# CSIRO at ImageCLEFmedical Caption 2022

Leo Lebrat<sup>®,1,2</sup>, Aaron Nicolson<sup>®,3</sup>, Rodrigo Santa Cruz<sup>1,2</sup>, Gregg Belous<sup>3</sup>, Bevan Koopman<sup>3</sup> and Jason Dowling<sup>3</sup>

#### Abstract

We describe the participation of team CSIRO in the ImageCLEFmedical Caption task of 2022. This task consisted of two subtasks: concept detection and caption prediction. Concept detection involved identifying medical concepts within a given medical image. To accomplish this, we employed an ensemble of DenseNets with threshold tuning. CSIRO placed third amongst the participating teams with an F1 score of 0.447. For caption prediction, the task was to compose a coherent caption for a given medical image. We employed an encoder-to-decoder model with the Convolutional vision Transformer (CvT) as the encoder and DistilGPT2 as the decoder. CSIRO placed third amongst the participating teams with a BLEU score of 0.311.

## **Keywords**

Medical concept detection, Medical caption prediction, Multimodal learning

# 1. Introduction

Interpreting medical images is a complex and labour-intensive task. To become proficient requires a significant amount of training. A radiologist must be able to interpret medical concepts and their interplay from the image. On top of this, the workload of radiologists has increased significantly over the last couple of decades — mostly due to increases in crosssectional imaging and under-staffing [1, 2]. This leads to the need for an automated system that can produce textual descriptions of medical images. Such a system could improve the efficiency of interpretation and report creation — potentially reducing the workload and improving patient care. Such a system could also increase the diagnostic accuracy of non-specialist clinicians who have a lower diagnostic confidence [3, 4].

The ImageCLEFmedical Caption task of 2022 [5, 6] is a step in this direction with its two subtasks: concept detection and caption prediction. For these subtasks, participants were required to develop methods from the provided dataset. The dataset was formed from a largescale collection of figures from open access biomedical journal articles (PubMed Central). All

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

aaron.nicolson@csiro.au (A. Nicolson)

6 0000-0002-7163-1809 (A. Nicolson); 0000-0001-5577-3391 (B. Koopman); 0000-0001-9349-2275 (J. Dowling)

© 2022 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)

Equal contribution.

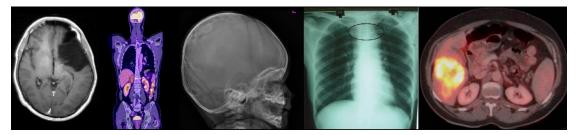
<sup>&</sup>lt;sup>1</sup>Imaging and Computer Vision Group, Commonwealth Scientific and Industrial Research Organisation, Pullenvale, Queensland, 4069, Australia.

<sup>&</sup>lt;sup>2</sup>Queensland University of Technology, Brisbane, Queensland, 4000, Australia.

<sup>&</sup>lt;sup>3</sup>Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, Herston, Queensland, 4006, Australia.

images in the dataset were accompanied by a caption, which form the labels for the caption prediction task. Unified Medical Language System (UMLS) concepts were extracted from each caption, forming the labels for the concept detection task.

In this article, we detail the methodology of our submissions for these two subtasks (Subsections 4.2 and 5.2). For concept detection, we make use of an ensemble of deep convolutional models. For caption prediction, we leverage encoder-to-decoder models, where the encoder is a computer vision model and the decoder is a natural language model. The remainder of this article includes a description of the two subtasks (Section 3), followed by a description and analysis of the data (Section 3). Following this, we describe the methodology and discuss the results for the concept detection task (Subsections 4.2 and 4.3). We then describe the methodology and discuss the results for the caption prediction task (Subsections 5.2 and 5.3) before concluding the article (Section 6).



**Figure 1:** A sample of images extracted from the dataset. The dataset includes an extensive range of modalities and anatomical regions, as well as both high and poor quality images. The licenses for the images from left to right are CC BY [7], CC BY-NC [8], CC BY [9], CC BY [10], and CC BY-NC [11].

# 2. Task Description

**Concept Detection Task**. A precursor to automatic medical image captioning is the identification of concepts in medical images.<sup>1</sup> This task is arduous given that the system must contend with 8 374 possible concepts. The concepts can be further applied for context-based image and information retrieval purposes.

**Caption Prediction Task**. Building upon the concept detection task, a system must not only detect concepts from a medical image, but also understand their interplay. From this understanding, the system must then compose a coherent natural language caption, akin to what a radiologist might write to describe their interpretation of an image.

# 3. Dataset Description & Analysis

Here, we describe the dataset for the concept detection and caption prediction subtasks. The dataset is a subset of the extended Radiology Objects in COntext (ROCO) dataset [12]. ROCO

<sup>&</sup>lt;sup>1</sup>Concepts are taken from some pre-defined medical terminology. In this case, the concepts are from the UMLS medical thesaurus.

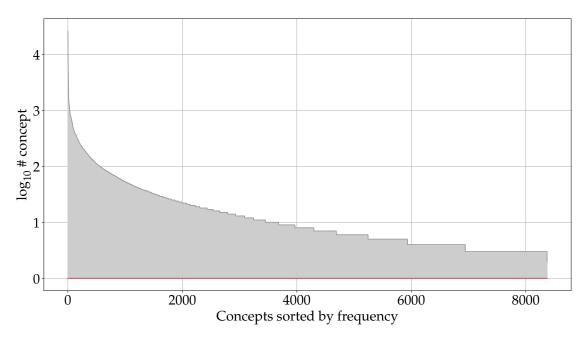


Figure 2: Number of occurrences of the concepts in the training split.

