decoder could generate was set to 128. Beam search was also used, with a beam size of eight. Additionally, all n -grams of size three were only allowed to occur once.

3.2. Pre-training and fine-tuning

Next, we describe the pre-training and fine-tuning strategies for the submissions. Here, fine-tuning refers to training on the task’s dataset. Pre-training refers to training on other, external datasets we selected; this was done before the fine-tuning stage. Teacher Forcing (TF) with categorical cross entropy loss was used to fine-tune each seq2seq model on the task’s dataset [16]. We also investigate additionally fine-tuning the seq2seq models (which have already been fine-tuned using TF) with Self Critical Sequence Training (SCST) [13]. Submissions that were fine-tuned with TF and then SCST have a tick in the “SCST” column of Figure 2.

For pre-training the seq2seq models, we used the Radiology Objects in COntext (ROCO) medical image captioning dataset (described in Section 3.2.1) with TF—before fine-tuning on the task’s dataset. A tick in the “ROCO” column of Figure 2 signifies if this stage of pre-training was conducted for a submission.

The *AdamW* optimiser [17] was used for gradient descent optimisation during pre-training and fine-tuning. A learning rate of $1e - 7$ was used for fine-tuning on the task’s dataset with SCST. A learning rate of $5e - 5$ and a linear warm-up of 10,000 training steps from a learning rate of zero was used when pre-training the MITs, pre-training on ROCO with TF, and when fine-tuning on the task’s dataset with TF. All other hyperparameters for *AdamW* were set to their defaults. For the pre-training strategy in [11], L2 regularisation (with a term of 0.9) helped to prevent overfitting the ViT during pre-training. Motivated by this, we investigated L2 regularisation for pre-training only (i.e., for pre-training the MITs and when pre-training on ROCO using TF). We investigated an L2 regularisation term with two of the submissions, as shown by the legend in Figure 2.

A mini-batch size of 64 was used to pre-train each MIT, to pre-train on ROCO with TF, and to fine-tune on the task’s dataset with TF. A mini-batch size of eight was used to fine-tuning on the task’s dataset with SCST. For epoch selection at each stage of pre-training and fine tuning in Figure 2, early stopping with a patience of five was used. The validation micro F1 score was the monitored metric for early stopping with each MIT and the validation BLEU score (BLEU is described in Section 2.2) was the monitored metric for early stopping with the seq2seq models. For submission vit2mrt-0.1.1_5_e131, the early stopping criteria was not enforced until epoch 50, as the BLEU score did not increase from zero until after this epoch. When fine-tuning on the task’s dataset using SCST, the maximum amount of subwords that the decoder could generate was set to 32 due to memory restrictions. Moreover, greedy search was used (i.e., a beam size of 1) when generating the baseline for SCST [13].

3.2.1. Pre-training datasets

A number of external medical image datasets were used to pre-training the MIT, to take advantage of any stored knowledge about medical images when fine-tuning on the task’s dataset. Specifically, four medical image multi-label classification datasets were identified; these are shown in Table 1. While CheXpert and MURA included validation sets, test sets were not available with any of the four datasets. We refrain from using these validation sets as the CheXpert validation set has been used as a test set previously [18]. Instead, 5% of each training set was selected and removed to form a validation set. Together, the datasets have 325 classes, 482,197 training examples, and 25,377 validation examples. Given the number of classes, the weights of the classification head for the pre-trained ViT checkpoint were replaced with randomly-initialised learnable weight matrix $W_{head}^{encoder} \in \mathbb{R}^{768 \times 325}$, before pre-training on the medical image multi-label classification task to form MIT. The number of classes and the number of examples for each dataset are detailed in the table

Table 1

Summary of the number of examples in the training and validation splits for the four medical image multi-label classification datasets, as well as the number of classes.

Dataset	Training split	Validation split	Classes
PadChest	152,787	8,041	254
CheXpert	212,244	11,170	42
ChestX-ray14	82,198	4,326	15
MURA	34,968	1,840	14
Total	482,197	25,377	325

