Advanced Topics in Deep Learing: The Rise of Transformers Project Report

Alessandro Carminati
Politecnico di Milano
alessandro.carminati@polimi.it

Abstract

Technological development in the medical imaging field increased the complexity and the load of information a radiologist must face. In the context of the project of the Ph.D. course "Advanced Topics in Deep Learning: The Rise of Transformers" we develop a Contrastive Language Image Pre-Training model to support radiologists and automatize part of their tasks. We train it on a medical imaging dataset and we test it on different tasks, such as few-shot classification and association between medical images and texts. We highlight that in certain tasks we exploit only parts of this model as ready-to-use building blocks for new models, while in other tasks we exploit this model as a whole. We suggest how to use this model for further applications in medical imaging. Examples of these applications include finding tumors using segmentation analysis and generating new images of rare cases for research purposes. We stress that these applications require long training times, high computational power, and huge datasets.

1. Introduction

Technological developments bring great innovation in the medical field, although they increase the complexity and the load of information a specialist must face. In particular, radiologists are required to analyze and generate textual descriptions (captions) of images obtained from several medical imaging modalities, such as Computer Tomography (CT), ultrasound, and Magnetic Resonance Imaging (MRI).

The project of the Ph.D. course "Advanced Topics in Deep Learning: The Rise of Transformers" aims at supporting radiologists in these tasks by developing automatic tools able to work both with images and captions. In particular, the project consists of implementing a Contrastive Language Image Pre-Training (CLIP) model [6], a multimodal neural network that can be used as a starting point for a wide range of applications, such as zero/few-shot learing, segmentation analysis, classification, and image generation.

This project exploits two datasets. The Radiology Objects in COntext (ROCO) dataset [5] consists of 128x128 RGB images with several medical imaging modalities, such as Computer Tomography (CT), Magnetic Resonance Imaging (MRI), and angiography. Each image is associated with a caption and a list of UMLS Concept Unique Identifiers (CUIs) [1]. The Pneumonia dataset [2] consists of chest x-ray images with various resolutions. Each image is associated with a label representing if the observed patient has pneumonia.

2. Hackathon

In the hackathon of 1 February 2023, we teamed up with Lapo Frascati (lapo.frascati@polimi.it), a PhD student in Information Technology. Due to our limited experience in the deep learning field, following the suggestions from the course's teachers, we developed two multiclass classifiers trained on ROCO to grasp better the concepts behind the CLIP model:

- A transformer-based classifier associating the captions to the CUIs.
- A Convolutional Neural Network-based classifier associating the images to the CUIs.

Each image and caption was associated with more than one CUI. For simplicity, we considered only the first CUI for each image and text.

We split ROCO in training, validation and test sets. We measured the accuracy of both classifiers using the test set, obtaining 88% for the transformer-based one, and 20% for the CNN-based one. We speculated that higher training times and a more precise model validation for the CNN-based classifier could have resulted in higher accuracy, but we were not able to prove this statement due to computational and time constraints.

In the last part of the hackathon we studied the structure of the CLIP model and we implemented its custom loss function.

3. Implementation and Applications

3.1. Train and test sets

A visual inspection of the ROCO dataset shows that observations with similar indexes are often related to the same type of medical imaging and body parts. Thus, we shuffled the dataset: in this way, any dataset slice contains a wide variety of images and captions. Then, we divided the ROCO dataset into three parts: training, validation, and test sets, containing 64, 16, and 20% of the entire dataset, respectively.

The Pneumonia dataset is already shuffled and divided into two parts: the training and the test sets, containing 624 (390 with pneumonia) and 5216 (3875 with pneumonia) images, respectively.

In the following, all the training is executed on the training sets, while all the results are obtained on the test sets.

