NeuralDynamicsLab at ImageCLEFmedical 2022

Georgios Moschovis[1,3], Erik Fransén[1,2]

[1]KTH Royal Institute of Technology, Lindstedtsvägen 5, 114 28 Stockholm, Sweden
[2]Science for Life (SciLife) Laboratory, Tomtebodavägen 23A, Gamma 6, 171 65 Solna, Sweden
[3]Corresponding author

Abstract

Diagnostic Captioning is described as the automatic text generation from a collection of X-RAY images and it can assist inexperienced doctors and radiologists to reduce clinical errors or help experienced professionals to increase their productivity. Therefore, tools that would help doctors and radiologists produce higher quality reports in less time could be of high interest for medical imaging departments, as well as significantly impact deep learning research within the biomedical domain. With our participation in ImageCLEFmedical 2022 Caption evaluation campaign, we have attempted to address both concept detection and caption prediction tasks by developing baselines based on Deep Neural Networks; including image encoders, classifiers and text generators. Our group, *NeuralDynamicsLab* at KTH Royal Institute of Technology, within the school of Electrical Engineering and Computer Science, ranked 4th in the former and 5th in the latter task.

Keywords

Neural networks, Speech and language technology, Natural Language Processing (NLP), Deep learning, Generative deep networks, Convolutional neural networks (CNN), Text generation, Information retrieval, Diagnostic captioning, Image captioning, concept prediction, classification, image encoders, transformers, Encoder-Decoder architecture, abstractive summarization

1. Introduction

One of the most exciting technological aspects nowadays is Machine Learning's impressive potential in transforming the world we live in, primarily due to its exciting resurgence through Deep Learning (DL). The increasing size of biomedical data has allowed researchers demonstrate the evolving capabilities of Deep Learning in biomedical applications, through the development of advanced computing and imaging systems in biomedical engineering, machine learning-based biomedical data mining algorithms [1] and baselines for Diagnostic Captioning that has recently attracted researchers' attention, towards the goal of reducing the time required by a doctor or radiologist to produce medical texts and the amount of clinical errors, but also increasing the throughput of medical imaging departments [2].

In this work, we attempted to develop Diagnostic Captioning baselines, based on novel Deep Learning approaches, to investigate to what extent deep networks are capable of automatically

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
geomos@kth.se (G. Moschovis); erikf@kth.se (E. Fransén)
https://www.linkedin.com/in/georgios-moschovis-96428029/ (G. Moschovis); https://www.kth.se/profile/erikf/ (E. Fransén)
0000-0003-0547-0581 (G. Moschovis); 0000-0003-0281-9450 (E. Fransén)

CEUR Workshop Proceedings (CEUR-WS.org)

generating a diagnostic text from a set of medical images and how much their interpretation of medical images can assist doctors and radiologists produce better quality diagnoses; also at an increased throughput [2]. Towards this objective, the first step is concept detection that boils down to predicting relevant tags for X-RAY images, while the end goal is caption generation. In ImageCLEFmedical 2022 evaluation campaign, we experimented with addressing both concept detection and caption prediction tasks in order to get a quantitative measure of our proposed architectures' performance [3].

2. Dataset

In this section, we describe the data provided in ImageCLEFmedical 2022 evaluation campaign. Precisely, we provide details about the ImageCLEFmedical 2022 concept detection and caption prediction datasets that include images from different radiological image modalities but without including imaging modality information.

The dataset provided for both subtasks of ImageCLEFmedical 2022 evaluation campaign [4] consists of 90920 images that constitute a subset of the extended Radiology Objects in COntext (ROCO) dataset [5], without imaging modality information. As in previous editions, the dataset originates from biomedical articles of the PMC OpenAccess subset. After merging the initially provided train and validation data, we shuffle them after manually setting the seeds to eliminate randomness in consecutive runs while tuning our hyperparameters and then keep 80% as our training set, 10% as our validation set used for hyperparameter tuning and the remaining 10% as our development set used for model selection. Since the dataset is large we perform neither cross-validation nor data-augmentation. We experimented with adding noise to the images, in the form of random rotations and translations, which however did not provide any additional benefit in our baselines' quantitative evaluation.

Regarding the concept detection subtask, there are 8374 tags of concepts that are assigned to the X-RAY images, while each image in any of the training, validation or development set is assigned 5 tags on average. Regarding the caption prediction subtask, the total number of captions in the training set is 72736, the total number of unique captions is 70879 and the average caption length is 108 words, including 28 unique words. In the validation set the total number of captions is 9092, the total number of unique captions is 8984, the average caption length is 107 words, including 26 unique words. In the development set the total number of captions is 9092, the total number of unique captions is 8977 and the average caption length is 108 words, including 28 unique words. These counts verify that the aforementioned sets are balanced in terms of their statistics.

"The concepts were generated using a reduced subset of the Unified Medical Language System (UMLS) 2020 AB release, which includes the sections (restriction levels) 0, 1, 2, and 9". [4] The UMLS is a set of files and software that collects multiple health and biomedical vocabularies and standards to enable interoperability between computer systems. To improve the feasibility of recognizing concepts from the images, concepts were filtered based on their semantic type and concepts with very low frequency were removed. In each caption, tokens containing numbers and all punctuation were removed, captions were converted to lower-case and lemmatization was applied using spaCy toolkit [3].

3. Methods and results

In this section, we describe the core components of the methods utilized to encode the X-RAYs with dense embeddings in our work and explain in detail the baseline networks that we proposed in ImageCLEFmedical 2022 evaluation campaign, in order of performance, for both subtasks that are based on the aforementioned core components that rely on pre-trained architectures, extremely popular in computer vision.

