Medical Image Captioning with Medical Pre-Trained Transformers

Diego Esteban Quintero Rey

Ingeniería de Sistemas y Computación

Tomado de: Doctor analyzing an x-ray image on the computer screen - stock photo 3225773 | Crushpixel

Visión de Máquina

Docente: Flavio Augusto Prieto Ortiz

Universidad Nacional de Colombia

Medical Image Captioning (MIC)

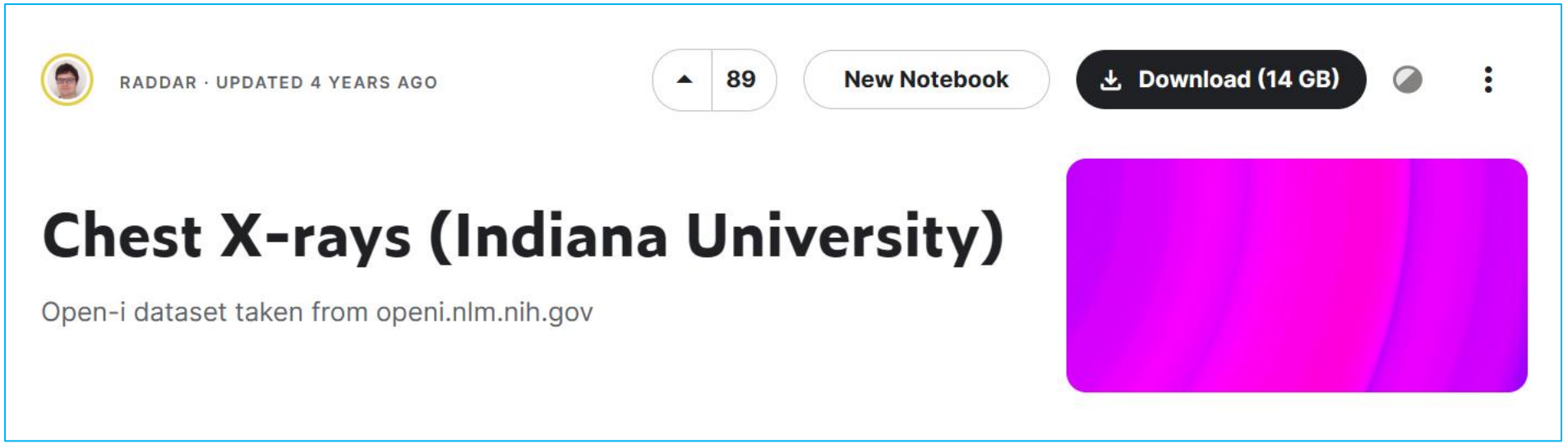
In the medical domain, image captioning consists in generating medical reports to highlight the most important clinical findings observed in the image [1].