3.2. Building and training the CLIP model

3.2.1 CLIP model structure

The two main parts of the CLIP model are the Transformer network [8] and the Convolutional Neural Network (CNN). The Transformer network's input is a vectorized caption T. First, each word w of T is embedded in a 128-dimensional space and the result is added to the 128-dimensional embedding of the position of w in the caption. The result is fed to to a sequence of transformer encoder blocks, each composed of a 4-head self-attention layer, two normalization layers, a feed-forward network, and two dropout layers. We refer to Figure 1 in [8] for a graphical representation of the transformer encoder block. A global averaging pooling layer reduces the output of the last transformer block to a 1-dimensional vector T_f. The CNN's input is a 128x128 RGB image I, encoded in a matrix with shape (128,128,3). The matrix contains integer values ranging from 0 to 255: they are normalized to have data in a more manageable range. Then the matrix is fed to to a series of convolutional blocks, composed of a 2D convolutional layer and a maximum pooling layer. In each convolutional block the number of filters is increased. We expect that this choice allows for the generation of an high number of very specific filter in the latest convolutional blocks. After the convolutional blocks, the matrix is flattened to obtain the vector I_f.
Then, the model computes the cosine similarity s between T_f and I_f. During training, if T is the real caption of I, the loss function of CLIP aims at maximizing s, otherwise, it aims at minimizing s.

3.2.2 Training and validation

We trained the CLIP model on the ROCO dataset, using the Adam optimizer [3] with a batch size of 64 elements. We selected the model's hyperparameters with a visual inspection of the validation and the training loss in each epoch: a lower validation loss was appreciated, while if the training loss was too low the model was probably overfitting and therefore was discarded. In particular, we selected a CLIP model with 5 convolutional blocks and 5 transformer encoder blocks, using an initial learning rate of 5e-5 for training.

We believe that the method for the hyperparameters choice we exploited has some critical points.

First, all the comparisons between trained models are mainly based on the loss on the same validation set. This factor makes the validation set very important for the hyperparameters choice, with the risk of selecting models that are overfitting the validation set. Such a problem can be reduced by employing k-fold cross-validation to obtain k estimates of the validation loss on k different datasets. Then, hyperparameters choice can be obtained using nonparametric statistical tests on estimates of the validation losses.

Second, in deep learning, training a model multiple times with the same hyperparameters can lead to very different performances, due to the complexity of the model's structure. This makes multiple training runs necessary for a better estimate of the validation loss. As in the first case, also with multiple runs the use of nonparametric statistical tests can give quantitative information for the hyperparameters choice.

In both cases, the proposed improvements were not adopted due to computational constraints.

3.3. Performance on the proxy task

We first test the trained model on the proxy task, i.e., on predicting which caption is paired with an image and vice versa. We evaluate the results using similarity matrices and we analyze singular examples.

Considering that the images and the captions have the same order, high percentages on the diagonal of the similarity matrix represent a correct execution of the proxy task. In Figure 1 we observe a similarity matrix with such structure: most of the cells on the diagonals have bright colors, representing high probabilities.

In Figure 2 we observe the images associated with a caption. The leftmost image, the correct one, has a 38.39% probability to be associated with the caption. We see that the second and third leftmost images, which have a cumulative probability of 38.89%, are similar to the correct image. Indeed, the three images with the highest percentage are x-rays of the abdomen, and the first part of the caption says: "xray of abdomen take on ...".

In Figure 3 we observe another example of the association between images and a caption, related to an abdominal CT scan. The distribution of probability in the four leftmost images, all abdominal CT scans, is more flat than the one in

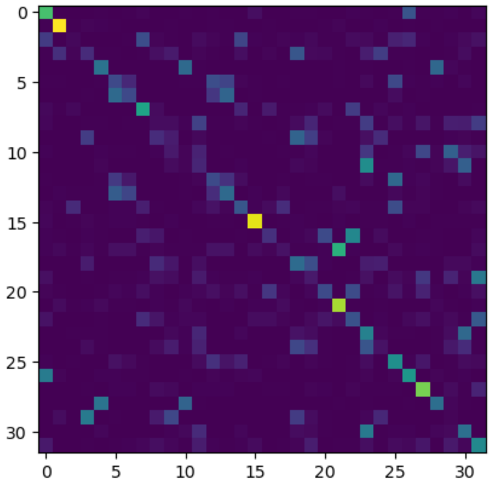

Figure 1. Similarity matrix obtained from CLIP proxy task.

the preceding example. In this case, the CLIP model finds it difficult to associate the caption to the correct abdominal CT scan image and gives the second highest probability to the correct image.