- **PadChest** includes 160,828 chest X-rays obtained from 67,000 patients of the San Juan Hospital (Spain) from 2009 to 2017 [19]. Each X-ray has an associated report produced by a radiologist. From these reports, labels were extracted manually by trained physicians (for 27% of the X-rays) and automatically (for 73% of the X-rays) using a supervised method. The labels covered six different position views, 174 different radiographic findings, 19 differential diagnoses, and 104 anatomic locations. The labels were then organised into a hierarchical taxonomy and mapped to Unified Medical Language System (UMLS) Concept Unique Identifier (CUI) codes. For our work, the considered 254 classes were derived from the labelCUIs, LocalizationsCUIs, Modality_DICOM, and the ViewPosition_DICOM labels described in [19, Table 11], where the Digital Imaging and Communications in Medicine (DICOM) fields were extracted from the X-ray. We found that 33 of the images were corrupt and were thus excluded.^{1,2}
- The **CheXpert** training set contains 223,414 chest radiographs of 65,240 patients from the Stanford Hospital [18]. The studies were performed between October 2002 and July 2017. An automatic system was used to extract 14 observations from the associated radiology

¹The corrupt filenames are available at: https://github.com/anicolson/supplementary/blob/main/padchest_corrupt.txt

²PadChest is available at: <http://bimcv.cipf.es/bimcv-projects/padchest/>

reports (*no finding, enlarged cardiom, cardiomegaly, lung lesion, lung opacity, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, and support devices*). Each observation class was rated as either positive, uncertain, or negative, thus resulting in 42 classes.³

- **ChestX-ray14** contains 86,524 chest X-rays (collected from 1992 to 2015) concerned with common thorax diseases [20]. It consists of 15 disease labels (*atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, consolidation, edema, emphysema, fibrosis, pleural thickening, and hernia*). The labels were mined from the associated radiological reports of the X-rays.⁴
- The **MURA** training set comprises of 36,808 musculoskeletal radiographs from 13,457 studies of 11,184 patients, where each study is manually labelled by radiologists as either normal or abnormal [21]. Each radiograph concerns one of seven sections of the body (*elbow, finger, hand, humerus, forearm, shoulder, or wrist*), where each is classed as either normal or abnormal, resulting in 14 total classes. MURA is a multi-class classification task unlike the previous datasets.⁵

The **ROCO** dataset was used to pre-train the seq2seq models before fine-tuning, as depicted in Figure 2. ROCO comprises of image-caption pairs from PubMed Central articles, where compound, multi-pane, and non-radiology images were removed using an automatic system [22]. The ROCO training and validation sets contain 65,450 and 8,180 examples, respectively. The ROCO dataset contains several medical imaging modalities including computer tomography, ultrasound, X-ray, fluoroscopy, positron emission tomography, mammography, magnetic resonance imaging, and angiography.⁶

4. Results and discussion

The BLEU scores for each submission on the validation and test sets of the task’s dataset are shown in Table 2.⁷ Submission vit2mrt-0.1.1_5_e131 attained the highest test score. This submission had the simplest configuration—no regularisation, no SCST, no ROCO pre-training, and no MIT pre-training. This indicates that the additional steps considered for the other configurations were not suited to the task’s dataset. One possible explanation is the small size of the task’s dataset.

Another observation is the large discrepancy between the validation and test scores. This indicates a significant difference between the examples of the two sets or that the submissions were overfitted to the training and/or validation sets. Additionally, the validation score was an inconsistent predictor of which submission would achieve the highest test score. While submission vit2mrt-0.1.2_2_e46 attained the highest test score, it was outperformed by multiple submissions in terms of validation score. In fact, submission vit2mrt-0.1.3_5_e3 attained the

³CheXpert is available at: <https://stanfordmlgroup.github.io/competitions/chexpert/>

⁴MURA is available at: <https://nihcc.app.box.com/v/ChestXray-NIHCC>

⁵MURA is available at: <https://stanfordmlgroup.github.io/competitions/mura/>

⁶ROCO is available at: <https://github.com/razorx89/roco-dataset>

⁷Note that submission vit2mrt-0.1.2_2_e46 was a preliminary submission where the incorrect epoch was selected with early stopping due to a rounding error of the monitored metric score.

highest validation score, which had a configuration that employed ROCO pre-training and SCST, but no regularisation or MIT pre-training.

Using L2 regularisation during pre-training had a negative impact on performance, where both `mit2mrt-0.1.7_1_e0` and `mit2mrt-0.1.8_1_e1` produced worse validation and test scores than `mit2mrt-0.1.5_1_e1`. A regularisation term of 0.9 (`mit2mrt-0.1.8_1_e1`) was able to attain higher validation and test scores than a term of 0.5 (`mit2mrt-0.1.7_1_e0`); however, this could be due to submission `mit2mrt-0.1.8_1_e1` completing more epochs of fine-tuning, as shown in Figure 2.

