The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 1, No. 1. Online ISSN: 2831-4794
Available at https://www.isods.org/publication/

A Transformer-based Approach for Automated Findings using X-Ray Images

Cuong Do*[a], Pham The Bao*[b], Lan Vu*[c]
[a]Department of Mathematics, University of Connecticut, 341 Mansfield Road U1009, Storrs, Connecticut 06269-1009
[b]Information Science Faculty, Saigon University, Hochiminh City, Vietnam
[c]Department of Epidemiology, Hanoi University of Public Health, 1A Duc-Thang road, North Tu-Liem District, Hanoi, Vietnam

ABSTRACT

In this study we investigate an automated approach for generating findings from X-Ray images based on a multi-layer transformer decoder and various convolutional neural network architectures for image extraction. We further fine tune the convolutional neural network architectures. Cosine similarity based on BERT embeddings is selected as the measure of performance evaluation to account for semantic similarity. This approach could be useful in situations where there is a lack of expertise in reading medical images, or when a preliminary result or second opinion is needed.

Keywords: Transformers, Encoder, Decoder, Convolutional Neural Networks, VGG-16, VGG-19, Inception, Xception, ResNet, DenseNet.

1. INTRODUCTION

Image captioning (Image to text) [1-2] is the task of generating a written description of an image using computer algorithms. It is a common problem in the fields of computer vision and natural language processing and has practical applications in areas such as visual search, virtual assistants, and automatic image tagging.

An early approach to image captioning involved training a model to predict a fixed set of predefined captions for a given image. More recent methods use neural networks to generate a caption in a more flexible and open-ended manner, using a combination of convolutional neural networks and recurrent neural networks [3-4]. These models are typically trained on large datasets of images and their corresponding captions and can generate descriptive captions for a variety of images. In recent years, there have been significant advances in the task of image captioning using transformers [2]. Transformers are a type of neural network architecture that has been successful in natural language processing tasks, such as language translation and language generation [5-6].

One significant advancement in image captioning using transformers is the development of large, pre-trained models such as Universal Sentence Encoder (USE) [7], Bidirectional Encoder Representations from Transformers (BERT) [8], and Generative Pre-trained Transformer (GPT) [9-11]. Both BERT and GPT are based on the transformer architecture, but they differ in the specific components they use from the transformer. BERT uses only the encoder blocks, while GPT uses only the decoder blocks. These models have been trained on massive amounts of data and can be fine-tuned for the task of image captioning. Using these pre-trained models allows for faster and more accurate image captioning results, as the models have already learned a lot of the underlying structure and patterns in language.

Another advancement is the use of attention mechanisms in image captioning with transformers [12]. Attention mechanisms allow the model to focus on specific parts of an image while generating a caption, allowing for more accurate and detailed descriptions of the image. Overall, the use of transformers in image captioning has led to significant improvements in the quality and accuracy of image captions, and it is expected that this trend will continue as the technology continues to advance.

In the context of medical images, image captioning can be used to provide a concise summary of the key features and findings in a medical image [13-14], which can be useful for a variety of purposes in hospital operations. A potential use of image captioning in hospitals is to assist radiologists and other medical professionals in interpreting and communicating the findings from medical images. For example, image captioning could be used to generate a summary of the key features and findings in an X-ray, CT, or MRI scan, which could be useful for quickly identifying and communicating important information to other members of the healthcare team. Another potential use is to provide a more accessible and user-friendly interface for viewing medical images. For example, image captioning could be used to generate a textual description of an image that could be read aloud to a patient or displayed on a screen for a patient who has difficulty reading or viewing images due to visual impairments or other factors.