Precisely, we provide details about the ImageCLEFmedical 2022 concept detection and caption prediction datasets and on how we designed backbone networks as generic image encoders that rely on Convolutional Neural Networks (CNN) architectures that are popular for vision tasks on generic images, such as classification and semantic segmentation, while they are shared within all baselines, in both ImageCLEFmedical Caption tasks. Furthermore, we describe the components of each model and give details on the selected hyper-parameters. For all our models, we have set in advance all the random seeds equal to 0, the CUDNNs backends as deterministic and disabled the CUDNNs backends benchmark to ensure consistency of the aforementioned splits in consecutive runs for hyper-parameter selection. This procedure has been applied for both subtasks of the evaluation campaign.

3.1. Backbone Networks: image encoders

One of the principal components in the proposed architectures that is shared for both subtasks includes the image encoders. They constitute existing state-of-the-art architectures, pretrained on ImageNet classification dataset [6], which are obtained from `torchvision` models library to perform inference, while any additional components such as a multi-label classification head or a caption generation architecture are appended to the output of the image encoder; in this content these models are referred to as “backbone networks”. The goal of these networks is to encode the images into dense numerical representations. Since Deep Learning became popular and what is called the Deep Learning Community was given birth, different initialization strategies for the weights and the biases were proposed. We used Glorot initialization shown below [7] to initialize the weights of the classification heads and experimented with non-pretrained image encoders that we initialized using the same strategy and fully-finetuned them, their performance however was inferior in concept prediction.

$$\text{Glorot: } W_{i,j} \sim \mathcal{U}\left(-\sqrt{\frac{6}{f_{\text{in}} + f_{\text{out}}}}, \sqrt{\frac{6}{f_{\text{in}} + f_{\text{out}}}}\right)$$

Some Convolutional Neural Network (CNN) encoders that have been attempted to use include variants of AlexNet [8], ResNet [9], DenseNet [10], VGG [11] and EfficientNet [12], which are obtained from `torchvision` models library as mentioned above. We also experimented with another architectural choice that is Vision Transformers (ViT) [13], the performance obtained was poor however compared to CNN encoders. That outcome is in line with the observation in [14] that Vision Transformers and "Hybrid-ViT architectures are inferior to the CNN-based ones". The above summarize the first step in the design of image encoders that is model selection based on their performance on a development set.

Moreover, model selection shall be followed by a model collaboration design principle, based on ensemble learning. In this case, we have used the aforementioned models as *members of the ensemble* or *weak learners* in a pool of encoders trained with different parameter values (e.g. learning rates, decision thresholds for the positive class, number of epochs), as well as based on different architectures, to seek for diversity and exploit the "Wisdom of the crowd" [15] for the fine-tuned models. In this context, we take into consideration the “votes” of all the different CNNs by averaging their outputs to make decisions on the generated tags or make guesses on the assigned captions.

3.2. Concept prediction subtask

As mentioned in section 3.1 “backbone networks” refer to image encoders, which are state-of-the-art architectures, pretrained on ImageNet classification dataset [6], shared for both subtasks. In the case of concept prediction, an additional classification head that is either a Perceptron or a Multi-layered Perceptron was added on top of these “backbone networks" and its weights were initialized using Glorot initialization strategy [7].

3.2.1. Pre-trained DenseNet161 with fine-tuned classification head, learning rate 10^{-3}, Adam optimizer and gradient clipping

The first two models correspond to a DenseNet161 convolutional network that is pretrained on ImageNet classification dataset and its head is a Perceptron, which is further fine-tuned on the ImageCLEFmedical 2022 data using sigmoid activation function in the output units that equal the number of concepts -thus 8374 nodes, a constant learning rate equal to 10^{-3} and the negative F_1 score as a minimization criterion. For each image, we assign it the concepts that have predicted probabilities above 50%, while the tags obtain their numerical IDs in their order of appearance before shuffling them. Furthermore, we clip the gradients computed during training to be in $[-1, 1]$, to increase numerical stability.

When performing stochastic or minibatch Gradient Descent, and the loss changes quickly at one direction and slowly at another, Gradient Descent will progress slowly along the shallow dimension and jitter along the steep one. To overcome this issue, we used Adam optimizer [16], so that progress along steep directions is damped and meanwhile progress along flat directions is accelerated. Adam uses exponentially decaying average to discard history but also momentum as an estimate of the first-order gradient. It has bias corrections for first-order and second-order moments and converges rapidly after finding a local convex bowl. If t represents the current time step, Adam updates are equal to:

$$\begin{aligned}
\mathbf{w}^{(t+1)} &= \mathbf{w}^{(t)} - \epsilon \frac{\mathbf{v}^{(t)}}{\delta + \sqrt{\mathbf{r}^{(t)}}}, \delta, \epsilon \in \mathbb{R}^{+} \\
\mathbf{v}^{(t+1)} &= \rho_1 \mathbf{v}^{(t)} + (1 - \rho_1)\mathbf{g}^{(t)}, \rho_1 \in \mathbb{R}^{+} \\
\mathbf{r}^{(t+1)} &= \rho_2 \mathbf{r}^{(t)} + (1 - \rho_2)\left(\mathbf{g}^{(t)}\right)^2, \rho_2 \in \mathbb{R}^{+}
\end{aligned}$$

Our best performing model (with submission ID 181750) is an instance of the aforementioned architecture trained in all the provided data, thus after merging again the training, validation and development sets that are described in section 2 and achieves $F_1 = 0.43601$. The next model corresponds to the same network architecture but is trained only in training set (with submission ID 181715) and achieves a score $F_1 = 0.43567$. For the latter case, where we have measured performance in all sets, we present plots with the evolution of F_1 score and accuracy during training in Figure 1(a).