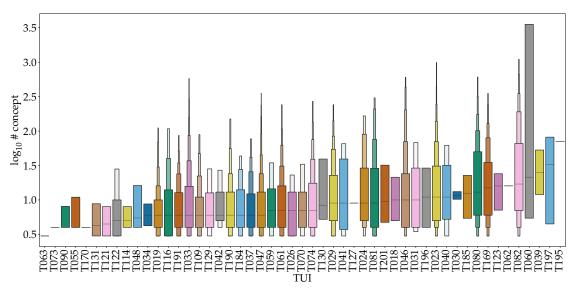


Figure 3: Number of occurrences of the concepts grouped by Type Unique Identifier (TUI).

originates from biomedical articles of the PubMed Central Open Access subset. The images of the dataset were split into training ( $n=83\,275$ ), validation ( $n=7\,645$ ), and test ( $n=7\,601$ ) sets. The concepts were generated using a reduced subset of the UMLS 2020 AB release for the concept prediction task, which includes the sections (restriction levels) 0, 1, 2, and 9. To

<sup>&</sup>lt;sup>2</sup>https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

**Table 1** Exemplars of common concepts.

Concept	Name	#	Concept	Name	#
	T060: Diagnostic Procedure		T080: Qualitative Concept		
C0040405	X-Ray Computed Tomography	25989	C0444706	Measured	1337
C1306645	Plain x-ray	24389	C0019409	Heterogeneity	958
C0024485	Magnetic Resonance Imaging	14622	C0392756	Reduced	831
C0041618	Ultrasonography	11147	C0442800	Enlarged	774
	T082: Spatial Concept			T130: Indicator or E	Diagnostic Aid
C0205131	Axial	3187	C0009924	Contrast Media	1406
C0238767	Bilateral	2722	C0016911	gadolinium	289
C0205129	Sagittal	2012	C1522485	Tracer	104
C0205091	Left	1696	C0013343	Dyes	43
	• • •			•••	

**Table 2** Exemplars of rare concepts.

Concept	Name	#	Concept	Name	#
	T019: Congenital Abnormality			T131: Hazardous or Poisonous Substa	ince
C0265905	Agenesis of pulmonary artery	3	C0037390	Snuff Tobacco	3
C2959359	Abnormal ventriculoarterial connection	3	C0142056	Asbestos, Serpentine	3
C0266035	Enamel pearls	3	C0556615	Paint thinners	3
	• • • • • • • • • • • • • • • • • • • •		C0003947	Asbestos	6
	T048: Mental or Behavioral Dysfunction		C0007018	carbon monoxide	10
C0006012	Borderline Personality Disorder	3		T063: Molecular Biology Research Tec	chniaue
C0016142	Firesetting Behavior	3			1
C0686346	Gender Dysphoria	3	C0920677	Gene Delivery Systems	3
	•••		C4725722	Second-strand Library Sequencing	3
	T122: Biomedical or Dental Material			T073: Manufactured Object	
C0181075	Bone graft - material	3	C0183336	Sleeve	4
C0011324	Dental Amalgam	3		T170: Intellectual Product	
	• • • • • • • • • • • • • • • • • • • •		C0205442	Eighth	4

improve the feasibility of recognising concepts from the images, concepts were further filtered based on their semantic type.

The images in this dataset are very diverse, as depicted in Figure 1. An average of  $4.74\pm2.72$  concepts are present for each image (the maximum number of concepts present in a single image is 50, and the minimum is 1). The concept distribution within the training dataset is highly skewed, as depicted in Figure 2, 93.36% concepts appear in less than 0.15% of the images. Concepts are grouped into higher-level semantic types; for example, concept "C0024485: Magnetic Resonance Imaging" is of type "T060: Diagnostic Procedure". As a result of the skewed count distribution, there is also an imbalance across Type Unique Identifiers (TUIs), as displayed in Figure 3. To illustrate this imbalance, we provide examples of common and rare concepts in Tables 1 and Table 2, respectively.

# 4. Concept Detection

In this section, we describe our participation pertaining the concept detection task. First, we describe the evaluation metric. Following this, we describe the methodology, as well as the model development on the task's dataset. Finally, we discuss the performance of our best model versus those of the other participants.

#### 4.1. Evaluation Metrics

The F1 score between the predicted and ground truth concepts was the primary metric for concept detection. This was calculated as follows:

- 1. The Python scikit-learn f1\_score function (v0.17.1-2) [13] was used to compute the F1 score between the predicted and ground truth boolean arrays. The default 'binary' averaging method was used.
- 2. All F1 scores were summed and averaged over the number of elements in the test set (7 601), giving the final score.

# 4.2. Methodology & Model Development

# 4.2.1. 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, 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  $\mathbb{R}^{3 \times 224 \times 224}$  and normalised according to the mean and standard deviation defined by the ImageNet checkpoints [14]. During training, the image was rotated at an angle sampled from  $\mathcal{U}[-50^\circ, 50^\circ]$  and randomly flipped horizontally or vertically with an independent probability of 0.4.

### 4.2.2. Architecture Selection

To select the architecture of the model for concept detection, we performed a grid search, as seen in Table 3. For this, we considered DenseNet-121 (7.2M parameters) [15], ResNet-18 (11.4M parameters) [16], and VGG-16 (134.7M parameters) [17]. Each model was trained for thirty hours using early stopping with either the stochastic gradient descent optimiser or Adam optimiser [18]. We also considered several learning rates from 0.01 to  $10^{-5}$ , as detailed in Table 3. Finally, we investigated three different initialisation procedures:

- 1. Freezing all the convolutional weights obtained from ImageNet and training only the last fully connected layers of the network (denoted by MLP in Table 3)
- 2. Training all weights from scratch (denoted by scratch in Table 3)
- 3. Initialising the weights with an ImageNet checkpoint and fine-tuning all of the network's weights (denoted by ImageNet in Table 3).

We observed that the best results were produced for the *Adam* optimiser [18] with a learning rate in the range  $[10^{-5}, 5 \cdot 10^{-4}]$ , and fine-tuning all the weights of DenseNet-161 initialised with

**Table 3**Architecture selection for varied learning rates and optimisers. A higher colour saturation indicates a better score. Green indicates the F1 score while blue indicates recall. Did Not Finish (DNF) indicates that a model did not converge during training.