2. TRANSFORMERS AND RELATED ARCHITECTURES

Transformers are a type of neural network that are particularly effective at modeling sequential data, such as natural language [5-6]. One key advantage of transformers over other types of neural networks, such as recurrent neural networks (RNNs) or convolutional neural networks (CNNs), is that they are highly parallelizable, making them efficient to run on hardware such as GPUs and TPUs. This is because Transformers use attention mechanisms in place of recurrence, allowing computations to happen simultaneously. Additionally, transformers can capture long-range dependencies and contextual information in the data, allowing them to effectively process sequences with distant or unrelated elements. Unlike RNNs and CNNs, which make assumptions about the temporal or spatial relationships within the data, Transformers do not make any such assumptions, making them well-suited for processing sets of objects without inherent ordering.

The Universal Sentence Encoder (USE) is a natural language processing model developed by Google that is based on transformer architecture. It is designed to convert sentences or paragraphs into a numerical representation called an embedding, which can be used as input to machine learning models or other natural language processing tasks. The USE is trained on a large dataset of sentences and paragraphs in multiple languages and can generate embeddings that capture the meaning of the input text. It has been shown to perform well on a variety of tasks, including document classification, sentence classification, and semantic similarity comparison [7]. One of the key advantages of the USE is that it can generate embeddings for a wide range of languages, making it a useful tool for natural language processing tasks that involve multiple languages. It is also able to handle long input sequences, allowing it to be used for tasks such as document summarization. Overall, the USE is a powerful tool for natural language processing tasks that requires the encoding of text into numerical representations. It is particularly useful for tasks that involve multiple languages or long input sequences.

BERT, or Bidirectional Encoder Representations from Transformers, is a state-of-the-art natural language processing (NLP) model developed by Google. It was introduced in a research paper in 2018 and has since become widely used in various NLP tasks such as text classification, language translation, and question answering. One of the key features of BERT is its ability to process language in a “bidirectional” manner, meaning that it takes into account the context of a word both before and after it in a sentence. This is in contrast to many previous NLP models that only considered the context before a word (also known as "left-to-right" processing). The use of the Transformer architecture, which is highly efficient at processing sequential data, also allows BERT to process large amounts of text efficiently [8]. In addition to its impressive performance on various NLP tasks, BERT has also been used as a pre-trained model for fine-tuning on other NLP tasks, allowing researchers and developers to leverage its power without having to train a model from scratch. As a result, BERT has become a go-to model for many NLP applications and has sparked further research and development in the field [8].

GPT (Generative Pre-training Transformer) is a series of language generation models developed by OpenAI. The first model in the series, GPT (now referred to as GPT-1), was released in 2018 and was followed by GPT-2 in 2019 and GPT-3 in 2020 [9-11]. GPT models are based on transformer architecture, which allows them to process sequential data such as text efficiently and effectively. They are trained using unsupervised learning, which means that they are trained on large amounts of data without explicit labels or guidance on what to learn. This allows the models to learn the underlying structure and patterns in language and generate human-like text. GPT models have been successful in a variety of natural language processing tasks, including language translation, language summarization, and language generation. They have also been used in a wide range of applications, including chatbots, language translation systems, and content generation [9-11]. GPT-3 has recently received significant attention due to its large size and impressive performance on a variety of tasks. It is currently the largest publicly available language generation model and has demonstrated the ability to perform a wide range of tasks without explicit training on those tasks. The GPT-3.5 series is a collection of models that were trained on a combination of text and code from before the fourth quarter of 2021. On November 30th of 2022, GPT-3.5 was

introduced to the world through ChatGPT, a version of GPT-3.5 that has been fine-tuned for use as a general-purpose chatbot.

3. CONVOLUTIONAL NEURAL NETWORKS (CNNS)

Convolutional neural networks (CNNs) are a type of deep learning algorithm that are particularly effective for image data. There are several popular CNN architectures that have been developed in previous research, including VGG, Inception, Xception, ResNet, and DenseNet.

VGG architectures were introduced by Simonyan and Zisserman in 2014 and use 3x3 convolutional layers stacked on top of each other in increasing depth. Variants of this architecture include VGG11, VGG16, and VGG19, which refer to the number of weight layers in the network [15]. Inception, introduced by Szegedy et al. in 2014, is an architecture that uses dense modules or blocks instead of stacking convolutional layers. Popular variants of this architecture include Inception-v1 and Inception-v3, with Inception-v3 being a 48-layer deep CNN that uses batch normalization [16-17].