3.2.2. Pre-trained DenseNet161 with fine-tuned classification head, learning rate 5×10^{-4}, AdamW optimizer and gradient clipping

The next model corresponds to another DenseNet161 convolutional network that is pretrained on ImageNet classification dataset and its head is a Perceptron, which is further fine-tuned on the ImageCLEFmedical 2022 data using sigmoid activation function in the output units that equal the number of concepts -thus 8374 nodes, a constant learning rate equal to 5×10^{-4} and the negative F_1 score as a minimization criterion. For each image, we assign it the concepts that have predicted probabilities above 50%, while the tags obtain their numerical IDs in their order of appearance before shuffling them. Furthermore, we clip the gradients computed during training to be in $[-1, 1]$, to ensure numerical stability.

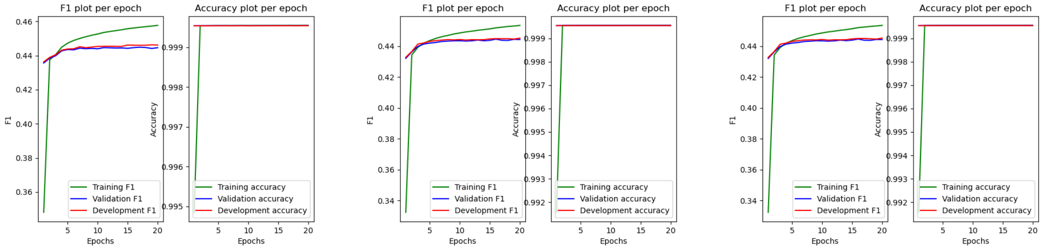
In this occasion we have used an improved version of Adam optimizer, called AdamW [17], where weight decay is performed only after controlling the parameter-wise step size and thus yields models that generalize much better. Compared to Adam optimizer that we discussed in section 3.2.1, as well as other adaptive gradient algorithms, where the potential benefit of weight decay regularization is limited because "the weights do not decay multiplicatively but by an additive constant factor" [17], AdamW optimizer may overcome this issue, while training much faster than stochastic or minibatch Gradient Descent.

Our model is an instance of the aforementioned network architecture, it is trained only in training set (with submission ID 181753) and achieves a score $F_1 = 0.43558$, although we would expect training with AdamW to perform better. Since the gain of re-training the model after merging all the splits is almost negligible, as we already noticed in section 3.2.1, the remaining models are not re-trained in the entire dataset. Once again, we present plots with the evolution of F_1 score and accuracy in Figure 1(b).

3.2.3. Pre-trained DenseNet161 with fine-tuned classification head, learning rate 5×10^{-4} and Adam optimizer

The subsequent model is yet another DenseNet161 convolutional network that is pretrained on ImageNet classification dataset and its head is a Perceptron, which is further fine-tuned on the ImageCLEFmedical 2022 data using sigmoid activation function in the output units that equal the number of concepts -thus 8374 nodes, a constant learning rate equal to 5×10^{-4} and the negative F_1 score as a minimization criterion. For each image, we assign it the concepts that have predicted probabilities above 50%, while the tags obtain their numerical IDs in their order of appearance before shuffling them and train the network using Adam optimizer; as we have excessively described in section 3.2.1.

Our model is an instance of the aforementioned network architecture (with submission ID 182152) and achieves a score $F_1 = 0.43539$, however, in this baseline we omit clipping the gradients, in contrast with the models described above in sections 3.2.1 and 3.2.2. Furthermore, as for both previous best-performing models we present plots with the evolution of F_1 score and accuracy below in Figure 1(c).



(a) (b) (c)

Figure 1: F_1 and accuracy scores plots per epoch for the models described (a) in section 3.2.1, (b) in section 3.2.2, as well as (c) in section 3.2.3. We observe that the classifications heads, which we finetune on ImageCLEFmedical 2022 data, appear to be sufficiently regularized (thus there is no overfitting) and to have used their maximum capacity.

3.2.4. Ensemble of pre-trained DenseNet CNNs with fine-tuned classification heads

The proceeding model and the best performing mixture of individual networks corresponds to the 10 best performing DenseNet CNNs, including instances of DenseNet161 and DenseNet121 architectures, and indicates our quest for diversity and to consequently exploit the "Wisdom of the crowd" [15] notion.

In this context, we take into account the “votes” of all the different CNNs to make decisions on the assigned tags. The voting scheme consists of averaging the probabilities computed by the different weak learners before assigning to each image the concepts that have average predicted probabilities above 50%, while the tags as usual obtain their numerical IDs in their order of appearance before shuffling them. We also experimented with using alternative voting policies, such as computing the union or intersection of the assigned tags by each weak learner, where assignments are defined by the predicted probabilities being above 50%, in the pool of finetuned networks, but they performed poorly.

Table 1 summarizes the architecture of all individual networks in the pool of encoders. This includes the type of Backbone Network, the optimizer, the value of learning rate and whether it is decaying per epoch, the batch size and the submission ID of the individual network, for the aforementioned weak learners in sections 3.2.1, 3.2.2, 3.2.3 that performed better than the ensemble altogether and thus were submitted individually. Note that the classification head is always a Perceptron which is further fine-tuned in the ImageCLEFmedical 2022 data using sigmoid activation function in the output units that equal the number of concepts. Moreover, when linear decay is applied, the learning rate is updated by: $\eta_{t+1} = \eta_0 \times \frac{1-t}{T}$ where t represents the current time step, T the total number of epochs and η_0 is the learning rate at the beginning of training procedure. The performance of this mixture of experts (with submission ID 182338) equals $F_1 = 0.43496$.