Learning rate	0.	01	0.0	005	0.0	001	0.0	0005	0.0	001	0.00	0005	0.00	0001
	F1	Recall												
DenseNet MLP ADAM	0.2348	0.2096	0.2823	0.1993	0.3148	0.1992	0.3160	0.2000	0.3153	0.2000	0.3155	0.1990	0.3123	0.1949
DenseNet MLP SGD	0.2817	0.1674	0.2561	0.1487	0.1560	0.0850	0.0844	0.0442	0.0372	0.0229	DNF	DNF	DNF	DNF
DenseNet scratch ADAM	0.3106	0.1974	0.3076	0.1945	0.3191	0.2016	0.3215	0.2069	0.3210	0.2086	0.2255	0.1328	0.2255	0.1328
DenseNet scratch SGD	0.1654	0.0977	0.0980	0.0549	0.0988	0.0550	DNF							
DenseNet ImageNet ADAM	0.3095	0.1966	0.3126	0.1974	0.3238	0.2143	0.3283	0.2123	0.3289	0.2147	0.3310	0.2173	0.3290	0.2142
DenseNet ImageNet SGD	0.3156	0.1949	0.2514	0.1452	0.0552	0.0304	0.0863	0.0480	DNF	DNF	DNF	DNF	DNF	DNF
ResNet MLP ADAM	0.2393	0.1901	0.2816	0.1840	0.2982	0.1900	0.2991	0.1898	0.3020	0.1887	0.3003	0.1872	0.2954	0.1828
ResNet MLP SGD	0.2691	0.1599	0.2437	0.1414	0.1421	0.0775	0.1161	0.0625	0.0303	0.0157	DNF	DNF	DNF	DNF
ResNet scratch ADAM	0.2812	0.1751	0.2957	0.1839	0.2720	0.1689	DNF	DNF	0.2957	0.1858	0.2319	0.1420	0.3205	0.2046
ResNet scratch SGD	0.2799	0.1730	0.2251	0.1371	0.0480	0.0256	DNF	DNF	DNF	DNF	0.0010	0.4095	DNF	DNF
ResNet ImageNet ADAM	0.3142	0.1985	0.3109	0.1964	0.3229	0.2048	0.3243	0.2089	0.3258	0.2133	0.3273	0.2114	0.3267	0.2112
ResNet ImageNet SGD	0.3073	0.1889	DNF	DNF	DNF	DNF	DNF	DNF	0.0330	0.0235	DNF	DNF	DNF	DNF
VGG MLP ADAM	0.3002	0.2046	0.3040	0.2062	0.3114	0.2045	0.3097	0.1986	0.3121	0.1973	0.3118	0.1972	0.3108	0.1954
VGG MLP SGD	0.2978	0.1834	0.2917	0.1781	0.2590	0.1540	0.2253	0.1302	0.1763	0.0990	0.0767	0.0412	0.0021	0.0558
VGG scratch ADAM	DNF	0.2738	0.1665	0.2559	0.1550	0.2710	0.1669							
VGG scratch SGD	0.2804	0.1742	0.2510	0.1551	0.1477	0.0825	0.0534	0.0279	0.0013	0.3967	0.0440	0.0308	0.1315	0.0789
VGG ImageNet ADAM	DNF	DNF	DNF	DNF	DNF	DNF	0.3157	0.1967	0.3237	0.2046	0.3256	0.2069	0.3274	0.2088
VGG ImageNet SGD	0.3179	0.1992	0.3172	0.1978	0.3085	0.1920	0.2981	0.1844	0.1701	0.0945	DNF	DNF	DNF	DNF

an ImageNet checkpoint. We also benchmarked more recent architectures, such as EfficientNet-B7 [19] and RegNetY-8GF [20], but they did not offer an improvement in performance over DenseNet-161.<sup>3</sup>

### 4.2.3. Improving Performance on Underrepresented Concepts

The concept detection task is a multi-label classification problem. For this purpose, the final activation function of each model was set to the sigmoid function. Moreover, Binary Cross-Entropy (BCE) was used as the loss function. In order to improve the recall for the underrepresented classes in the training set described in Section 3, we experiment with three different approaches, namely, weighted BCE loss, preferential sampling, and threshold optimisation. We also investigated an ensemble of models.

**Weighted BCE loss**: Given a ground truth concept vector  $\mathbf{y} \in \{0, 1\}^N$  and a prediction vector  $\tilde{\mathbf{y}} \in [0, 1]^N$ , the **p**-weighted BCE loss is defined as:

$$\frac{1}{N} \sum_{i=1}^{N} \left( \mathbf{p}_{i} \mathbf{y}_{i} \log(\tilde{\mathbf{y}}_{i}) + (1 - \mathbf{y}_{i}) \log(1 - \tilde{\mathbf{y}}_{i}) \right). \tag{1}$$

We trialled different weighting vectors  $\mathbf{p}$ , ranging from continuous weighting that is inversely proportional to the frequency of the concept within the training set, to quantified weighting for

<sup>&</sup>lt;sup>3</sup>Each implementation and checkpoint is from https://github.com/pytorch/vision/tree/main/torchvision/models and the input images are re-sized to the architecture's requirement.

different percentiles of the count distribution that more heavily penalises mispredictions of rare concepts. However, we found through 10-fold cross-validated that weighted BCE loss did not consistently improve the F1 score.

**Preferential sampling**: For the second approach, we over-sampled training examples that contained rare concepts by providing them with a larger sampling probability. This was in an attempt to balance the number of times a model would observe each concept during training. However, we found through 10-fold cross-validation that preferential sampling did not consistently improve the F1 score.

**Threshold optimisation**: Given a model's prediction, one has to convert the vector of probabilities into a set of concepts. This is classically performed using thresholding (where a threshold of 0.5 is typically used for each class). In this challenge, we propose to fine-tune those thresholds based on the performance on a holdout dataset  $\mathcal{D}$ ; this is done by solving the following optimisation problem:

$$\underset{\mathbf{t} \in [0,1]^N}{\operatorname{arg}} \sum_{\mathbf{y}^i \in \mathcal{D}} \operatorname{F1-score} \left( \mathbb{1}_{\tilde{\mathbf{y}}^i \ge \mathbf{t}}, \mathbf{y}^i \right), \tag{2}$$

where  $\mathbb{1}_{\bullet \geq \mathbf{t}}$  is a piecewise comparison operator that returns 1 if the i-th entry of  $\bullet$  is greater or equal to  $\mathbf{t}_i$ , else, 0 is returned. Empirically, we noticed that optimising the threshold for rare concepts did not generalise well ('100% threshold optimisation' in Table 4), as a sufficient number of data points is required for a robust estimation. Following this, we instead optimised the threshold for concepts that appear sufficiently in the holdout dataset; for the concepts that occured in the Top-10% ('Top-10% threshold optimisation' in Table 4), as well as the concepts that occur in the Top-1% ('Top-1% threshold optimisation' in Table 4). We also compared setting the thresholds for each concept to a constant ('Fixed threshold' in Table 4).