Tomado de: Digital X-ray Services in San Diego - Imaging Healthcare Specialists

Proposal

1. Encoder: Medical pre-trained vision transformer [4]
2. Decoder: Medical pre-trained transformer [5]

Dataset

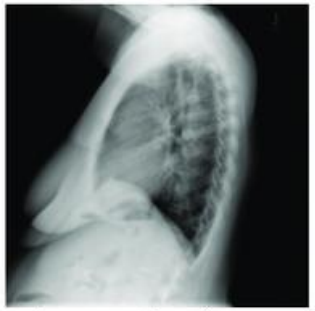


7466 unique values	Frontal	51%
	Lateral	49%

3822 Frontal Images
3305 Frontal Images with Reports

Dataset

frontal view



lateral view

Medical Image Report

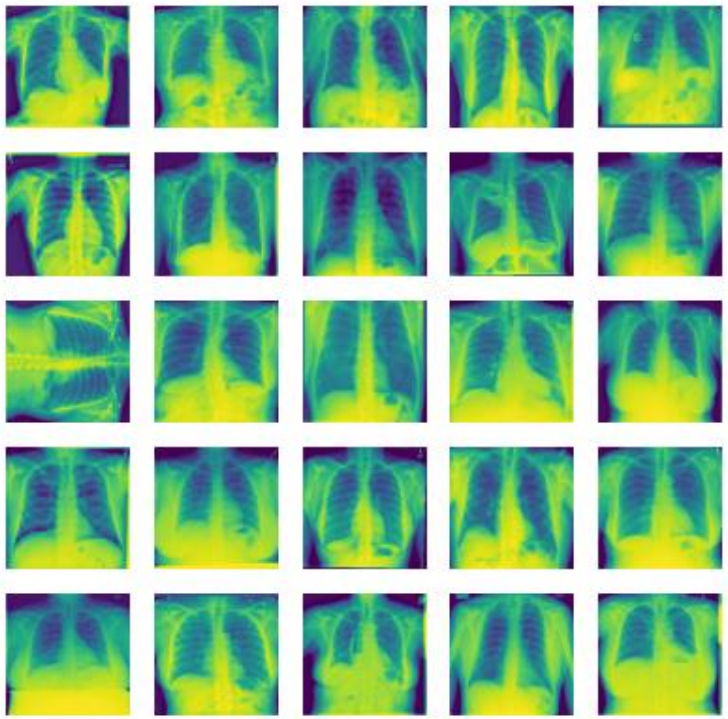
Findings: Heart size and pulmonary vascularity appear within normal limits. There is mild tortuosity to the descending thoracic aorta. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. No discrete nodules or adenopathy are noted. Degenerative changes are present in the spine.

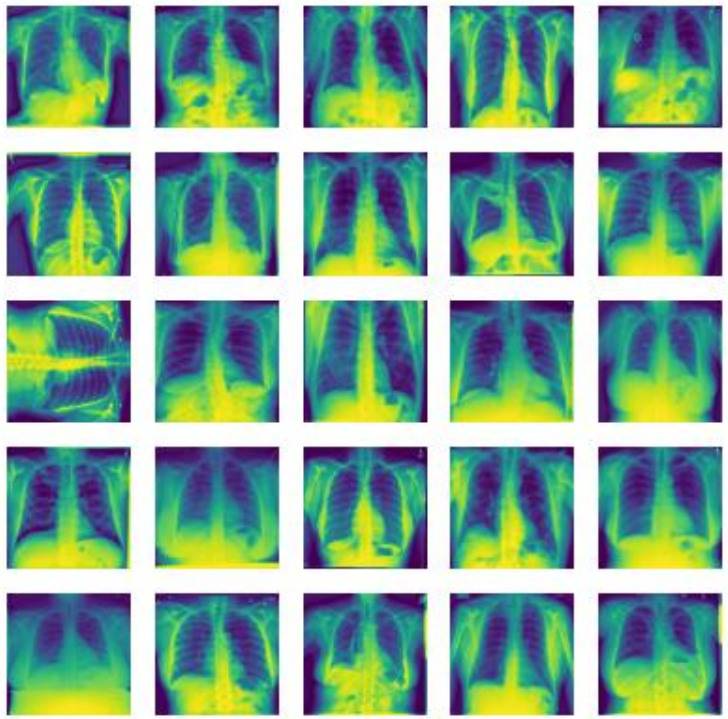
Impression: No evidence of active disease.