Table 1
Summary of weak learners' architecture and training regime in model 182338

Backbone Net.	Optimizer	Learning Rate	Linear Decay	Batch size	Epochs	Subm. ID
DenseNet121	AdamW	5×10^{-4}	False	60	20	-
DenseNet121	AdamW	10^{-3}	False	60	20	-
DenseNet121	AdamW	10^{-4}	False	60	20	-
DenseNet161	Adam	10^{-3}	False	120	20	181750, 181715
DenseNet161	AdamW	10^{-3}	True	120	20	-
DenseNet161	Adam	5×10^{-4}	False	120	20	-
DenseNet161	Adam	5×10^{-4}	False	120	20	181753
DenseNet161	AdamW	5×10^{-4}	False	120	20	182152
DenseNet161	AdamW	10^{-4}	False	120	50	-
DenseNet161	AdamW	10^{-4}	False	120	20	-

3.2.5. Ensemble of pre-trained DenseNet CNNs with fine-tuned classification heads

Although Dense Convolutional Networks (DenseNet CNNs) appear to outperform other network architectures, which is in line with their extensive use in biomedical applications that include X-RAYs processing [18], we also experimented with a plethora of CNNs backbone networks as we have mentioned in section 3.1. Consequently, the ensuing three models constitute ensembles that include different architectures within their members, with varying hyperparameter values to encourage diversity of training regimes. During the voting process we average the probabilities computed by the softmax layer of all different week learners before assigning to each image the tags that have average predicted probabilities above 50%.

Table 2
Summary of weak learners' architecture and training regime in model 181546

Backbone Net.	Optimizer	Learning Rate	Linear Decay	Batch size	Epochs	Subm. ID
AlexNet	AdamW	10^{-4}	False	60	20	-
AlexNet	AdamW	5×10^{-5}	False	60	20	-
DenseNet121	AdamW	5×10^{-4}	False	60	20	-
DenseNet121	AdamW	10^{-3}	False	60	20	-
DenseNet121	AdamW	10^{-4}	False	60	20	-
DenseNet161	Adam	10^{-3}	False	120	20	181750, 181715
DenseNet161	AdamW	10^{-3}	True	120	20	-
DenseNet161	Adam	5×10^{-4}	False	120	20	-
DenseNet161	Adam	5×10^{-4}	False	120	20	181753
DenseNet161	AdamW	5×10^{-4}	False	120	20	182152
ResNet50	AdamW	10^{-4}	False	60	20	-
ResNet101	AdamW	10^{-4}	False	60	20	-
VGG-13	AdamW	10^{-4}	False	60	20	-
VGG-16	AdamW	10^{-4}	False	60	20	-

Table 3
Summary of weak learners' architecture and training regime in model 182155

Backbone Net.	Optimizer	Learning Rate	Linear Decay	Batch size	Epochs	Subm. ID
AlexNet	AdamW	10^{-4}	False	60	20	-
AlexNet	AdamW	5×10^{-5}	False	60	20	-
DenseNet121	AdamW	5×10^{-4}	False	60	20	-
DenseNet121	AdamW	10^{-3}	False	60	20	-
DenseNet161	Adam	10^{-3}	False	120	20	181750, 181715
DenseNet161	AdamW	10^{-3}	True	120	20	-
ResNet50	AdamW	10^{-4}	False	60	20	-
ResNet50	AdamW	5×10^{-5}	False	60	20	-
ResNet101	AdamW	10^{-4}	False	60	20	-
ResNet101	AdamW	5×10^{-4}	False	60	20	-
VGG-13	AdamW	10^{-4}	False	60	20	-
VGG-13	AdamW	5×10^{-5}	False	60	20	-
VGG-16	AdamW	10^{-4}	False	60	20	-
VGG-16	AdamW	5×10^{-5}	False	60	20	-

Table 4
Summary of weak learners' architecture and training regime in model 182154

Backbone Net.	Optimizer	Learning Rate	Linear Decay	Batch size	Epochs	Subm. ID
AlexNet	AdamW	10^{-4}	False	60	20	-
AlexNet	AdamW	5×10^{-5}	False	60	20	-
DenseNet121	AdamW	5×10^{-4}	False	60	20	-
DenseNet121	AdamW	10^{-3}	False	60	20	-
DenseNet121	AdamW	10^{-4}	False	60	20	-
DenseNet161	Adam	10^{-3}	False	120	20	181750, 181715
DenseNet161	AdamW	10^{-3}	True	120	20	-
DenseNet161	Adam	5×10^{-4}	False	120	20	-
ResNet50	AdamW	10^{-4}	False	60	20	-
ResNet101	AdamW	10^{-4}	False	60	20	-
VGG-13	AdamW	10^{-4}	False	60	20	-
VGG-16	AdamW	10^{-4}	False	60	20	-

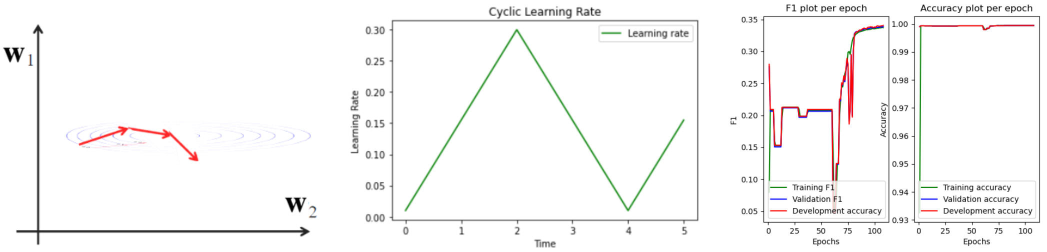
Our three following mixtures of experts (with submission IDs 181546, 182155, 182154) and achieve a score $F_{1,1} = 0.43404$, $F_{1,2} = 0.43130$, $F_{1,3} = 0.42957$ respectively. Tables 2, 3, 4 summarize the architecture of all individual networks in each pool of encoders. Their format is identical to that used in section 3.2.4 and consequently they also refer to the hyper-parameter values for each of the weak learners.

