NeuralDynamicsLab at ImageCLEFmedical 2022

Georgios Moschovis[1] and Erik Fransén[1, 2]
{geomos, erikf}@kth.se

1. KTH Royal Institute of Technology,
Division of Computational Science and Technology,
School of Electrical Engineering and Computer Science,
Lindstedtsvägen 5, 114 28 Stockholm, Sweden
2. Science for Life (SciLife) Laboratory,
Tomtebodavägen 23A, 171 65 Solna, Sweden

Research Question

To what extent Deep Neural Networks are capable of **automatically** generating a diagnostic text from a set of medical images but also how much their interpretation of these medical images can assist medical professionals **reduce their amount of clinical errors**, as well as help them **increase their productivity by ameliorating the quality and speed** in producing medical diagnoses, which is associated to an **increased throughput** of medical imaging departments.

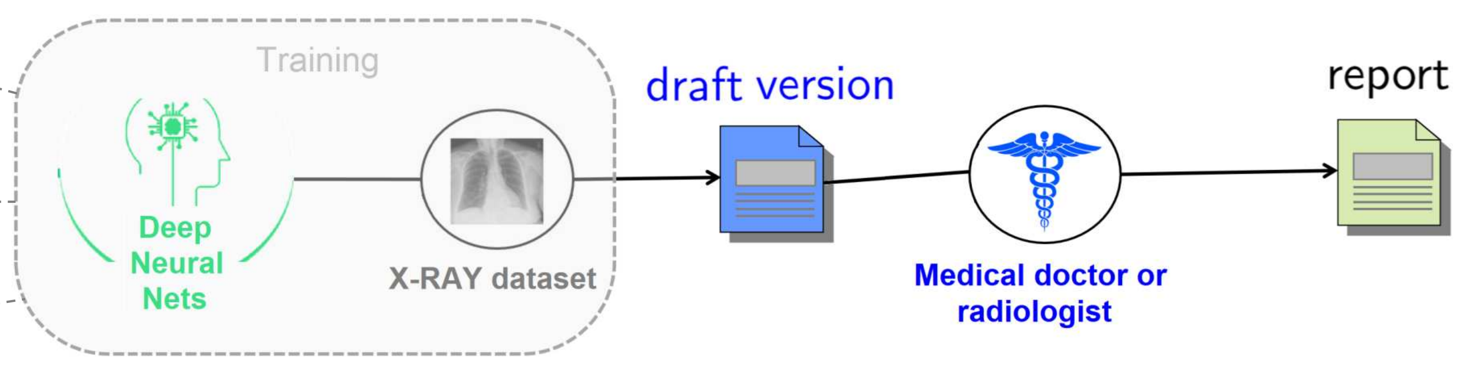
Data

90920 medical images; 80% training set, 10% validation set, 10% development set

Proposed Method

Image encoders: State-of-the-art CNN architectures, pretrained on ImageNet for classification, which have been obtained through `torchvision` models' library to perform inference; in order to encode the medical images into **descriptive dense numerical representations**. **They are shared for both subtasks.**

AlexNet | ResNet 50 101
DenseNet 121 161 | VGG 13 16
EfficientNet B3 B5



Concept Prediction

8374 tags of concepts assigned to the medical images. Each image in the training, validation, or development set is assigned **5 tags on average** based on a reduced subset of the Unified Medical Language System 2020 AB release.

In all baselines we use pre-trained encoders and train Perceptron heads initialized using Glorot, apart from the latter two, where we fully fine-tune a DenseNet161 and use the tags of the visually most similar image respectively.

Backbone Network	Training Regime	Learning Rate	Test F_1
DenseNet161	Adam optimizer and gradient clipping	constant 10^{-3}	0.43601
DenseNet161	Adam optimizer and gradient clipping[1]	constant 10^{-3}	0.43567
DenseNet161	AdamW optimizer and gradient clipping	constant $5 \cdot 10^{-4}$	0.43558
DenseNet161	Adam optimizer without gradient clipping	constant $5 \cdot 10^{-4}$	0.43539
DenseNet variants	Ensemble of best-performing DenseNets	per weak learner	0.43496
Various networks	Ensemble of diverse configurations[2]	per weak learner	0.43404
Various networks	Ensemble of diverse configurations[2]	per weak learner	0.43130
Various networks	Ensemble of diverse configurations[2]	per weak learner	0.42957
DenseNet161	Full fine-tuning with AdamW optimizer	cyclical	0.31687
VGG-16	Nearest Neighbor baseline (1-NN)	–	0.25061

[1] Model training occurred in 80% of the data; apart from the best performing DenseNet161 where we merge the training, validation, development sets and train in all the provided data (all 90920 medical images).
[2] Configuration search involves optimizers, learning rates, number of epochs, batch sizes, weight decay.

Caption Generation

Training set: 72736 captions, 70879 unique captions, average length 108 words,
Validation set: 9092 captions, 8984 unique captions, average length 107 words, and
Development set: 9092 captions, 8977 unique captions, average length 108 words

In (1+k)-NN we keep the caption of the visually most similar image as is and pass the remaining k ones to Pegasus summarizer; then concatenate. In k-NN with RAG we pass the captions of all k most similar images to RAG-token; then concatenate all of them with RAG-token's generation.

Backbone Network	Training Regime	Neighbors	Length	BLEU
AlexNet	(1+k)-NN retriever with Pegasus	k = 9	15 tokens	0.29166
AlexNet	(1+k)-NN retriever with Pegasus	k = 4	15 tokens	0.28343
AlexNet	(1+k)-NN retriever with Pegasus	k = 3	15 tokens	0.27855
AlexNet	(1+k)-NN retriever with Pegasus	k = 2	15 tokens	0.27007
AlexNet	(1+k)-NN retriever with Pegasus	k = 4	5 tokens	0.25521
AlexNet	(1+k)-NN retriever with Pegasus	k = 3	5 tokens	0.25334
AlexNet	k-NN retriever with RAG-token	k = 1	–	0.25127
VGG-16	k-NN retriever with RAG-token	k = 1	–	0.23958
AlexNet	k-NN retriever with RAG-token	k = 1	–	0.24064
VGG-16	Nearest Neighbor baseline	k = 1	–	0.22757