ResNet, introduced by Microsoft Research, is a deep CNN that addresses the issue of accuracy degradation as the network depth increases by using skip connections. It allows for the design of even deeper CNNs without compromising model generalization, and the ResNet-50 model is a 50-layer deep CNN with 26M parameters [18]. Xception, a 71-layer deep CNN proposed by Chollet in 2017, is an extension of the Inception architecture that uses depth-wise separable convolution layers [19].

DenseNet, introduced by Huang et al. in 2018, is an architecture that alleviates the vanishing gradient problem, strengthens feature propagation, encourages feature reuse, and reduces the number of parameters. It is composed of dense blocks in which the layers are densely connected, allowing each layer to receive more supervision from the loss function due to the shorter connections. A transition layer decreases the spatial size of the input. DenseNet-201 is a 201-layer deep CNN [20].

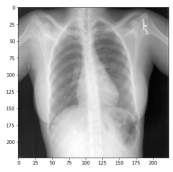
4. DATASET

The Indiana University Open Access chest X-ray collection, originally from NIH (https://openi.nlm.nih.gov/), consists of a set of 7470 chest X-ray images that were originally in DICOM (Digital Imaging and Communications in Medicine) format, table 1, which is a widely used standard for storing and transmitting medical images and information in hospitals and other healthcare settings. These images were processed to remove the top and bottom 0.5% of DICOM pixel values, scale the remaining pixel values to the range of 0-255, and resize the images to 2048 pixels. There are two image categories: frontal and lateral chest X-rays. Frontal X-rays images are selected to be included in model development. The total size of the dataset is 14GB.

uid	filename	projection	MeSH	Problems	image	indication	comparison	findings	impression
1	1_IM-0001-4001.dcm.png	Frontal	normal	normal	Xray Chest PA and Lateral	Positive TB test	None.	The cardiac silhouette and mediastinum size ar...	Normal chest x-XXXX.
2	2_IM-0652-1001.dcm.png	Frontal	Cardiomegaly/borderline;Pulmonary Artery/enlarged	Cardiomegaly;Pulmonary Artery	Chest, 2 views, frontal and lateral	Preop bariatric surgery.	None.	Borderline cardiomegaly. Midline sternotomy XX...	No acute pulmonary findings.
4	4_IM-2050-1001.dcm.png	Frontal	Pulmonary Disease, Chronic Obstructive;Bullous...	Pulmonary Disease, Chronic Obstructive;Bullous...	PA and lateral views of the chest XXXX, XXXX a...	XXXX-year-old XXXX with XXXX.	None available	There are diffuse bilateral interstitial and a...	1. Bullous emphysema and interstitial fibrosis...
6	6_IM-2192-1001.dcm.png	Frontal	normal	normal	PA and Lateral Chest. XXXX, XXXX at XXXX	Evaluate for infection	XXXX, XXXX	Heart size and mediastinal contour are within ...	No acute cardiopulmonary findings.

Fig. 1: An excerpt from the X-ray image description

Table 1. An excerpt of X-Ray images and findings from the training data

Images	Findings
	The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.
	Borderline cardiomegaly. Midline sternotomy XXXX. Enlarged pulmonary arteries. Clear lun gs. Inferior XXXX XXXX XXXX.
	There are diffuse bilateral interstitial and alveolar opacities consistent with chronic obstructiv e lung disease and bullous emphysema. There are irregular opacities in the left lung apex, that could represent a cavitary lesion in the left lung apex.There are streaky opacities in the right u pper lobe, XXXX scarring. The cardiomediastinal silhouette is normal in size and contour. Th ere is no pneumothorax or large pleural effusion.
	Heart size and mediastinal contour are within normal limits. There is no focal airspace consoli dation or suspicious pulmonary opacity. No pneumothorax or large pleural effusion. Mild deg enerative change of the thoracic spine.