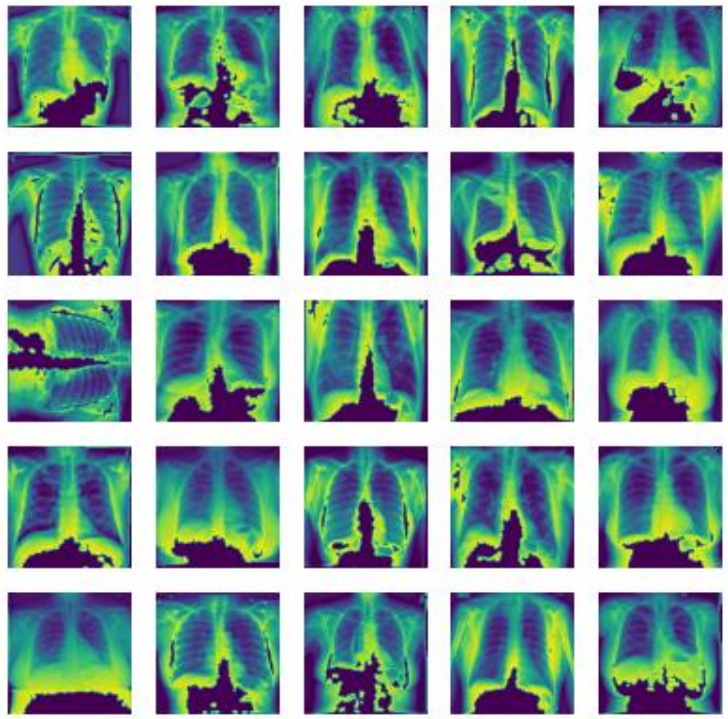
MTI tags: Deformity/thoracic vertebrae/mild

Preprocessing

1. Resize to 224×224
2. Conversion to grayscale
3. Histogram equalization
4. Histogram equalization + Adaptative masking







$$\theta = \min + 0.9 \cdot (\max - \min)$$

Encoder

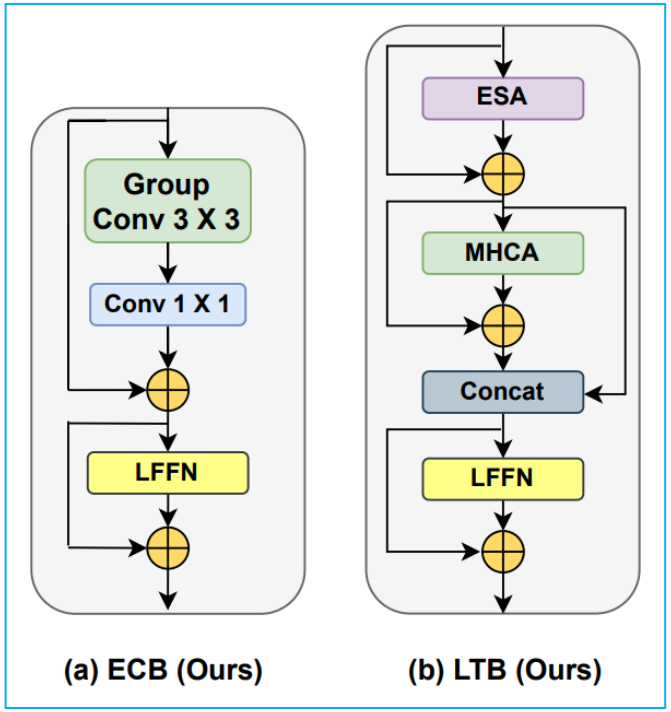
ELSEVIER

Computers in Biology and Medicine

Volume 157, May 2023, 106791

MedViT: A robust vision transformer for generalized medical image classification

Omid Nejati Manzari [a], Hamid Ahmadabadi [a], Hossein Kashiani [b], Shahriar B. Shokouhi [a], Ahmad Ayatollahi [a]