Note that the classification head is always a Perceptron which is further fine-tuned in the ImageCLEFmedical 2022 data using sigmoid activation function in the output units that equal the number of concepts. Moreover, when linear decay is applied, the learning rate is updated by: $\eta_{t+1} = \eta_0 \times \frac{1-t}{T}$ where t represents the current time step, T the total number of epochs and η_0 is the initial learning rate.

3.2.6. Fully fine-tuned DenseNet161 with cyclical learning rate and AdamW optimizer

The succeeding model corresponds to a DenseNet161 convolutional network that is now fully-finetuned on the ImageCLEFmedical 2022 data using sigmoid activation function in the output units that equal the number of concepts -thus 8374 nodes, scheduled learning rate [19] and the negative F_1 score as a minimization criterion. For each image, we assign it the concepts that have predicted probabilities above 50%, while the tags obtain their numerical IDs in their order of appearance before shuffling them.

One important aspect of minibatch or stochastic gradient descent relates to the choice of the learning rate η that controls the size of the update, which will occur to the gradients in every iteration. Constant learning rates have been traditionally used to train Deep Neural Networks based on back-propagation algorithm, although do not guarantee optimal convergence rate according to the Stochastic Approximation Theory [20], precisely the network parameters hover around a minimum at an average distance proportional to the learning rate and to a variance that is dependent on the objective function and the exemplar set [21]. To this end, cyclical learning rates have been proposed as a new method for setting the learning rate by cyclically varying its value between reasonable boundary values, which increases classification accuracy when training CNNs with generic images [22].



(a) (b) (c)

Figure 2: (a) Schematic illustration of the error landscape with a high learning rate, (b) example plot of a cyclical learning rate with $\eta_{\text{min}} = 0.01$, $\eta_{\text{max}} = 0.30$, $n_s = 2$ and (c) F_1 and accuracy scores plots per epoch for the model described in section 3.2.6.

A high value of η will make the network make large steps above the minimum of the error function but never converge to it, as illustrated in Figure 2(a). A small value of η will delay convergence, preventing the network to find a minimum of the error function if the number of epochs is limited. A cyclical learning rate linearly ranges between two values η_{min} and η_{max}. One maximization of the learning rate followed by a minimization is called a cycle. In Figure 2(b) hereunder we present an example of cyclical learning rate, where $\eta_{\text{min}} = 0.01$, $\eta_{\text{max}} = 0.30$, $n_s = 2$ and we denote as $2n_s$ the time required for a cycle of our learning rate to complete. In our model we set $\eta_{\text{min}} = 10^{-5}$, $\eta_{\text{max}} = 0.1$, $n_s = 4$ for the first 80 epochs and then set it to a constant value $\eta = 10^{-3}$ for 30 additional epochs.

This network (with submission ID 182156) achieves a score $F_1 = 0.31687$, which is a rather lower score compared to the pre-trained models on ImageNet classification dataset [6], achieving more than 10% higher F_1 results on the test set. Moreover, we present plots with the evolution of F_1 score and accuracy per training epoch of the model in Figure 2(c) that is quite unstable while varying the learning rate.

3.2.7. Nearest Neighbours Baseline

The ensuing model is a generalization of the 1-NN baseline proposed in [23]. We further either remind or inform the reader that for every image in the test set, the 1-NN baseline assigns the tags of the visually most similar image from the training set as the output and consequently for every image, $\hat{x}$, in the test set, the 1-NN baseline will output the set of concepts, say y^*, of the most similar image, say x^*, from the training set as output [2]. Therefore, if we denote by $\mathbf{e}(.)$ the output of the employed image encoder among those mentioned in section 3.1, 1-NN predicts $(\hat{x}, \hat{y}) = (\hat{x}, y^*)$ that satisfies $(x^*, y^*) = \arg\min_{x^*} \cos\left(\mathbf{e}(\hat{x}), \mathbf{e}(x^*)\right)$. Our generalized Nearest Neighbours baseline takes into account $k \in \mathbb{Z}^+$ neighbours instead and not necessarily only the one with closest representation. Our model (with submission ID 182331) uses $k = 1$ with a VGG-16 encoder pre-trained on ImageNet classification dataset and achieves only $F_1 = 0.25061$ that indicates the importance of fine-tuning.

3.3. Performance summary

Table 5 below summarizes several characteristics of the proposed baselines for concept detection, in order of performance with respect to F_1 scores, together with their respective submission IDs. We observe that DenseNet161 image encoders with finetuned classification heads are the top performing configurations and outperform other CNN architectures, which is in accordance with their extensive use X-RAYs processing [18], while fully finetuning the backbone networks and using retrieval based heuristics that capture representations' similarities, such as the 1-NN baseline [23], achieve lower scores.