#### 4.3. Results & Discussion

# 4.3.1. Final architectures

The models for our concept detection submissions were ensembles of DenseNets (where each was a DenseNet-161 initialised with the weights of an ImageNet checkpoint) trained according to the following recipe. We first start by combining the training and validation sets, splitting that into a holdout set used for threshold fine-tuning (10%). We randomly split the remaining data into an 80%/20% split for training/validation. We train all the parameters of each DenseNet-161 using the *Adam* optimiser with a batch-size of 8 and a learning rate of  $10^{-5}$ , along with early-stopping. The monitored metric was the validation F1-score. The average training time was 90 hours, where the selected epoch for each model was  $\approx 210$ . We form an ensemble of these models using majority voting. Finally our submissions, which use different thresholding and ensemble approaches, are presented in Table 4. Our best scoring submission was an ensemble of 43 DenseNet-161 models with Top-1% threshold optimisation.

# 4.3.2. Participant Rankings

The rankings amongst the participants of the concept detection task are shown in Table 5. We managed an F1 score of 0.447, placing us third amongst the participants. Despite our best

**Table 4**Submission history and associated F1 scores for each model.

Model	Threshold selection	F1
Ensemble of 43 DenseNet-161	Top- $1\%$ threshold optimisation	0.447
Ensemble of 43 DenseNet-161	Fixed threshold: 0.30	0.446
Ensemble of 11 DenseNet-161	Fixed threshold: 0.30	0.445
Ensemble of 11 DenseNet-161	Fixed threshold: 0.25	0.444
Ensemble of 5 DenseNet-161	Fixed threshold: 0.3	0.442
Ensemble of 5 DenseNet-161	Top- $10\%$ threshold optimisation	0.407
Ensemble of 5 DenseNet-161	100% threshold optimisation	0.406
Single DenseNet-161	Fixed threshold: 0.30	0.437
Single DenseNet-161	Fixed threshold: 0.50	0.433
Single DenseNet-161	100% threshold optimisation	0.396

efforts to account for rare concepts, our best model only predicted 107 out of the 8 734 available concepts on the test set, as shown in Tables A1 and A2. Some of the most commonly predicted concepts included modality (e.g., 'X-Ray Computed Tomography', 'Plain x-ray', and 'Magnetic Resonance Imaging'), body location (e.g., 'Chest', 'Abdomen', and 'Neck'), body part (e.g., 'Bone structure of cranium', 'Lower Extremity', and 'Pelvis'), and colour ('Yellow color', 'Green color', and 'Blue color'). One research direction that could be explored is a mixture of experts [21] with models that focus on different diagnostic procedures (e.g., TUI T060).

**Table 5**Final ranking for the concept detection task. A higher colour saturation indicates a better score.

Group Name	F1 Score	Secondary F1	Rank
AUEB-NLP-Group	0.451	0.791	1
fdallaserra	0.451	0.822	2
CSIRO	0.447	0.794	3
eecs-kth	0.436	0.856	4
vcmi	0.433	0.863	5
PoliMi-ImageClef	0.432	0.851	6
SSNSheerinKavitha	0.418	0.654	7
IUST_NLPLAB	0.398	0.673	8
Morgan_CS	0.352	0.628	9
kdelab	0.310	0.412	10
SDVA-UCSD	0.308	0.552	11

# 5. Caption Prediction

In this section, we first describe the evaluation metrics and the methodology for the caption prediction tasks. We then discuss the results of our models and end with a discussion about the performance of our best model versus those of the other participants.

**Table 6**Caption prediction metrics.

Metric	Official metric	Description
CLEF-BLEU	/	Average score of BLEU-1, BLEU-2, BLEU-3, and BLEU-4 [22].
CLEF-ROUGE-1	✓	ROUGE-n with unigrams [23].
CLEF-METEOR	✓	METEOR v1.5 [24].
CLEF-CIDEr	✓	CIDEr [25].
CLEF-SPICE	✓	SPICE [26].
CLEF-BERTScore	✓	BERTScore with microsoft/deberta-xlarge-mnli [27].
BLEU-1	×	BLEU-n with unigrams [22].
BLEU-2	×	BLEU- <i>n</i> with bigrams [22].
BLEU-3	×	BLEU- <i>n</i> with trigrams [22].
BLEU-4	×	BLEU-n with four-grams[22].
METEOR	×	METEOR v1.5 [24].
ROUGE-L	×	ROUGE with longest common subsequence-based statistics [28].
CIDEr	×	CIDEr [25].

### 5.1. Evaluation Metrics

The metrics for evaluating the caption prediction task are shown in Table 6. For the official metrics used by the competition organisers, which we designate as CLEF-\*, the following formatting was applied to the predicted and ground truth captions before evaluation:

- 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. **Lemmatization:** Lemmatization was next applied using spaCy's Lemmatizer (with spaCy model en\_core\_web\_1g).

For the remaining metrics, only the first two formatting steps were applied (lower-cased and remove punctuation), and only to the ground truth captions (the models were relied upon to generate captions with the correct formatting, as they were trained on ground truth captions with the same formatting). These non-official metrics were only applied to the validation set of the caption prediction task.

# 5.2. Methodology

First, we describe the image pre-processing, followed by the caption formatting and generation, the models, and model fine-tuning.

### 5.2.1. Caption Formatting and Generation

For the training and validation sets, the ground truth captions were converted to lower-case and had punctuation removed. When generating the captions during validation and testing, a beam search with a beam size of four and a maximum number of 128 subwords was used.

### 5.2.2. Encoder-to-Decoder Models

The encoder-to-decoder models investigated for caption prediction are listed below. The input to the encoder is a medical image. The output of the encoder is fed to the cross-attention module of the decoder, which then generates a caption in an autoregressive fashion. 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. It also should be noted that the image pre-processing that we used for caption prediction differs slightly to that used in Subsection 4.2.1. The image was instead resized to a size of  $\mathbb{R}^{3\times384\times384}$ . During training, the image was rotated at an angle sampled from  $\mathcal{U}[-5^{\circ}, 5^{\circ}]$  and no random horizontal or vertical flipping was applied.