Each image comes with text information such as uid, file name, projection, MeSH, problems, image, indication, comparison, findings, and impression. (See Fig. 1). Table 1 shows several examples of X-Ray images and corresponding findings. The data is split into the training, validation, and test sets at the ratios of 60%, 20%, and 20%.

5. MODELING APPROACH

Transfer learning [21] is applied that involves reusing the trained weights of a pre-trained model for a different task. In a CNN, the shallow layers learn low-level features from images, while the deep layers learn high-level features. The shallow layers can be used as a base model and the weights are frozen during training, with the new deep layers being trained for the specific task. Using a pre-trained model can save time and resources compared to training a model from scratch. We have tested several architectures for extracting image features, including VGG, Inception, Xception, ResNet, and DenseNet, in using their pretrained weights as well as fine-tuned.

Fig. 2 illustrates the model with a multi-layer transformer decoder architecture on the right, which is like the GPT model, and a convolutional neural network on the left as the encoder. In the figure, MHA denotes Multi-Head Attention, FF denotes Feed Forward for the feedforward layer. In each layer of the multi-layer transformer decoder, there are a self-attention layer that allows each output location to attend to the previous outputs via masking future positions, a cross attention layer that allows each output location to attend to the input image, and a feedforward network layer that further

processes each output location independently. Residual connections are used followed by layer normalization. The output of the model is a multiclass classification over the output vocabulary, using a sparse soft-max cross entropy loss with logits. The input captions are tokenized and positionally encoded, with a vocabulary size of 5000.

The scaled dot-product attention is calculated as [6]:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

In which $Q \in \mathbf{R}^{Lxd}$, $K \in \mathbf{R}^{Lxd}$, and $V \in \mathbf{R}^{Lxd}$ are matrices for Query, Key, and Value; d_k is the scaling factor, where L is the sequence length, and d is the features depth. In the original paper, the queries, keys, and values are projected h times for h different heads, and later concatenated, and once again projected to the final values [6]:

$$MHAttention(Q,K,V) = concat(head_1, head_2, \dots, head_h)W^O \quad (2)$$

Where W^O is the matrix to project to the final value vector, and $head_i$ is the head attention for the i^{th} head [6]

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (3)$$

Typically, h=8 for 8 different heads. The first MHA is in the self-attention mode, where all Q, K, and V are the same as the input. The second MHA takes features from images as V and K, and text features from the layer below it as Q.