Table 5
Summary of our configurations' characteristics and statistics

Backbone Network	Section described	Type of model	F1 scores	Submission ID
DenseNet161	Section 3.2.1	Deep Network Head	0.43601	181750
DenseNet161	Section 3.2.1	Deep Network Head	0.43601	181750
DenseNet161	Section 3.2.2	Deep Network Head	0.43558	181753
DenseNet161	Section 3.2.3	Deep Network Head	0.43539	182152
DenseNet variants	Section 3.2.4	Ensemble of Networks	0.43496	182338
Various networks	Section 3.2.5	Ensemble of Networks	0.43404	181546
Various networks	Section 3.2.5	Ensemble of Networks	0.43130	182155
Various networks	Section 3.2.5	Ensemble of Networks	0.42957	182154
DenseNet161	Section 3.2.6	Deep Network (full)	0.31687	182156
VGG-16	Section 3.2.7	Nearest Neighbour	0.25061	182331

3.4. Caption generation subtask

In ImageCLEFmedical 2022 evaluation campaign, "the first step to automatic image captioning and scene understanding boils down to identifying the presence and location of relevant concepts within a large corpus of medical images that is followed by caption generation in captioning. Based on medical images content, the concept prediction task provides the building blocks for scene understanding by identifying the individual components, referred to as image tags, from which captions are composed. The assigned concepts can be further applied for context-based image and information retrieval purposes" [3].

"On the basis of the vocabulary $\mathcal{V}$ identified during concept prediction task, as well as the visual information of their interaction in the image, caption generation task refers to composing coherent captions for each entire image. For the medical captioning task, rather than the mere coverage of visual concepts, detecting the interplay of visible elements can be crucial for strong performance" [3]. In the following, we describe our proposed models for Diagnostic Captioning, in which the generalized Nearest Neighbours baseline that we introduced in section 3.2.7 has a crucial role despite it performing poorly as is.

3.4.1. $(1 + k)$-NN image retriever with Pegasus summarizer

Our best performing models extend the Nearest Neighbours baseline for caption generation. Precisely, 1-NN [23] constitutes one of the model components, where for every image in the test set, it will produce the diagnostic text of the visually most similar image from the training set as the output and consequently it will assign the corresponding caption, say y^*, of the most similar image, say x^*, from the training set as output [2]. Thus, if we denote by $\mathbf{e}(.)$ the output of the employed image encoder among those mentioned in section 3.1, 1-NN predicts $(\hat{x}, \hat{y}) = (\hat{x}, y^*)$ that satisfies $(x^*, y^*) = \arg\min_{x^*} \cos(\mathbf{e}(\hat{x}), \mathbf{e}(x^*))$. This prediction constitutes the first part of the models' generated caption.

In the generalized baseline however, apart from the neighbour with the closest representation, we retrieve the top-$(k + 1)$ nearest neighbours, concatenate their outputs, excluding that of the most similar image and feed them as input to an abstractive summarizer; Pegasus [24] that is based on the transformer architecture [25], one idea that revolutionized Natural Language Processing and is trained with a Masked Language Modelling objective, which became popular within the research community though BERT [26].

For our models we employed a pre-trained AlexNet CNN on ImageNet classification dataset as our image encoder and merged our training, validation and development sets that are described in section 2, in order to benefit from an extensive set of train data to compute similarities with the test images. For each of them we keep the caption of the visually most similar image, concatenate the captions of the k proceeding ones and give them as input to Pegasus summarizer, which we allow to produce a summary of maximum length n tokens to eliminate repetitions. We exclude phrases as "All images are copyrighted." and "Images courtesy of AFP, EPA, Getty" that were probably included in Pegasus' training set from our generated summaries. The predicted captions constitute the concatenation of 1-NN baseline and Pegasus summarizer outputs. Table 6 below presents all configurations' hyper-parameter values, namely k and n, their submission IDs and BLEU scores in decreasing order [27].

Table 6
Summary of our configurations' hyper-parameters and statistics

Backbone Network	Captions k	Tokens n	BLEU scores	Submission ID
AlexNet	$k = 9$	$n = 15$	0.29166	182337
AlexNet	$k = 4$	$n = 15$	0.28343	182286
AlexNet	$k = 3$	$n = 15$	0.27855	182284
AlexNet	$k = 2$	$n = 15$	0.27007	182285
AlexNet	$k = 4$	$n = 5$	0.25521	182271
AlexNet	$k = 3$	$n = 5$	0.25334	182272

3.4.2. k-NN image retriever with Retrieval Augmented Generation

It has been impressive to researchers how nowadays general-purpose sequence-to-sequence models are getting really powerful, they manage to capture the world knowledge in parameters, they achieve strong results on loads of tasks and are applicable for almost everything. However, they often hallucinate, may usually struggle to access, and apply knowledge and are difficult to update. On the other hand, modern Information Retrieval (IR) is great as well, as externally reviewed knowledge can be useful for a huge variety of NLP tasks. Modern IR provides a precise and accurate knowledge access mechanism, it is trivial to update, whereas by "modern" IR we refer to dense retrieval that starts to outperform traditional IR. On the negative side though, it still needs retrieval supervision or heuristics such as BM25, as well as some –task specific– way to integrate into downstream tasks.

The goal of Retrieval Augmented Generation (RAG) [28], which was used as model component, pretrained on Wikipedia with a FAISS index [29] built on 42% of PubMed 2022 including recent publications related to the fields of neuroscience and computational biology; is to combine the strengths of sequence-to-sequence models and explicit knowledge retrieval. Obviously, RAG is also blended with the 1-NN baseline; namely its outputs are concatenated with the caption of the visually most similar image from the training set to produce caption predictions. This model uses either a pre-trained AlexNet or VGG-16 CNN on ImageNet classification dataset as backbone network and, despite it containing a non-parametric memory, additional to storing information in the parameters of a sequence-to-sequence generative model that is a Bidirectional Auto-Regressive Transformers (BART) generator [30], after merging our training, validation and development sets that are described in section 2 to take advantage of more input-output pairs $(\boldsymbol{x}, \boldsymbol{t})$, achieves a lower BLEU score than its predecessors described in 3.4.1 according to Table 7 below. These results could possibly improve if we store extracts from patients' previous diagnoses instead of biomedical articles.