The two captions “chest radiography after of treatment” and “chest xray on pod show no remarkable finding except a little increase opacity on left chest wall” are associated with the image in Figure 4 with probabilities of 45.62% and 45.55%, respectively. The second caption is the correct one for the image. From this example, we can see that captions that are badly written and that lack information can generate confusion in the CLIP model.

3.4. Classification of the medical imaging type

We observed that a recurring information in the captions is the type of medical imaging. We found six types of medical imaging in the ROCO dataset: Computer Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound, Fluoroscopy, X-rays, and Angiography. We exploited the CLIP model to classify the images based on their medical imaging type: given image I, we compute the similarity matrix between I_f and the embedding of different prompts related to the medical imaging types. We took as ground truth the captions: if the name n of a medical imaging type is in the caption associated with an image, the ground truth is that the image is of type n. For this task, we removed from the ROCO dataset the images associated with captions that contained multiple or no medical imaging types. Using the prompts “ct scan”, “mri”, “ultrasound”, “fluoroscopy”, “xray”, and “angiography”, we obtained a 79.7% accuracy.

We also exploited the same classification task with an ensemble of prompts, where the final probabilities for each class were the mean of the probabilities obtained with each prompt vector. In this way, we tried different prompts: for instance, we added body parts (e.g., “abdominal ultrasound”) and we used abbreviations or extended names (e.g., “computed tomography”, “magnetic resonance imaging”). Using the ensemble of prompts we obtained a 79% accuracy.

The choice to use grammatically poor prompts stems from the analysis of the captions. Indeed, the captions often do not have the structure of a phrase, they mostly contain keywords related to the observed image.

We believe that there are some limits in our approach. First, the ground truth mechanism discards a good part of the dataset and can be wrong in certain cases. Second, we have limited knowledge related to the medical imaging field, which translates into poor prompt engineering. We believe that a specialist could provide a more precise ground truth and could generate prompts based on his domain knowledge that improve the accuracy of the model.

It is possible to extend this classification task to a wide range of classes, making the CLIP model a flexible tool in the medical imaging field. For instance, it could be used to analyze automatically medical imaging datasets and categorize the images based on their characteristics, such as the body part they represent.

3.5. Zero and Few-Shot Learning

We tested the performance of the CLIP model on the Pneumonia dataset, which was not used in the training phase. We exploited the CLIP model to obtain the probabilities of associating each image with the prompts ”normal” and ”pneumonia”, but we obtained poor performances. We believe that this result is related to a poor choice of prompts, as described in Section 3.4, and to the fact that the term ”pneumonia” is rarely used in the captions of ROCO.

We generated a new classifier starting from the CNN part of the CLIP model: starting from image I, we obtain the embedding I_f from the CNN, then I_f is fed to a sequence of two dense layers with ReLU activation function and a logistic regressor. The result of the logistic regressor is the probability that the x-ray comes from a patient with pneumonia. We trained this model for different sizes of the training dataset, keeping the CNN parameters frozen. In Figure 5 we represent the relationship between the size of the training dataset and the accuracy of the classifier. We observe that the model requires 100 samples to obtain an accuracy of around 81% and that with a higher size of the training dataset it reaches an accuracy of 84%.

This example highlights the importance of CLIP as a pretrained model: its blocks have already been trained and are ready-to-use, therefore the models that exploit them reduce their data and computational requirements for training.