The impact of the medical image multi-label classification task for MIT pre-training was inconclusive. Comparing submission `vit2mrt-0.1.1_5_e131` to `mit2mrt-0.1.9_1_e138`, using an MIT produced higher validation scores but lower test scores than ViT. Oppositely, employing an MIT over a ViT resulted in a lower validation score, but a higher test score—when comparing submission `vit2mrt-0.1.3_5_e3` to `mit2mrt-0.1.5_1_e1`. It should be emphasised that the medical images that largely make up the medical image multi-label classification task dataset are chest X-rays—whereas those in the task’s dataset, as well as ROCO, are more varied in modality and location. This suggests that the datasets used for the medical image multi-label classification task were not suited to the subsequent stages of pre-training and fine-tuning depicted in Figure 2.

The impact of using ROCO to pre-train the seq2seq models before fine-tuning was also inconclusive. Comparing submission `vit2mrt-0.1.1_5_e131` to `vit2mrt-0.1.2_3_e91`, ROCO pre-training increased the validation score and decreased the test score. However, pre-training on ROCO substantially reduced the number of epochs until convergence during fine-tuning. Note that medical image captioning datasets derived from PubMed Central articles have been previously criticised for their significant amount of noise [4]. This could mean that pre-training on ROCO may be harmful to performance.

SCST has been effective for medical image captioning previously [3]. Here, we note that SCST is sensitive and can be difficult to attain stable training. Comparing submission `vit2mrt-0.1.1_5_e131` to `vit2mrt-0.1.4_2_e0`, SCST substantially decreased the validation and test scores. This was likely due to the learning rate being too high for this configuration. Oppositely, SCST

Table 2

Results for each submission of the AEHRC CSIRO team for the ImageCLEFmed Caption 2021 task. The submissions are sorted by their test scores. The best validation and test scores are in boldface.

Submission identifier	Val. BLEU	Test BLEU
<code>vit2mrt-0.1.1_5_e131</code>	0.836	0.432
<code>mit2mrt-0.1.5_1_e1</code>	0.856	0.430
<code>vit2mrt-0.1.3_5_e3</code>	0.860	0.426
<code>vit2mrt-0.1.2_3_e91</code>	0.842	0.423
<code>mit2mrt-0.1.9_1_e138</code>	0.844	0.419
<code>mit2mrt-0.1.8_1_e1</code>	0.830	0.416
<code>vit2mrt-0.1.4_2_e0</code>	0.736	0.415
<code>mit2mrt-0.1.7_1_e0</code>	0.817	0.405
<code>vit2mrt-0.1.2_2_e46</code>	0.821	0.388

improved both validation and test scores when comparing submission vit2mrt-0.1.2_3_e91 to vit2mrt-0.1.3_5_e3, indicating that the learning rate was suitable for this configuration.

5. Conclusion

For ImageCLEFmed Caption 2021, the performance of submission vit2mrt-0.1.1_5_e131 placed the AEHRC CSIRO team third—with a score 0.078 worse than the first placed participant. This indicates that utilising pre-trained Transformer checkpoints to initialise the encoder and decoder of a seq2seq model is a promising approach for this task. However, the impact of the selected pre-training data was unclear; pre-training the seq2seq model with ROCO produced inconclusive results. Instead of ROCO, an image-caption dataset derived from real medical images and their associated radiologists' reports—such as MIMIC-CXR [23]—is recommended. The impact of the medical image multi-label classification task was also inconclusive, where its medical images were likely too dissimilar to those from ROCO and the task's dataset. The impact of SCST and L2 regularisation were clearer, with SCST providing a small improvement when configured correctly and the used L2 regularisation terms resulting in a decrease in performance. In future work, we aim to conduct a more thorough investigation of the proposed approach—to better adapt it to medical image captioning. At the same time, our overall approach has been intentionally kept general so that it might be applied to tasks other than medical image captioning.