- ViT2BERT ViT (86M parameters) is the encoder [29]. 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 uncased BookCorpus [30] and Wikipedia articles using self-supervised learning [31]. Both ViT and BERT are 12 layers with a hidden size of 768.
- **ViT2BERT** (**remove stopwords**) Identical to ViT2BERT, except that stopwords are additionally removed from the ground truth captions of the training and validation sets.
- **ViT2PubMedBERT** Identical to ViT2BERT, except that PubMedBERT (110M parameters) is the decoder. Its main difference to BERT is the pre-training data: uncased abstracts from PubMed (4.5B words) and articles from PubMed Central (13.5B words).
- **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 [32]. 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 [33]. It has three stages, with a combined 21 layers.
- **CvT2DistilGPT2** (retain aspect ratio) Identical to CvT2DistilGPT2, except that the image is first resized using bilinear interpolation so that its smallest side has 384 pixels and its largest side is set such that it maintained the aspect ratio. Next, the resized image is cropped to a size of  $\mathbb{R}^{3\times384\times384}$ . The crop location was random during training and centred during testing.
- **CvT2DistilGPT2·MIMIC-CXR** This is CvT2DistilGPT2 warm-started with a MIMIC-CXR checkpoint [34, 35]. The checkpoint was not additionally fine-tuned with reinforcement learning on MIMIC-CXR.
- **CvT2DistilGPT2·MIMIC-CXR (no. repeat n-gram size: 2)** Identical to CvT2DistilGPT2·MIMIC-CXR, except that a penalty was applied during caption

generation to the probability of tokens to prevent an n-gram from appearing more than once in a caption (the penalty was realised by setting a token's probability to zero). An n-gram size of two was used.

CvT2DistilGPT2·MIMIC-CXR (no. repeat n-gram size: 3) — Identical to CvT2DistilGPT2·MIMIC-CXR (no. repeat n-gram size: 2), except that an n-gram size of three was used.

### 5.2.3. Fine-tuning

Teacher forcing was used for fine-tuning [36]. Each model was implemented in PyTorch version 1.10.1 and trained with  $4\times NVIDIA$  P100 16GB 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 [37]; an initial learning rate of 1e-5 and 1e-4 for the encoder and all other parameters, respectively, following [38]. All other hyperparameters for AdamW were set to their defaults. A model's best epoch was selected using the highest validation BLEU-4 score. The epochs that were selected based on this criterion for each model were: epoch 5 for ViT2BERT and ViT2BERT (remove stopwords), epoch 7 for ViT2PubMedBERT and CvT2DistilGPT2, and epoch 8 for ViT2DistilGPT2 and CvT2DistilGPT2·MIMIC-CXR.

### 5.3. Results & Discussion

### 5.3.1. Model Performance

Here, we evaluate the performance of each encoder-to-decoder model to determine our best submission for the caption prediction task. The results of each encoder-to-decoder model are presented in Tables 7 (auxiliary metrics) and 8 (official metrics). One important consideration for this task is the multiple formatting steps applied to the predicted and ground truth captions before evaluation is performed with the official metrics, as described in Subsection 5.1. Hence, we wanted to determine if training with formatted ground truth captions is advantageous. Removing stopwords from the ground truth captions during training of ViT2BERT improved the validation CLEF-BLEU and CLEF-ROUGE-1 scores, as well as the test CLEF-BLEU score. However, it drastically decreased the test CLEF-BLEU score. Due to this, we abandoned this formatting strategy. Multiple strategies had a negligible impact on performance. This included the choice of encoder (ViT vs. CvT), as well as maintaining the aspect ratio of the medical images during image pre-processing.

When examining the performance of the decoders, BERT attained the highest test CLEF-BLEU score, while PubMedBERT scored the highest on test CLEF-ROUGE-1 (i.e., ViT2PubMedBERT vs. ViT2BERT and ViT2DistilGPT2). This indicates that the natural language understanding pre-training tasks of BERT and PubMedBERT are more transferable to the caption prediction tasks than the natural language generation pre-training strategies of DistilGPT2.

**Table 7**Caption prediction validation scores for each of the encoder-to-decoder models using the non-official metrics. (The row in grey indicates that the ground truth captions were different during evaluation and are thus uncomparable with the other rows). A higher colour saturation indicates a better score.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
ViT2BERT	0.223	0.124	0.066	0.038	0.091	0.202	0.192
⊢remove stopwords	0.155	0.080	0.041	0.021	0.078	0.169	0.230
ViT2PubMedBERT	0.228	0.126	0.068	0.039	0.091	0.204	0.203
ViT2DistilGPT2	0.214	0.118	0.064	0.036	0.087	0.196	0.203
CvT2DistilGPT2	0.215	0.119	0.064	0.036	0.087	0.198	0.202
→retain aspect ratio	0.215	0.119	0.065	0.037	0.088	0.198	0.208
CvT2DistilGPT2·MIMIC-CXR	0.221	0.122	0.067	0.039	0.090	0.201	0.213
→no. repeat n-gram size: 2	0.197	0.110	0.061	0.035	0.090	0.204	0.233
⇒no. repeat n-gram size: 3	0.204	0.115	0.063	0.037	0.092	0.205	0.229

Table 8

Caption prediction validation and test scores for each of the encoder-to-decoder models on the official metrics. Note that only the CLEF-BLEU and CLEF-ROUGE-1 scores were made available for each of the submissions, the remaining official metrics were only used with the best submission in Table 10. A higher colour saturation indicates a better score. Yellow designates scores on the validation set, while blue indicates scores on the test set.

Model	Valid	lation Set	Test Set		
	CLEF-BLEU	CLEF-ROUGE-1	CLEF-BLEU	CLEF-ROUGE-1	
ViT2BERT	0.004	0.182	0.311	0.181	
→remove stopwords	0.005	0.188	0.297	0.186	
ViT2PubMedBERT	0.004	0.188	0.309	0.188	
ViT2DistilGPT2	0.004	0.183	0.306	0.181	
CvT2DistilGPT2	0.005	0.181	0.309	0.182	
⇒retain aspect ratio	0.005	0.183	0.310	0.181	
CvT2DistilGPT2·MIMIC-CXR	0.006	0.188	0.310	0.181	
→no. repeat n-gram size: 2	0.005	0.195	0.308	0.197	
→no. repeat n-gram size: 3	0.006	0.194	0.311	0.197	

When comparing CvT2DistilGPT2 to CvT2DistilGPT2·MIMIC-CXR, it can be seen that warm-starting with the MIMIC-CXR checkpoint (a chest X-ray checkpoint) improved the validation score for each metric. As highlighted in Table 1, 'X-ray' was the second most represented modality in the dataset. This gives one reason as to why warm-starting with the MIMIC-CXR checkpoint was beneficial. However, the gains experienced on the validation set did not translate to an improvement in the test scores, with CvT2DistilGPT2 and CvT2DistilGPT2·MIMIC-CXR performing similarly.