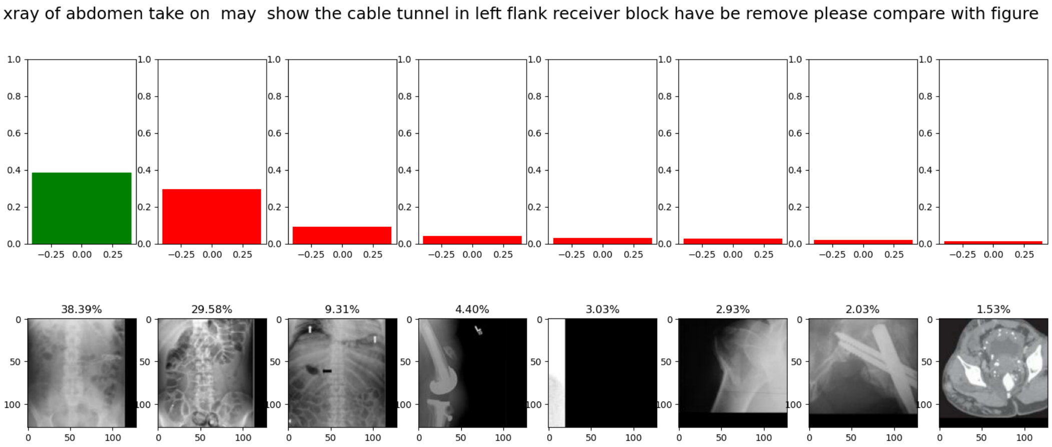

Figure 2. Images associated with a caption. The correct image is the one with the highest probability.

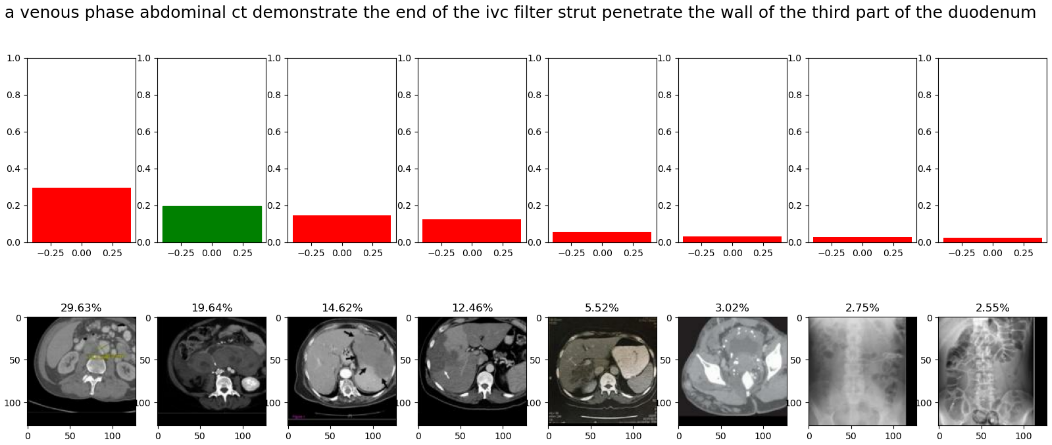

Figure 3. Images associated with a caption. The correct image is the one with the second highest probability.

4. Further Developments

In this Section, we propose further applications for the CLIP model, which were not explored due to technical or time-related constraints.

4.1. CUIs classification

It is possible to use the transformer network and the CNN of CLIP as encoders for a multi-label classifier, associating an image and a caption to some CUIs. The two networks generate two embeddings starting from an image and a caption and the concatenation of the embeddings is fed to a multi-label classifier block.

CUIs are 7-digit codes: their low interpretability makes it difficult for humans to work with them. The automatization of CUI-related tasks, such as the association between images and captions and CUIs, reduces the workload of radiologists: it removes the difficult, memory-based task of traducing the CUIs to the corresponding human-comprehensible concept.

4.2. Segmentation analysis

It is possible to use the transformer network of CLIP as a text encoder for a Language-Drive Semantic Segmentation (LSeg) [4] model. In this model, the transformer network encodes a list of segmentation labels, while an image encoder returns the embedding of the input image. Then, the image and the text embedding are multiplied to obtain, for each segmentation label, the part of the image related to that segment.

The LSeg model can be used as part of a tool for the automatic analysis of the content of medical imaging datasets, thanks to its ability to define which are the main objects and body parts in each image.