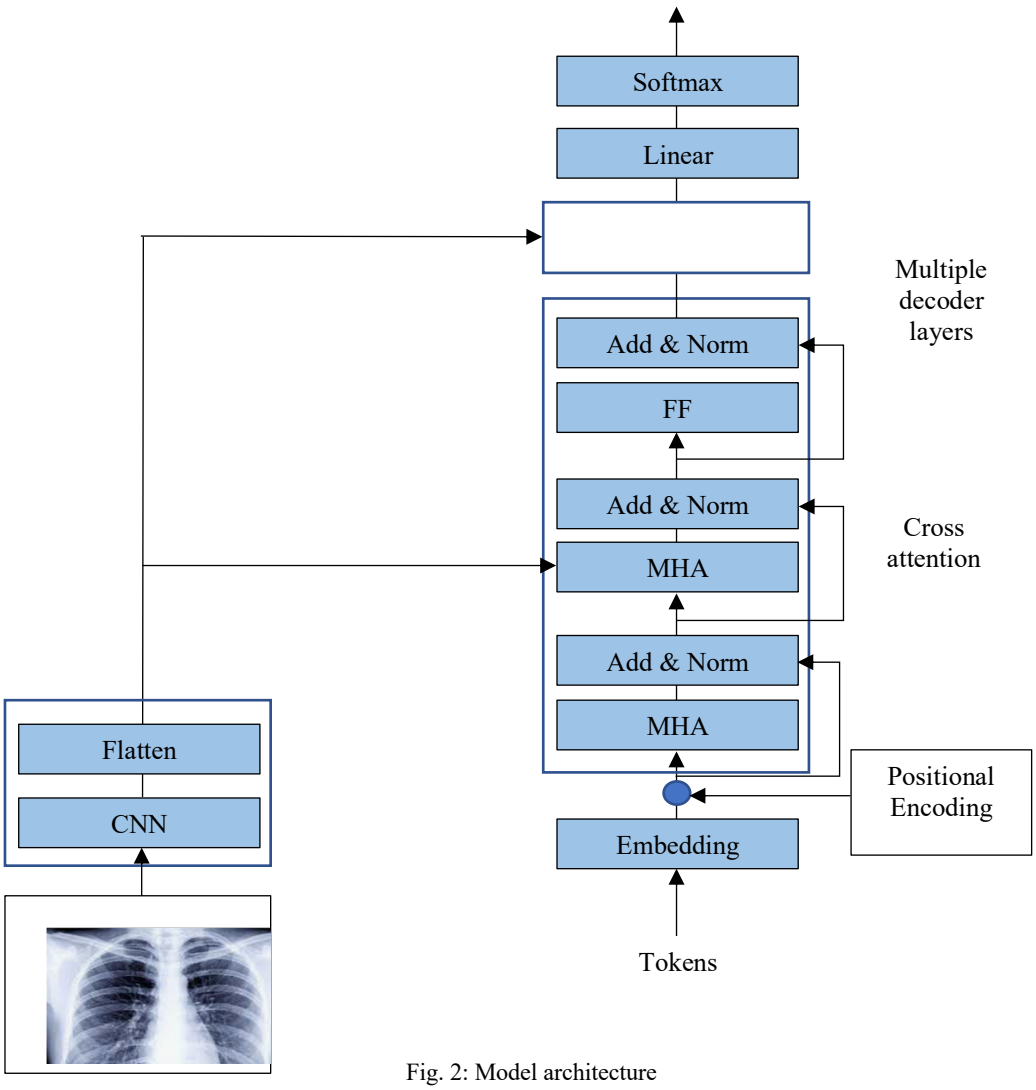


Fig. 2: Model architecture

We train the model on the training data and validate it on the validation data. To prevent overfitting, we use a stop-early callback. In the next section, we will evaluate the model performance on the test data. We use a temperature parameter t to adjust the softmax distribution, for example as t=0, greedy decoding is used, where the most likely next token at each step is selected; and as t=1, random sampling is used according to the logits.