References

- [1] O. Pelka, A. Ben Abacha, A. García Seco de Herrera, J. Jacutprakart, C. M. Friedrich, H. Müller, Overview of the ImageCLEFmed 2021 concept & caption prediction task, in: CLEF2021 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bucharest, Romania, 2021.
- [2] B. Ionescu, H. Müller, R. P'eteri, A. Ben Abacha, M. Sarrouiti, D. Demner-Fushman, S. A. Hasan, S. Kozlovski, V. Liauchuk, Y. Dicente, V. Kovalev, O. Pelka, A. G. S. de Herrera, J. Jacutprakart, C. M. Friedrich, R. Berari, A. Tauteanu, D. Fichou, P. Brie, M. Dogariu, L. D. Ştefan, M. G. Constantin, J. Chamberlain, A. Campello, A. Clark, T. A. Oliver, H. Moustahfid, A. Popescu, J. Deshayes-Chossart, Overview of the ImageCLEF 2021: Multimedia retrieval in medical, nature, internet and social media applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 12th International Conference of the CLEF Association (CLEF 2021), LNCS Lecture Notes in Computer Science, Springer, Bucharest, Romania, 2021.
- [3] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, M. Ghassemi, Clinically Accurate Chest X-Ray Report Generation, in: F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, J. Wiens (Eds.), Proceedings of the 4th Machine Learning for Healthcare Conference, volume 106 of *Proceedings of Machine Learning Research*, PMLR, Ann Arbor, Michigan, 2019, pp. 249–269. URL: <http://proceedings.mlr.press/v106/liu19a.html>.
- [4] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, A Survey on Biomedical Image Captioning,

- in: Proceedings of the Second Workshop on Shortcomings in Vision and Language, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 26–36. URL: <https://www.aclweb.org/anthology/W19-1803>. doi:10.18653/v1/W19-1803.
- [5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://www.aclweb.org/anthology/N19-1423>. doi:10.18653/v1/N19-1423.
- [6] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing (2020). [arXiv:2007.15779](https://arxiv.org/abs/2007.15779).
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, I. Polosukhin, Attention is All You Need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.
- [8] S. S. Yadav, S. M. Jadhav, Deep convolutional neural network based medical image classification for disease diagnosis, *Journal of Big Data* 6 (2019). doi:10.1186/s40537-019-0276-2.
- [9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision* 115 (2015) 211–252. doi:10.1007/s11263-015-0816-y.
- [10] S. Rothe, S. Narayan, A. Severyn, Leveraging Pre-trained Checkpoints for Sequence Generation Tasks, *Transactions of the Association for Computational Linguistics* 8 (2020) 264–280. doi:10.1162/tacl_a_00313.
- [11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2020). [arXiv:2010.11929](https://arxiv.org/abs/2010.11929).
- [12] W. Liu, S. Chen, L. Guo, X. Zhu, J. Liu, CPTR: Full Transformer Network for Image Captioning (2021). [arXiv:2101.10804](https://arxiv.org/abs/2101.10804).
- [13] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-Critical Sequence Training for Image Captioning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017. doi:10.1109/cvpr.2017.131.
- [14] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a Method for Automatic Evaluation of Machine Translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: <https://www.aclweb.org/anthology/P02-1040>. doi:10.3115/1073083.1073135.
- [15] T. Ridnik, E. Ben-Baruch, A. Noy, L. Zelnik-Manor, ImageNet-21K Pretraining for the Masses (2021). [arXiv:2104.10972](https://arxiv.org/abs/2104.10972).
- [16] R. J. Williams, D. Zipser, A Learning Algorithm for Continually Running Fully Recurrent Neural Networks, *Neural Computation* 1 (1989) 270–280. doi:10.1162/neco.1989.1.2.270.
- [17] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: International Conference on Learning Representations, 2019. URL: <https://openreview.net/forum?id=>

Bkg6RiCqY7.

- [18] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghighi, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, A. Y. Ng, CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, volume 33, Association for the Advancement of Artificial Intelligence (AAAI), 2019, pp. 590–597. doi:10.1609/aaai.v33i01.3301590.
- [19] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, PadChest: A large chest x-ray image dataset with multi-label annotated reports, *Medical Image Analysis* 66 (2020) 101797. URL: <https://www.sciencedirect.com/science/article/pii/S1361841520301614>. doi:10.1016/j.media.2020.101797.
- [20] X. Wang, Y. Peng, L. Lu, Z. Lu, R. M. Summers, TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2018. doi:10.1109/cvpr.2018.00943.
- [21] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, A. Y. Ng, MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs (2017). [arXiv:1712.06957](https://arxiv.org/abs/1712.06957).
- [22] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. M. Friedrich, Radiology Objects in COntext (ROCO): A Multimodal Image Dataset, in: D. Stoyanov, Z. Taylor, S. Balocco, R. Schnitman, A. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S.-L. Lee, S. Moriconi, V. Cheplygina, D. Mateus, E. Trucco, E. Granger, P. Jannin (Eds.), *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, Springer International Publishing, Cham, 2018, pp. 180–189.
- [23] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. ying Deng, R. G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific Data* 6 (2019). doi:10.1038/s41597-019-0322-0.