When observing the captions generated by CvT2DistilGPT2·MIMIC-CXR, for example, in the first row in Table 9, there were repetitions of n-grams. This was evident in the generated captions of the other models as well. To mitigate this issue, we applied a penalty to the probabilities of the subword tokens in order to prevent an n-gram from being generated more than once, which is detailed in Subsection 5.2.2. It can be seen that an n-gram size of three successfully removed the repetitions. While this improved the validation and test CLEF-ROUGE-1 scores in Table 8, alarmingly, it had a minimal impact on the CLEF-BLEU score. This highlights the fragility of

BLEU — the metric did not penalise the score due to the repetitions. A more aggressive schema, i.e., an n-gram size of two, attained a test CLEF-ROUGE-1 score similarly to an n-gram size of three; however, the CLEF-BLEU score was reduced. Hence, CvT2DistilGPT2·MIMIC-CXR with no repetitions for an n-gram size of 3 was our best-performing caption prediction model.

**Table 9**Generated captions for test example *092563*. CvT2DistilGPT2·MIMIC-CXR would repeatedly generate the same six-gram if a penalty was not applied to the word token probabilities during generation.

Model	Generated caption
CvT2DistilGPT2·MIMIC-CXR	angiogram of the left subclavian artery occlusion of the left subclavian artery occlusion of the left common carotid artery occlusion of the left subclavian artery
CvT2DistilGPT2·MIMIC-CXR →no. repeat n-gram size: 2	angiogram of the left subclavian artery after stent implantation
CvT2DistilGPT2·MIMIC-CXR →no. repeat n-gram size: 3	angiogram of the left subclavian artery occlusion of the proximal leave anterior descend artery

### 5.3.2. Participant Rankings

The rankings amongst the participants of the caption prediction task are shown in Table 10. IUST NLP LAB attained the highest CLEF-BLEU and CLEF-METEOR scores, placing them first amongst the participants. However, their system produced the second-worst CLEF-CIDEr and CLEF-SPICE scores and their mean ranking over all the metrics was 6.2. This indicates that their performance, while optimal for CLEF-BLEU, did not generalise to the remaining metrics.

**Table 10**Caption prediction scores for the best submission of each participant. The ranking is determined by the participants' CLEF-BLEU scores. The mean rank of each participant over all the metrics is given in the last column. A higher colour saturation indicates a better score.

Participant	CLEF- BLEU	CLEF- ROUGE-1	CLEF- METEOR	CLEF- CIDEr	CLEF- SPICE	CLEF- BERTScore	Mean Rank
	DLEU	ROUGE-1	METEUR	CIDER	SPICE	DERISCORE	
IUST NLP LAB	0.483	0.142	0.093	0.030	0.007	0.561	6.2
AUEB-NLP-Group	0.322	0.166	0.074	0.190	0.031	0.599	5.0
CSIRO	0.311	0.197	0.084	0.269	0.046	0.623	2.2
vcmi	0.306	0.174	0.075	0.205	0.036	0.604	4.3
eecs-kth	0.292	0.116	0.062	0.132	0.022	0.573	7.7
fdallaserra	0.291	0.201	0.082	0.256	0.046	0.610	2.8
kdelab	0.278	0.158	0.074	0.411	0.051	0.600	4.3
Morgan CS	0.255	0.144	0.056	0.148	0.023	0.583	7.5
MAI ImageSem	0.221	0.185	0.068	0.251	0.039	0.606	5.0
SSN Sheerin Kavitha	0.160	0.043	0.023	0.017	0.007	0.545	10.0

Comparing our results (CSIRO) to that of the other participants, we attained the third-highest

CLEF-BLEU and CLEF-SPICE scores, the second-highest CLEF-ROUGE-1, CLEF-METEOR, and CLEF-CIDEr scores, and the highest CLEF-BERTScore. We also attained the highest mean ranking over all the metrics at 2.2. This suggests that our system, when considering all metrics, outperformed the system of IUST NLP LAB.

The mean ranking of *fdallserra* suggests that, in fact, their system was second best. Moreover, their system attained the highest CLEF-ROUGE-1 score. This highlights the importance of considering multiple metrics when evaluating natural language generation systems, as purely relying on a single metric, for example, CLEF-BLEU, can be misleading. We thus commend the organisers of the caption prediction task for expanding on the number of metrics from previous years. However, it should be noted that the model for each team was selected based on the best CLEF-BLEU score, which could bias the mean rank.

### 6. Conclusion

In this work, we detailed our participation in the concept detection and caption prediction subtasks of ImageCLEFmedical Caption 2022. For concept detection, we demonstrate the effectiveness of the ensemble approach, as well as the performance gains from threshold tuning. Despite our efforts, only a small portion of the concepts were predicted on the test set. This could be due to the fact that a vast amount of the concepts are underrepresented in the training set. For caption prediction, the important role that processing word token probabilities during generation can play was highlighted. Here, we used a penalty to prevent n-gram repetitions, which dramatically increased our CLEF-ROUGE-1 score. In future work, we aim to improve performance on the caption prediction task by leveraging the concept detection task, following the aim of the ImageCLEFmedical Caption challenge.

# Acknowledgement

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

# References

- [1] R. J. McDonald, K. M. Schwartz, L. J. Eckel, F. E. Diehn, C. H. Hunt, B. J. Bartholmai, B. J. Erickson, D. F. Kallmes, The Effects of Changes in Utilization and Technological Advancements of Cross-Sectional Imaging on Radiologist Workload, Academic Radiology 22 (2015) 1191–1198. doi:10.1016/j.acra.2015.05.007.
- [2] J. X. Liu, Y. Goryakin, A. Maeda, T. Bruckner, R. Scheffler, Global Health Workforce Labor Market Projections for 2030, Human Resources for Health 15 (2017). doi:10.1186/s12960-017-0187-2.
- [3] J. H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, J. Brink, Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success, Journal of the American College of Radiology 15 (2018) 504–508. doi:10.1016/j.jacr.2017.12.026.