6. RESULTS AND DISCUSSION

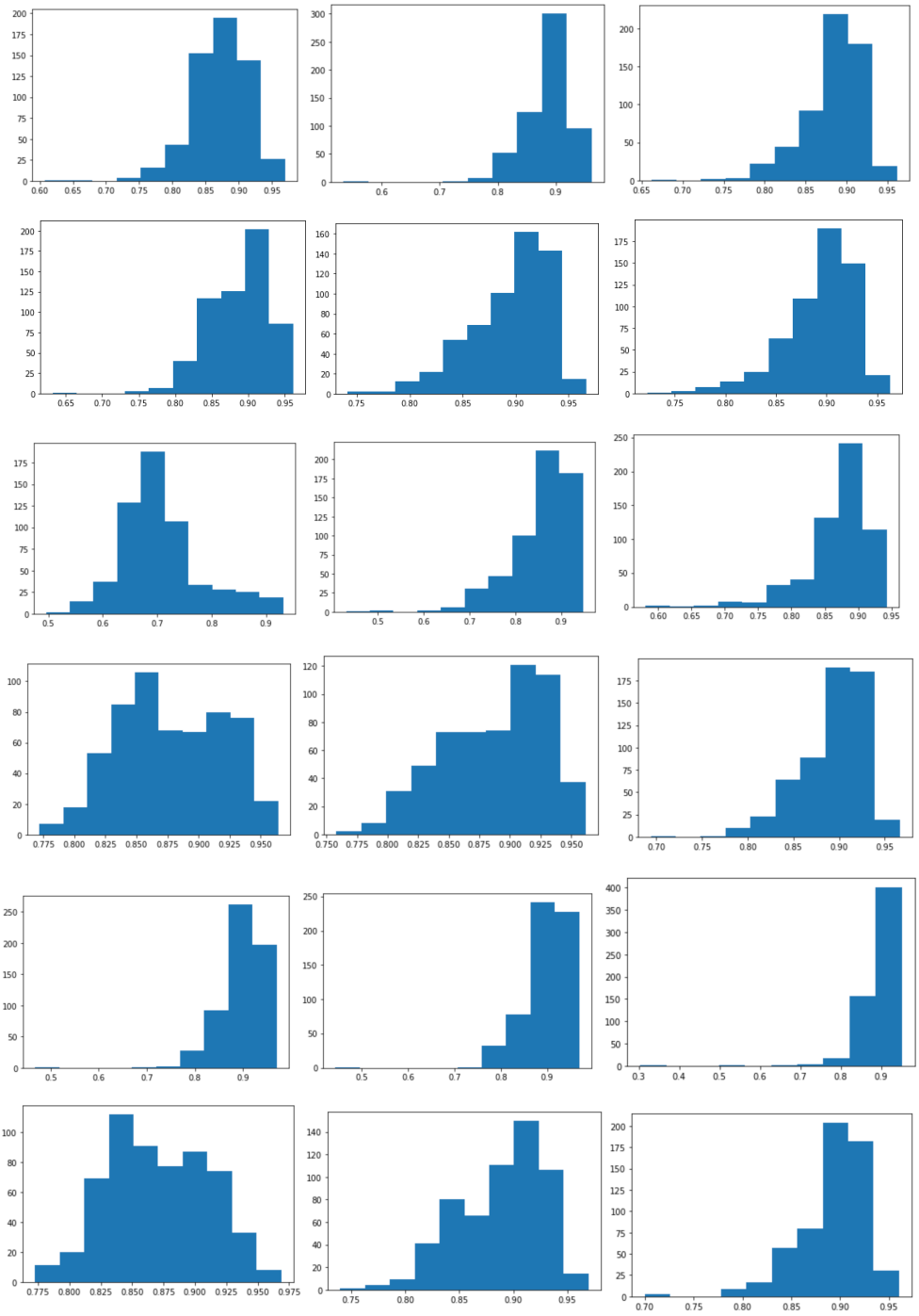


Fig. 3: Distributions of cosine similarity for Xception, ResNet50, InceptionV3, VGG16, VGG19, and DenseNet201, each on its rows with t=0, t=0.5, and t=1

In order to assess the semantic similarity of our models, we utilize cosine similarity based on BERT embeddings. Figure 3 illustrates the distribution of cosine similarity within the test set. Table 2 provides the mean cosine similarity for the model with each architecture of the image extraction part.

From Table 2, without further training VGG19 performs well with high mean cosine similarity at the temperate t=0 and t=0.5. Changing temperature does not seem to affect its performance. ResNet50 at t=0.5 performs just as well.

Table 2. Model performance using pretrained CNN architectures

Models	**Cosine Similarity (Mean)**		
	t=0	**t=0.5**	**t=1**
Xception	0.872	0.886	0.884
ResNet50	0.887	0.895	0.894
InceptionV3	0.707	0.856	0.870
VGG16	0.878	0.888	0.893
VGG19	0.896	0.896	0.895
DenseNet201	0.871	0.889	0.892

We further fine tune the convolutional neural network architectures. Table 3 provides the mean cosine similarity for the model with each of the fined-tune architecture for image extraction. From the table, it could be seen that InceptionV3 outperforms others at both temperatures t=0 and t=0.5. Next to InceptionV3 are DenseNet201, and Xception with lightly lower performance.