Table 7
Summary of our configurations' image encoders and statistics

Backbone Network	BLEU scores	Submission ID
AlexNet	0.25127	181712
VGG-16	0.23958	181860

In the RAG approach [28], dual memory components are pre-trained and pre-loaded with extensive knowledge to encapsulate information via the representations without further training; the generator p_θ acts as a parametric memory, with the retriever p_η embodying a non-parametric memory in the query encoder $\mathbf{q}(.)$, while also including a Dense Passage Retriever (DPR) [31]. To train the retriever p_η and generator p_θ end-to-end, we can treat the retrieved document as a latent variable $\boldsymbol{z}$, while the embedding of the closest document representation is represented as $\mathbf{d}(\boldsymbol{z})$). The Maximum Inner Product Search (MIPS) algorithm [32] is used to compute the top k retrieved documents with respect to $p_\eta(\boldsymbol{z}|\boldsymbol{x})$. Finally, the generated caption $\boldsymbol{y}$ is produced by marginalizing over the predictions.

$$p_\eta(\boldsymbol{z}|\boldsymbol{x}) = \exp\left(\mathbf{d}(\boldsymbol{z})^T \mathbf{q}(\boldsymbol{x})\right)$$

The generator p_θ is a sequence-to-sequence model, a BART [30] instance precisely, which conditions on the latent documents $\boldsymbol{z}$ together with each input $\boldsymbol{x}$ to generate each output. As an overall component, it produces $p_\theta\left(\boldsymbol{y}_i|\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}_{1:i-1}\right)$ to create a Language Model (LM) over the tokens vocabulary $\mathcal{V}$ given as input the latent documents $\boldsymbol{z}$ and queries $\boldsymbol{x}$, which are the outputs of 1-NN baseline. During training, we treat questions-answers as input-output pairs i.e. $(\boldsymbol{x}, \boldsymbol{t})$ and train RAG-token by directly minimizing the negative marginal log-likelihood of generating output sequences $\boldsymbol{y}$ on input sequences $\boldsymbol{x}$. If $\mathcal{D} = \{\boldsymbol{x}_j, \boldsymbol{t}_j\}_j$ is the complete dataset, our training objective is:

$$l_{\text{cross}}(\boldsymbol{x}, \boldsymbol{t}; \theta, \eta) = -\log p(\boldsymbol{y}|\boldsymbol{x}; \theta, \eta)$$

$$\sum_j l_{\text{cross}}(\boldsymbol{x}_j, \boldsymbol{t}_j; \theta, \eta) = \sum_j -\log p(\boldsymbol{y}_j|\boldsymbol{x}_j; \theta, \eta)$$

3.4.3. 1-NN image retrieval baseline

Last but not least, we attempted using the 1-NN baseline [2] as is to generate the diagnostic text within the captions, which however achieved a lower score than all the aforementioned approaches. Although at first, one could interpret this as RAG models, in which the generator acts as a parametric memory, whereas the retriever p_η embodies a non-parametric memory in the query encoder $\mathbf{q}$ examined in section 3.4.2, perform better than solely the 1-NN baseline; when the latter is combined with abstractive summarization techniques for the diagnostic texts of k additional visually similar images from the training set, where $k \in \mathbb{Z}^+$, it may perform better as it is indicated in section 3.4.1 and Table 6. Our models use a pre-trained AlexNet or VGG-16 CNN on ImageNet classification dataset as image encoder, our training, validation and development sets that are described in section 2 merged together and achieve a BLEU score according to Table 8.

Table 8
Summary of our configurations' image encoders and statistics

Backbone Network	BLEU scores	Submission ID
AlexNet	0.24064	181711
VGG-16	0.22757	181859

4. Directions for future work

In this work, we developed CNN-based image encoders trained end-to-end for tags assignment or combined with heuristics such as the 1-NN baseline for either concept prediction or caption generation, which although is really simple performs rather well if combined with abstractive summarization algorithms, as highlighted in section 3.4.1 as well as the study in [2], where this baseline itself performs well for the Indiana University chest X-ray Collection [33] (IU chest X-ray dataset). Future work could focus on the use of task-specific models for summarization, such as Bio-BERT [34], further fine-tuning on the number of neighbours k and the summary maximum length n in section 3.4.1 and consideration of potential associations between the two subtasks during 1-NN baseline extension.

Furthermore, although higher quantitative accuracy is most often better, there are categorical differences of the DC methods as well, which relate to their qualitative evaluation and indicate their practical usefulness. It is an open question how we may obtain practical information about the quality of the generated captions.

5. Ethical considerations

Development of Diagnostic Captioning systems based on novel DL architectures could have both positive and negative societal impacts. My proposed work, for example, may be used for analyzing medical image data in undeveloped regions or countries under development. This is related to the 3rd goal of United Nations Sustainability Goals (UNSG) about ensuring good health and well-being and the 10th goal about reduced inequalities. On the other hand, privacy issues might arise from the use of medical data and "concerns over the sensitive information security and privacy" [35] that may also be related to the General Data Protection Regulation (GDPR) and EU legislation.