CNN-Transformer Hybrid Model

Encoder: MedViT

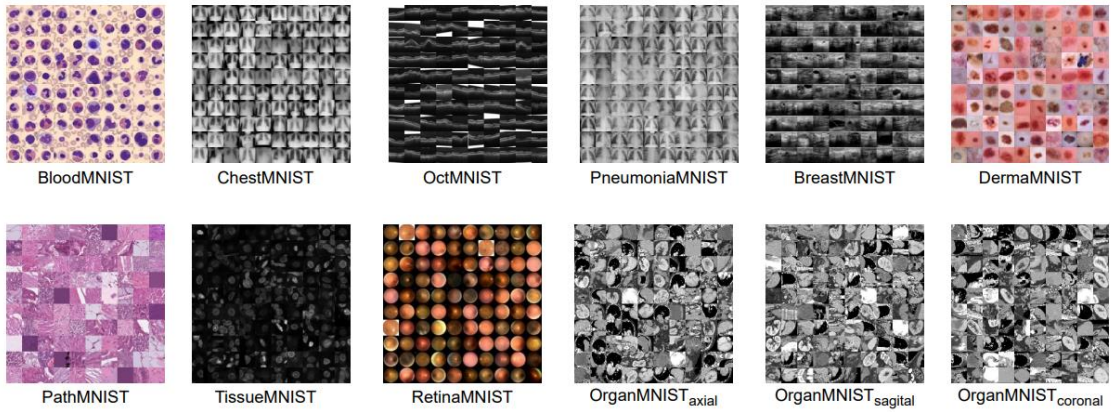
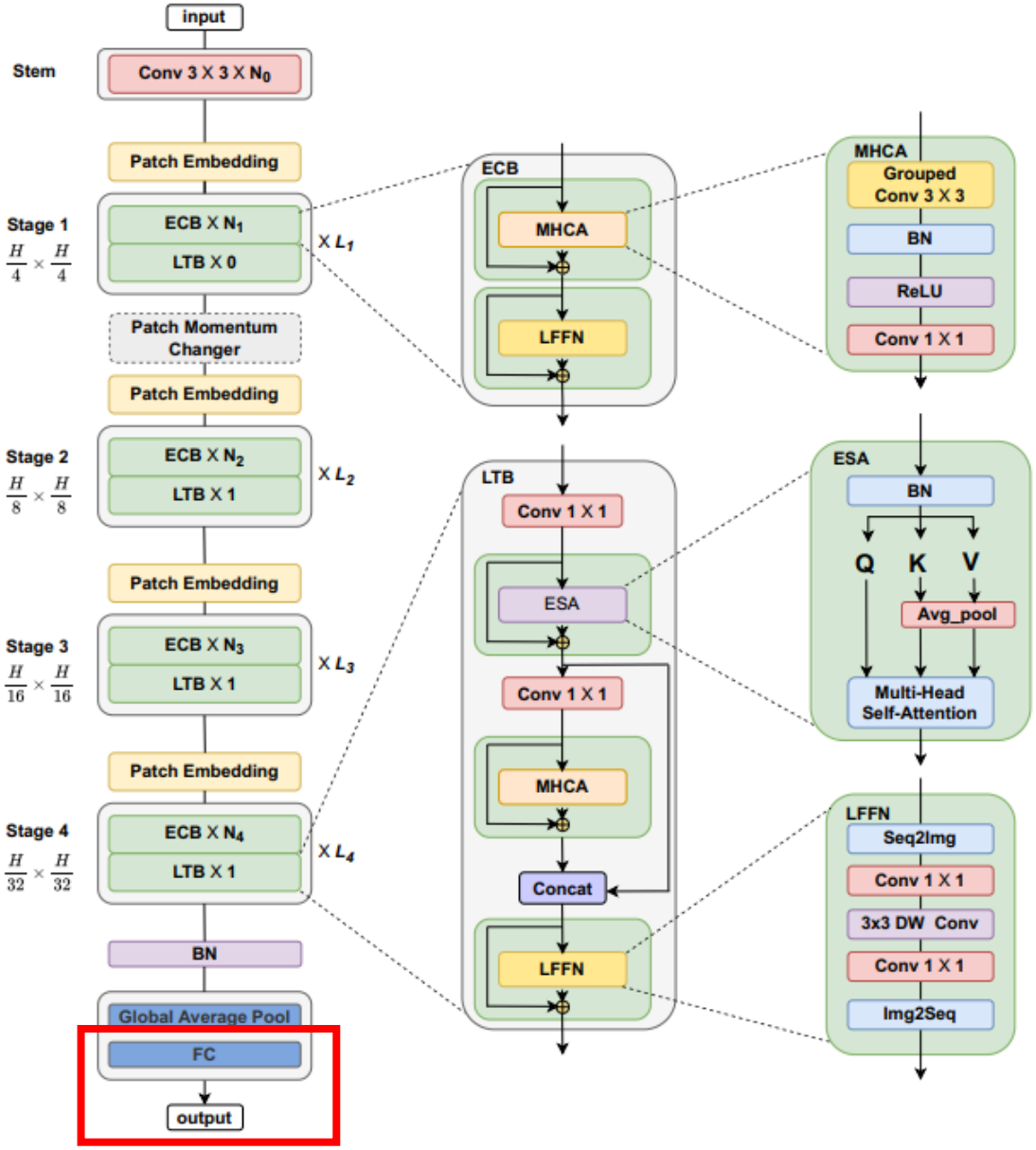