Table 3. Model performance using fine-tuned CNN architectures, tuned on the X-ray/description dataset

Models	**Cosine Similarity (Mean)**		
	t=0	**t=0.5**	**t=1**
Xception	0.906	0.906	0.893
ResNet50	0.887	0.896	0.883
InceptionV3	0.912	0.912	0.899
VGG16	0.895	0.897	0.895
VGG19	0.859	0.883	0.888
DenseNet201	0.911	0.909	0.898

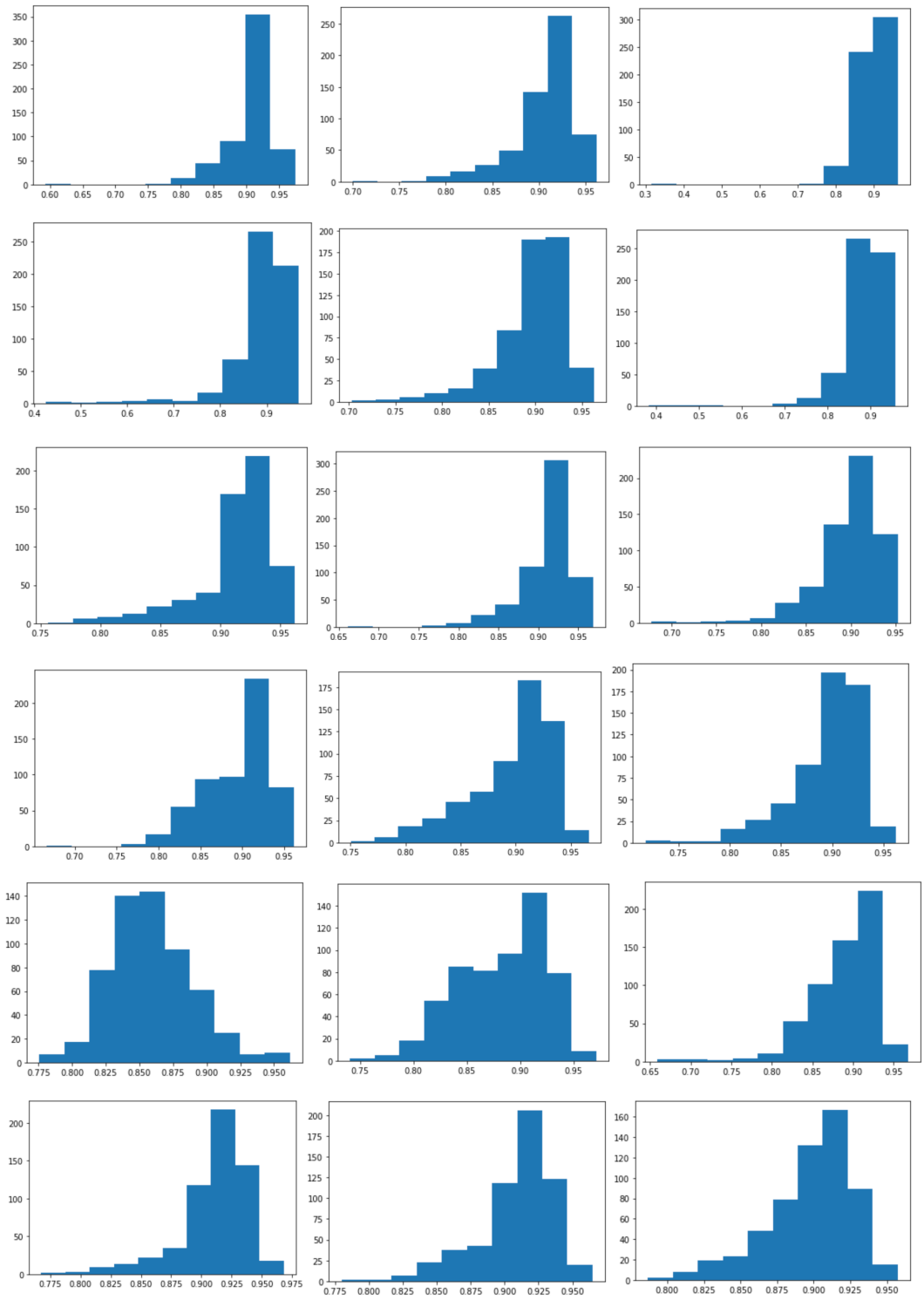


Fig. 4: Distributions of cosine similarity for fine-tuned Xception, ResNet50, InceptionV3, VGG16, VGG19, and DenseNet201, each on its rows with t=0, t=0.5, and t=1