- [4] A. Alexander, A. Jiang, C. Ferreira, D. Zurkiya, An Intelligent Future for Medical Imaging: A Market Outlook on Artificial Intelligence for Medical Imaging, Journal of the American College of Radiology 17 (2020) 165–170. doi:10.1016/j.jacr.2019.07.019.
- [5] B. Ionescu, H. Müller, R. Peteri, J. Rückert, A. Ben Abacha, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. Kozlovski, Y. D. Cid, V. Kovalev, L.-D. Ştefan, M. G. Constantin, M. Dogariu, A. Popescu, J. Deshayes-Chossart, H. Schindler, J. Chamberlain, A. Campello, A. Clark, Overview of the ImageCLEF 2022: Multimedia Retrieval in Medical, Social Media and Nature Applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 13th International Conference of the CLEF Association (CLEF 2022), LNCS Lecture Notes in Computer Science, Springer, Bologna, Italy, 2022.
- [6] J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2022 Caption Prediction and Concept Detection, in: CLEF2022 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Bologna, Italy, 2022.
- [7] J. A. da Silva, A. Alves, M. Talina, S. Carreiro, J. Guimarães, M. Xavier, Arachnoid cyst in a patient with psychosis: Case report, Annals of general psychiatry 6 (2007) 16–16. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933420/. doi:10.1186/1744-859X-6-16.
- [8] A.-K. Seidel, M. Pless, C. Michel, C. Soll, C. Hochuli, J. Gubler, A Rare Differential Diagnosis of an Adrenal Mass: A Case Report, Case reports in oncology 10 (2017) 981–986. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5731155/. doi:10.1159/000481501.
- [9] A. Al Kaissi, K. Klaushofer, F. Grill, Outward bulging of the right parietal bone in connection with fibrous dysplasia in an infant: a case report, Cases journal 1 (2008) 347–347. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2611982/. doi:10.1186/1757-1626-1-347.
- [10] H. Ni, A. Htet, Follicular thyroid carcinoma in a patient with myasthenia gravis and thymoma: a rare association, Ecancermedical science 6 (2012) 274–274. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3469247/. doi:10.3332/ecancer.2012.274.
- [11] S. Goja, M. K. Singh, R. J. Chaudhary, A. S. Soin, Robotic-assisted right hepatectomy via anterior approach for intrahepatic cholangiocarcinoma, Annals of hepato-biliary-pancreatic surgery 21 (2017) 80–83. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5449368/. doi:10.14701/ahbps.2017.21.2.80.
- [12] 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. Sznitman, 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.
- [13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.
- [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
- [15] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely Connected Convolutional

- Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017. doi:10.1109/cvpr.2017.243.
- [16] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016. doi:10.1109/cvpr.2016.90.
- [17] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: International Conference on Learning Representations, 2015.
- [18] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
- [19] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR, 2019, pp. 6105–6114.
- [20] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, P. Dollár, Designing network design spaces, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10428–10436.
- [21] S. Masoudnia, R. Ebrahimpour, Mixture of experts: a literature survey, Artificial Intelligence Review 42 (2014) 275–293.
- [22] K. Papineni, S. Roukos, T. Ward, W. 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.
- [23] C.-Y. Lin, E. Hovy, Automatic evaluation of summaries using N-gram co-occurrence statistics, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology NAACL '03, Association for Computational Linguistics, 2003. doi:10.3115/1073445.1073465.
- [24] S. Banerjee, A. Lavie, 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, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909.
- [25] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-Based Image Description Evaluation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- [26] P. Anderson, B. Fernando, M. Johnson, S. Gould, SPICE: Semantic Propositional Image Caption Evaluation, in: Computer Vision ECCV 2016, Springer International Publishing, 2016, pp. 382–398. doi:10.1007/978-3-319-46454-1\_24.
- [27] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating Text Generation with BERT, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.
- [28] C. Lin, F. J. Och, 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-04), Barcelona, Spain, 2004, pp. 605–612. URL: https://aclanthology.org/P04-1077. doi:10.3115/1218955.1219032.
- [29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. De-

- hghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv:2010.11929 [cs.CV] (2020). arXiv:2010.11929.
- [30] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, 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), 2015.
- [31] J. Devlin, M. 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.
- [32] A. Gokaslan, V. Cohen, OpenWebText Corpus, 2019. URL: http://Skylion007.github.io/ OpenWebTextCorpus.
- [33] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, L. Zhang, CvT: Introducing Convolutions to Vision Transformers, arXiv:2103.15808 [cs.CV] (2021). arXiv:2103.15808.
- [34] A. Nicolson, J. Dowling, B. Koopman, Improving Chest X-Ray Report Generation by Leveraging Warm-Starting (2022). arXiv:2201.09405.
- [35] A. Nicolson, J. Dowling, B. Koopman, Chest X-Ray Report Generation Checkpoints for CvT2DistilGPT2, 2022. doi:10.25919/64WX-0950.
- [36] 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.
- [37] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.
- [38] Z. Chen, Y. Song, T. Chang, X. Wan, Generating Radiology Reports via Memory-driven Transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020, pp. 1439–1449. doi:10.18653/v1/2020.emnlp-main.112.

# A. Predicted concepts

The 107 concepts predicted by our best model for the concept detection task are shown in Tables A1 and A2.