Such a model can also be a useful supporting system for the radiologists. It can help them find small particulars, reducing the time required for the visual analysis of each im-

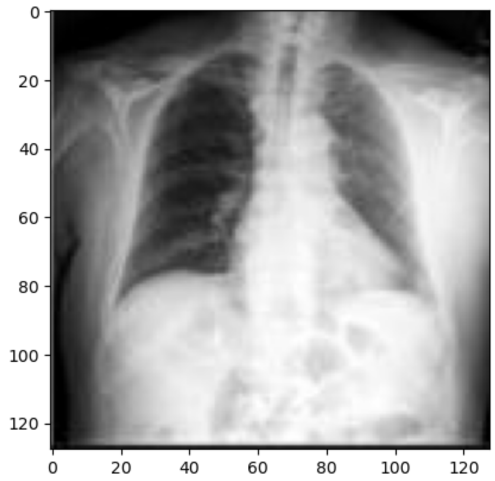

Figure 4. Radiography used for the CLIP proxy task.

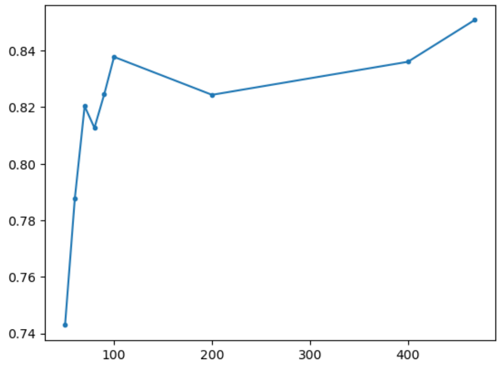

Figure 5. Relationship between the accuracy of the few-shot learning model and the training sample size.

age. For instance, suppose that LSeg can find the segment associated with the label "tumor": it could automatically find the presence and the dimensions of a tumor in an image and notify the specialist. Therefore, this type of application can have a positive impact on the medical diagnosis of tumors or other diseases visible in medical imaging.

4.3. Image generation

The CLIP model can be employed in building an image generation model, as depicted in [7]: the transformer network generates the embedding T_f starting from vectorized text T, and then T_f is fed to an autoregressive or diffusion prior to produce an image embedding I_f. I_f is used to condition a diffusion decoder which generates the final image.

In the medical imaging field, datasets are often unbalanced: in the real world certain diseases, conditions, and cures are rarer than others, and this difference can be reflected in the dataset. The image generation model can fill this gap by generating new versions of rare images.

Also, such a model can be used for didactic and research purposes: it is possible to generate images related to very extreme cases, which can be interesting in scientific research. For instance, an image of a patient with a combination of two rare diseases is very difficult, if not impossible, to find. Nevertheless, such an extreme case can be interesting for research purposes.

The generated images do not belong to any patient. This property makes their manipulation and sharing easier because they are not constrained by any prior with the patient.

We recognize that a medically reliable image generation model is extremely difficult to obtain and maintain: it requires a huge dataset containing a wide spectrum of diseases, patient characteristics, and medical imaging techniques, making its training computationally and time demanding. Also, such a model requires periodical updates to keep up-to-date with the new advancements in the medical field.

5. Conclusion

The project of the Ph.D. course "Advanced Topics in Deep Learning: The Rise of Transformers" required building a CLIP model and exploiting it in the context of medical imaging. In this article, we report the main choices in building and training this model, and we describe how we exploited this model for different tasks. We highlight that the CLIP model can be used as a whole, for instance exploiting prompt engineering for image classification purposes, or as a pre-trained model providing ready-to-use building blocks for other models, such as few-shot learners. Then, we propose further developments to this project, highlighting both their importance in real-world applications and their requirements.

References

[1] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32:267–270, 2004.

[2] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng, R. Hou, W. Shi, X. Fu, Y. Duan, V. A. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172:1122–1131, 2 2018.

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[4] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. Language-driven semantic segmentation, 2022.

[5] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich. Radiology objects in context (roco): A multimodal image

dataset. volume 11043 LNCS, pages 180–189. Springer Verlag, 2018.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. 2 2021.

[7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.