7. CONCLUSION

To summarize, we have investigated and compared various architectures using a combination of convolutional neural networks and transformers for automating medical image analysis. Several convolutional neural network architectures were used for image extraction for comparison to evaluate the influence on the result, in which pretrained weights were used as well as fine-tuned weights, compared to previous work that only used a specific architecture with or without fine tuning. Our results show that the approach using fine-tuned InceptionV3 achieved the best performance with a cosine similarity of 0.912 based on BERT embeddings on the test set. With these models, hospitals and healthcare providers may be able to use them for preliminary results, for second opinions, or as a main resource in cases where there is a shortage of expertise for medical image analysis. For future research, we plan to consider additional information in the text data in addition to findings, and to use and fine-tune transformer-based architectures such as USE, BERT, and GPT for sentence embeddings.

REFERENCES

[1] Vinyals, O. et al., "Show and tell: A neural image caption generator," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 3156-3164, 2015.

[2] Tsaniya, H. et al., "Transformer Approaches in Image Captioning: A Literature Review," 2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE), 2022, pp. 1-6, 2022.

[3] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R., "From show to tell: A survey on image captioning," arXiv preprint arXiv:2107.06912, 2021.

[4] Elhagry, A., & Kadaoui, K., "A thorough review on recent deep learning methodologies for image captioning," arXiv preprint arXiv:2107.13114, 2021.

[5] Tunstall, L. et al., "Natural Language Processing with Transformers, 1st Edition", O'Reilly Media, 2022.

[6] Vaswani, A. et al., "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010, 2017.

[7] Cer, D et all, "Universal sentence encoder," arXiv preprint arXiv:1803.11175, 2018.

[8] Devlin, J., Chang, Ming-W., Lee, K., and Toutanova, K., "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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics, 2019.

[9] Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I., "Improving language understanding by generative pre-training," 2018.

[10] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I., "Language Models are Unsupervised Multitask Learners," Available at: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf, 2018.

[11] Brown, T. et al., "Language models are few-shot learners," Advances in neural information processing systems, 33, 1877-1901, 2020.

[12] Chordia, S., Pawar, Y., Kulkarni, S., Toradmal, U., Suratkar, S., "Attention Is All You Need to Tell: Transformer-Based Image Captioning," in Advances in Distributed Computing and Machine Learning. Lecture Notes in Networks and Systems, vol 427. Springer, Singapore.

[13] Wang, X., Peng, Y., Lu, L., Lu, Z., & Summers, R. M, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 9049-9058), 2018.

[14] Benjamin Hou, Georgios Kaissis, Ronald M. Summers, and Bernhard Kainz, "RATCHET: Medical Transformer for Chest X-ray Diagnosis and Reporting". In Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27 – October 1, 2021, Proceedings, Part VII. Springer-Verlag, Berlin, Heidelberg, 293–303, 2021.

[15] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556 [cs], 2015.

[16] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going Deeper with Convolutions," arXiv:1409.4842 [cs], 2014.

[17] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., “Rethinking the Inception Architecture for Computer Vision,” arXiv:1512.00567 [cs], 2015.
[18] He, K., Zhang, X., Ren, S. and Sun, J., “Deep Residual Learning for Image Recognition,” arXiv:1512.03385 [cs], 2015.
[19] Chollet, F., “Xception: Deep Learning with Depthwise Separable Convolutions,” arXiv:1610.02357 [cs] (2017).
[20] Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K. Q., “Densely Connected Convolutional Networks,” arXiv:1608.06993 [cs], 2018.
[21] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," in Proceedings of the IEEE, vol. 109, no. 1, pp. 43-76, Jan. 2021.