**Table A1**Concepts predicted by our final model on the test set.

Concept	TUI	TUI Description	Name	#predictions
C0040405	T060	Diagnostic Procedure	X-Ray Computed Tomography	2804
C1306645	T060	Diagnostic Procedure	Plain x-ray	1997
C0024485	T060	Diagnostic Procedure	Magnetic Resonance Imaging	1504
C0041618	T060	Diagnostic Procedure	Ultrasonography	1097
C0817096	T029	Body Location or Region	Chest	1076
C0000726	T029	Body Location or Region	Abdomen	718
C0002978	T060	Diagnostic Procedure	angiogram	514
C0037303	T023	Body Part, Organ, or Organ Component	Bone structure of cranium	306
C0023216	T023	Body Part, Organ, or Organ Component	Lower Extremity	280
C0205129	T082	Spatial Concept	Sagittal	261
C0221205	T080	Qualitative Concept	Yellow color	240
C0030797	T023	Body Part, Organ, or Organ Component	Pelvis	195
C0238767	T082	Spatial Concept	Bilateral	192
C0242485	T169	Functional Concept	Measurement	157
C1140618	T023	Body Part, Organ, or Organ Component	Upper Extremity	134
C0037949	T023	Body Part, Organ, or Organ Component	Vertebral column	107
C0046056	T109	Organic Chemical	fluorodeoxyglucose F18	97
C0205131	T082	Spatial Concept	Axial	96
C0225860	T023	Body Part, Organ, or Organ Component	Left atrial structure	87
C0205143	T082	Spatial Concept	Angular	85
C1699633	T060	Diagnostic Procedure	PET/CT scan	84
C0225897	T023	Body Part, Organ, or Organ Component	Left ventricular structure	82
C0226032	T023	Body Part, Organ, or Organ Component	Anterior descending branch of left	80
C0023884	T023	Body Part, Organ, or Organ Component	Liver	80
C0225883	T023	Body Part, Organ, or Organ Component	Right ventricular structure	76
C0223883	T039	Physiologic Function	Uptake	75
C0245144	T023	Body Part, Organ, or Organ Component	Right atrial structure	73
C0223644 C0035190	T201	Clinical Attribute	Residual volume	69
C0033130	T029	Body Location or Region	Neck	62
C0309093	T1029	Organic Chemical	FLAIR (product)	62
C0006104	T023	Body Part, Organ, or Organ Component	Brain	58
C0003104	T060	Diagnostic Procedure		58
C0032743 C0012751	T080	8	Positron-Emission Tomography Distance	55
C0205132	T082	Quantitative Concept	Linear	53
		Spatial Concept	Pleural effusion disorder	49
C0032227	T047	Disease or Syndrome		
C0444706	T080	Qualitative Concept	Measured	36
C0021102	T074	Medical Device	Implants	32
C1302256	T082	Spatial Concept	Apical four chamber view	31
C0332575	T033	Finding	Redness	29
C0034052	T023	Body Part, Organ, or Organ Component	Pulmonary artery structure	27
C0005400	T023	Body Part, Organ, or Organ Component	Bile duct structure	27
C0005682	T023	Body Part, Organ, or Organ Component	Urinary Bladder	25
C0332583	T080	Qualitative Concept	Green color	24
C1261316	T023	Body Part, Organ, or Organ Component	Right coronary artery structure	23
C1260957	T080	Qualitative Concept	Blue color	23
C0178602	T081	Quantitative Concept	Dosage	23
C0015965	T018	Embryonic Structure	Fetus	23
C0728985	T023	Body Part, Organ, or Organ Component	Cervical spine	23
C3827002	T033	Finding	Ground-glass opacities	23

**Table A2** (Continued) Concepts predicted by our final model on the test set.

Concept	TUI	TUI Description	Name	#predictions
C0521530	T047	Disease or Syndrome	Lung consolidation	18
C0040578	T023	Body Part, Organ, or Organ Component	Trachea	17
C0037775	T082	Spatial Concept	Spatial Distribution	15
C1302222	T082	Spatial Concept	Parasternal long axis view	15
C1881277	T081	Quantitative Concept	Isodose	14
C0003483	T023	Body Part, Organ, or Organ Component	Aorta	13
C0031039	T047	Disease or Syndrome	Pericardial effusion	12
C0025584	T023	Body Part, Organ, or Organ Component	Metatarsal bone structure	12
C0026266	T047	Disease or Syndrome	Mitral Valve Insufficiency	12
C0006141	T023	Body Part, Organ, or Organ Component	Breast	11
C0454199	T081	Quantitative Concept	Planning target volume	10
C0030274	T023	Body Part, Organ, or Organ Component	Pancreas	10
C0026608	T026	Cell Component	Motor Endplate	10
C0015813	T023	Body Part, Organ, or Organ Component	Head of femur	10
C4331911	T169	Functional Concept	M-Mode Ultrasound Mode	8
C0025062	T047	Disease or Syndrome	Mediastinal Emphysema	8
C0038536	T046	Pathologic Function	Subcutaneous Emphysema	7
C3829578	T033	Finding	Hypoechoic Focus	7
C0024109	T023	Body Part, Organ, or Organ Component	Lung	7
C0032005	T023	Body Part, Organ, or Organ Component	Pituitary Gland	6
C0024687	T023	Body Part, Organ, or Organ Component	Mandible	5
C0032326	T047	Disease or Syndrome	Pneumothorax	5
C1711105	T109	Organic Chemical	b-Hexachlorocyclohexane	5
C0014876	T023	Body Part, Organ, or Organ Component	Esophagus	5
C0018563	T023	Body Part, Organ, or Organ Component	Hand	5
C0000962	T023	Body Part, Organ, or Organ Component	Bone structure of acetabulum	5
C0024091	T023	Body Part, Organ, or Organ Component	Bone structure of lumbar vertebra	5
C0221198	T033	Finding	Lesion	5
C0025526	T023	Body Part, Organ, or Organ Component	Metacarpal bone	4
C0226054	T023	Body Part, Organ, or Organ Component	Right pulmonary artery	4
C0442119	T082	Spatial Concept	Intraoral approach	4
C0449381	T033	Finding	Observation parameter	4
C0222601	T023	Body Part, Organ, or Organ Component	Left breast	4
C0227613	T023	Body Part, Organ, or Organ Component	Right kidney	4
C0018800	T033	Finding	Cardiomegaly	3
C0205082	T033	Finding	Severe (severity modifier)	3
C0016504	T023	Body Part, Organ, or Organ Component	Foot	3
C0522510	T080	Qualitative Concept	With intensity	3
C0227614	T023	Body Part, Organ, or Organ Component	Left kidney	3
C0030647	T023	Body Part, Organ, or Organ Component	Patella	2
C0005847	T023	Body Part, Organ, or Organ Component	Blood Vessel	2
C0016642	T061	Therapeutic or Preventive Procedure	Fracture Fixation, Internal	2
C0030288	T023	Body Part, Organ, or Organ Component	Pancreatic duct	2
C0230431	T023	Body Part, Organ, or Organ Component	Structure of right knee	2
C0021852	T023	Body Part, Organ, or Organ Component	Intestines, Small	2
C0230461	T023	Body Part, Organ, or Organ Component		2
C0022742	T023	Body Part, Organ, or Organ Component	Knee	2
C0003956	T023	Body Part, Organ, or Organ Component	Ascending aorta structure	2
C0040508	T061	Therapeutic or Preventive Procedure	Total Hip Replacement (procedure)	2
C0013931	T061	Therapeutic or Preventive Procedure	Embolization, Therapeutic Bone structure of tibia	2
C0040184	T023	Body Part, Organ, or Organ Component		2
C0013303	T023	Body Part, Organ, or Organ Component	Duodenum Ontia Nama	2
C0029130	T023	Body Part, Organ, or Organ Component	Optic Nerve	1
C0205128	T082	Spatial Concept	Vertical	1
C1395409	T030	Body Space or Junction	Cephalometric nasion point	1
C1285498	T190	Anatomical Abnormality	Vegetation	1
C1295725	T082	Spatial Concept	Perpendicular axis	1
C0227481	T023	Body Part, Organ, or Organ Component	Right lobe of liver	1