Figure 5: MedMNIST-2D Classification. MedMNIST is a collection of 12 pre-processed medical image datasets. It is designed to be educational, standardized, diverse and lightweight, which could be used as a general classification benchmark in medical image analysis.

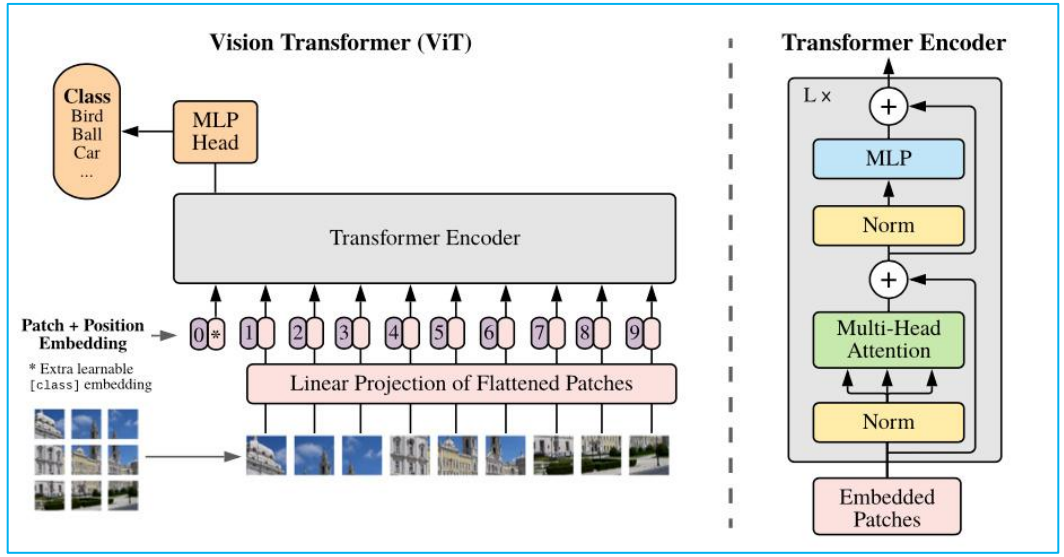


Standard ViT Model: To Connect to Decoder

google / **vit-base-patch16-224** ♡ like 356

Image Classification | Transformers | PyTorch | TensorFlow | JAX | Safetensors | imagenet-1k | imagenet-21k | vit

vision | Inference Endpoints | arxiv:2010.11929 | arxiv:2006.03677 | License: apache-2.0



AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy*,†, Lucas Beyer*, Alexander Kolesnikov*, Dirk Weissenborn*, Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby*,†

*equal technical contribution, †equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