References

[1] C. E. Lawson, J. M. Martí, T. Radivojevic, S. V. R. Jonnalagadda, R. Gentz, N. J. Hillson, S. Peisert, J. Kim, B. A. Simmons, C. J. Petzold, S. W. Singer, A. Mukhopadhyay, D. Tanjore, J. G. Dunn, H. Garcia Martin, Machine learning for metabolic engineering: A review, Metabolic Engineering 63 (2021) 34–60. URL: https://www.sciencedirect.com/science/article/pii/S109671762030166X. doi:https://doi.org/10.1016/j.ymben.2020.10.005, tools and Strategies of Metabolic Engineering.

[2] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, D. Papamichail, Diagnostic captioning: a survey, Knowledge and Information Systems (2022) 1–32. doi:10.1007/s10115-022-01684-7.

[3] 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.

[4] B. Ionescu, H. Müller, R. Peteri, J. Rückert, A. B. 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. Stefan, 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.

[5] 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. L. Martel, L. Maier-Hein, L. Duong, G. Zahnd, S. Demirci, S. Albarqouni, S. 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 - 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings, volume 11043 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 180–189. URL: https://doi.org/10.1007/978-3-030-01364-6_20. doi:10.1007/978-3-030-01364-6_20.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge, International Journal of Computer Vision 115 (2014). doi:10.1007/s11263-015-0816-y.

[7] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Y. W. Teh, M. Titterington (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, volume 9 of *Proceedings of Machine Learning Research*, PMLR, Chia Laguna Resort, Sardinia, Italy, 2010, pp. 249–256. URL: https://proceedings.mlr.press/v9/glorot10a.html.

[8] A. Krizhevsky, One weird trick for parallelizing convolutional neural networks, CoRR abs/1404.5997 (2014). URL: http://arxiv.org/abs/1404.5997. arXiv:1404.5997.

[9] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely Connected Convolutional Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.

[10] 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), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[11] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv 1409.1556 (2014).

[12] M. Tan, Q. V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, CoRR abs/1905.11946 (2019). URL: http://arxiv.org/abs/1905.11946. arXiv:1905.11946.

[13] 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, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[14] I. Athanasiadis, G. Moschovis, A. Tuoma, Weakly-Supervised Semantic Segmentation via Transformer Explainability, in: ML Reproducibility Challenge 2021 (Fall Edition), 2022. URL: https://openreview.net/forum?id=rcEDhGX3AY.

[15] J. Surowiecki, The Wisdom of Crowds, Anchor, 2005.

[16] D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, International Conference on Learning Representations (2014).

[17] I. Loshchilov, F. Hutter, Fixing Weight Decay Regularization in Adam, 2018. URL: https://openreview.net/forum?id=rk6qdGgCZ.

[18] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Y. Ding, A. Bagul, C. P. Langlotz, K. S. Shpanskaya, M. P. Lungren, A. Y. Ng, CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, CoRR abs/1711.05225 (2017). URL: http://arxiv.org/abs/1711.05225. arXiv:1711.05225.

[19] J. Konar, P. Khandelwal, R. Tripathi, Comparison of Various Learning Rate Scheduling Techniques on Convolutional Neural Network, in: 2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS), 2020, pp. 1–5. doi:10.1109/SCEECS48394.2020.94.

[20] H. Robbins, S. Monro, A stochastic approximation method, Annals of Mathematical Statistics 22 (1951) 400–407.

[21] C. J. Darken, J. E. Moody, Note on Learning Rate Schedules for Stochastic Optimization, in: NIPS, 1990.

[22] L. N. Smith, No More Pesky Learning Rate Guessing Games, CoRR abs/1506.01186 (2015). URL: http://arxiv.org/abs/1506.01186. arXiv:1506.01186.

[23] G. Liu, T. H. Hsu, M. B. A. McDermott, W. Boag, W. Weng, P. Szolovits, M. Ghassemi, Clinically Accurate Chest X-Ray Report Generation, CoRR abs/1904.02633 (2019). URL: http://arxiv.org/abs/1904.02633. arXiv:1904.02633.

[24] J. Zhang, Y. Zhao, M. Saleh, P. J. Liu, PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization, in: Proceedings of the 37th International Conference on Machine Learning, ICML'20, JMLR.org, 2020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser,

I. Polosukhin, Attention is All you Need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[26] 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://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[27] 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://aclanthology.org/P02-1040. doi:10.3115/1073083.1073135.

[28] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[29] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.

[30] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[31] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-t. Yih, Dense Passage Retrieval for Open-Domain Question Answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6769–6781. URL: https://aclanthology.org/2020.emnlp-main.550. doi:10.18653/v1/2020.emnlp-main.550.

[32] S. Mussmann, S. Ermon, Learning and Inference via Maximum Inner Product Search, in: M. F. Balcan, K. Q. Weinberger (Eds.), Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*, PMLR, New York, New York, USA, 2016, pp. 2587–2596. URL: https://proceedings.mlr.press/v48/mussmann16.html.

[33] D. Demner-Fushman, M. Kohli, M. Rosenman, S. Shooshan, L. Rodriguez, S. Antani, G. Thoma, C. McDonald, Journal of the American Medical Informatics Association : JAMIA 23 (2016) 304–310. doi:10.1093/jamia/ocv080, publisher Copyright: © 2015 Published by Oxford University Press on behalf of the American Medical Informatics Association 2015. This work is written by US Government employees and is in the public

domain in the US.

[34] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. doi:10.1093/bioinformatics/btz682.

[35] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, M. Saadi, Big data security and privacy in healthcare: A Review, Procedia Computer Science 113 (2017) 73–80. URL: https://www.sciencedirect.com/science/article/pii/S1877050917317015. doi:https://doi.org/10.1016/j.procs.2017.08.292, the 8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2017) / The 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2017) / Affiliated Workshops.