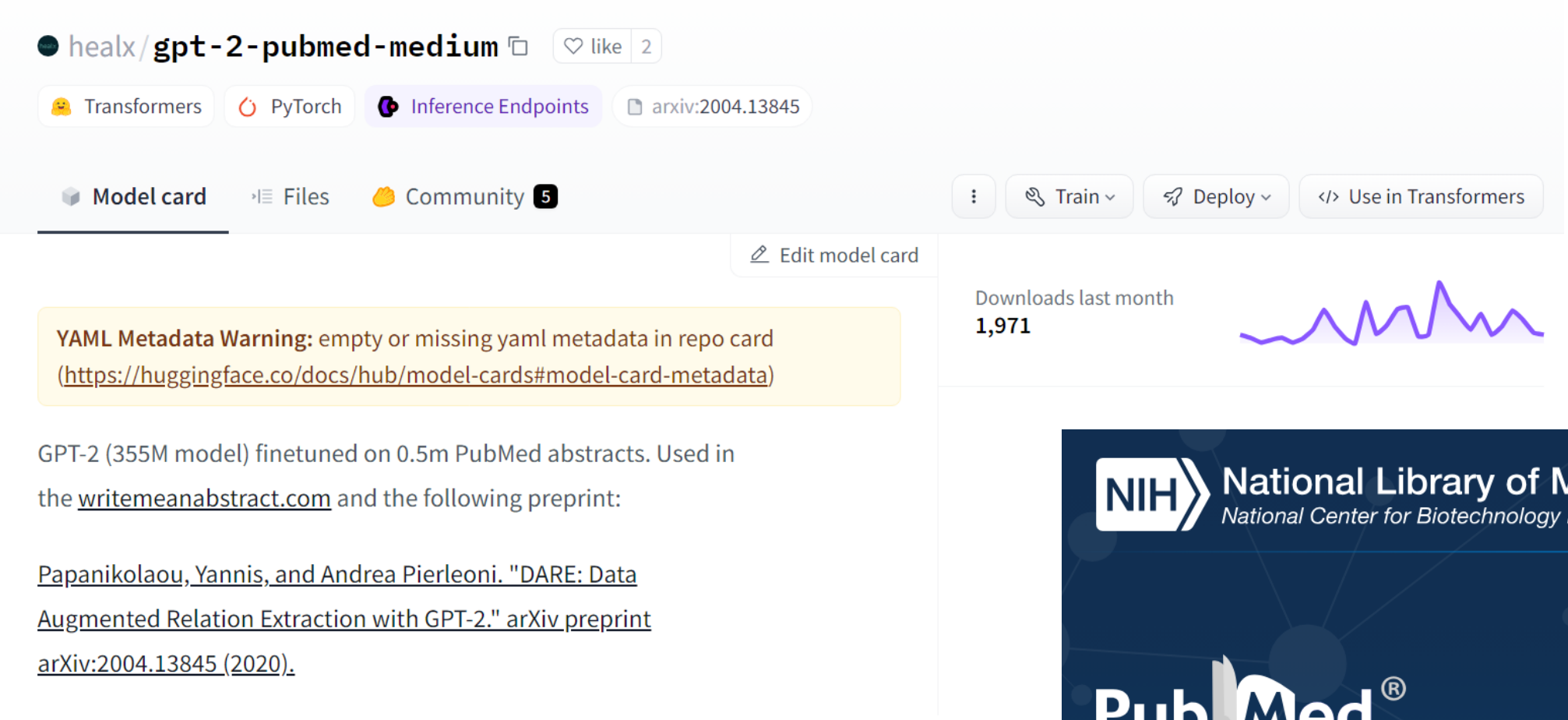
Decoder: GPT2

gpt2 ♡ like 1.46k

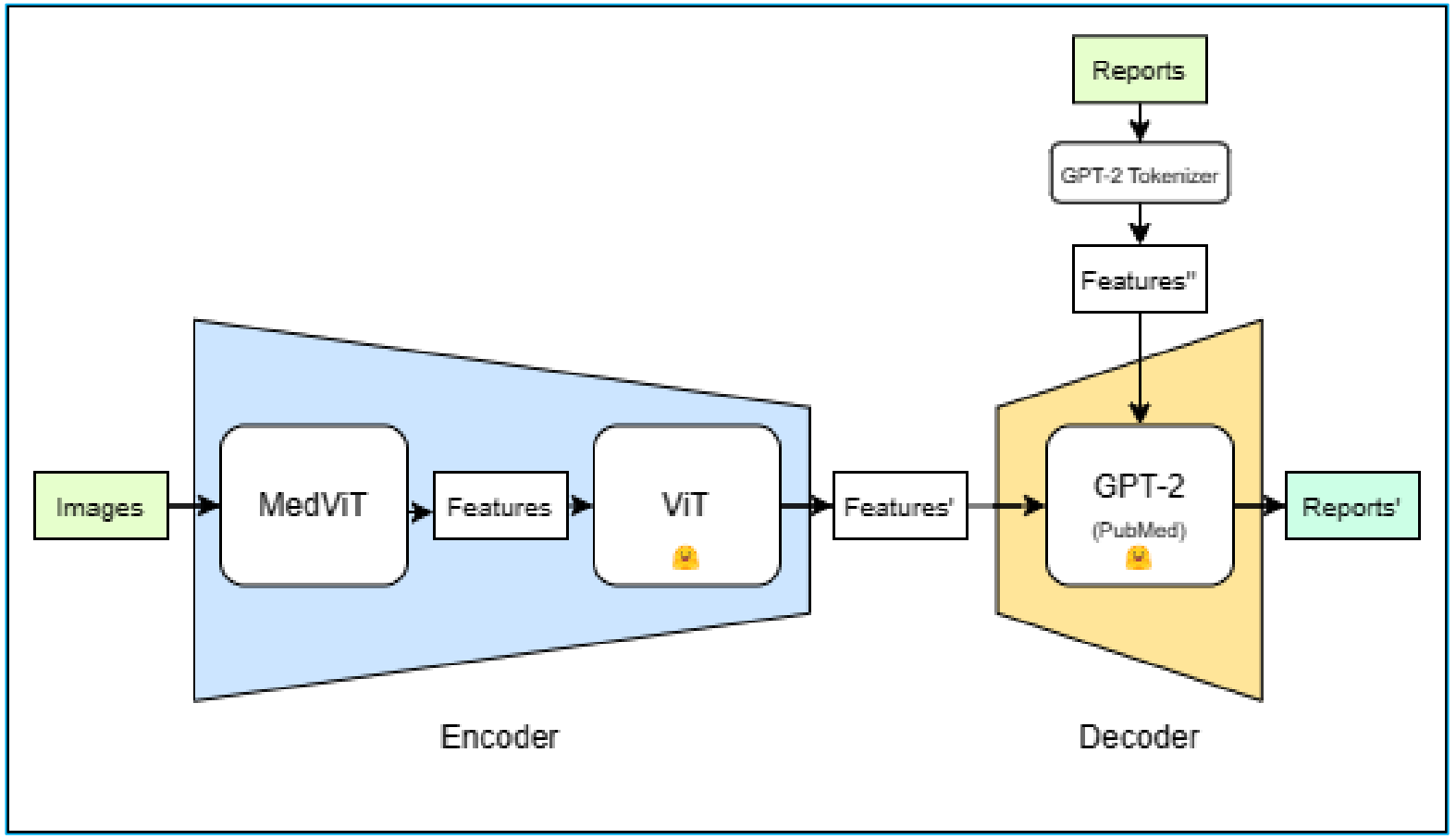
Text Generation · Transformers · PyTorch · TensorFlow · JAX · TF Lite · Rust · ONNX · Safetensors · English

doi:10.57967/hf/0039 · gpt2 · exbert · Inference Endpoints · text-generation-inference · License: mit

Decoder: GPT2 (PubMed)



Architecture: Vision Encoder Decoder Model



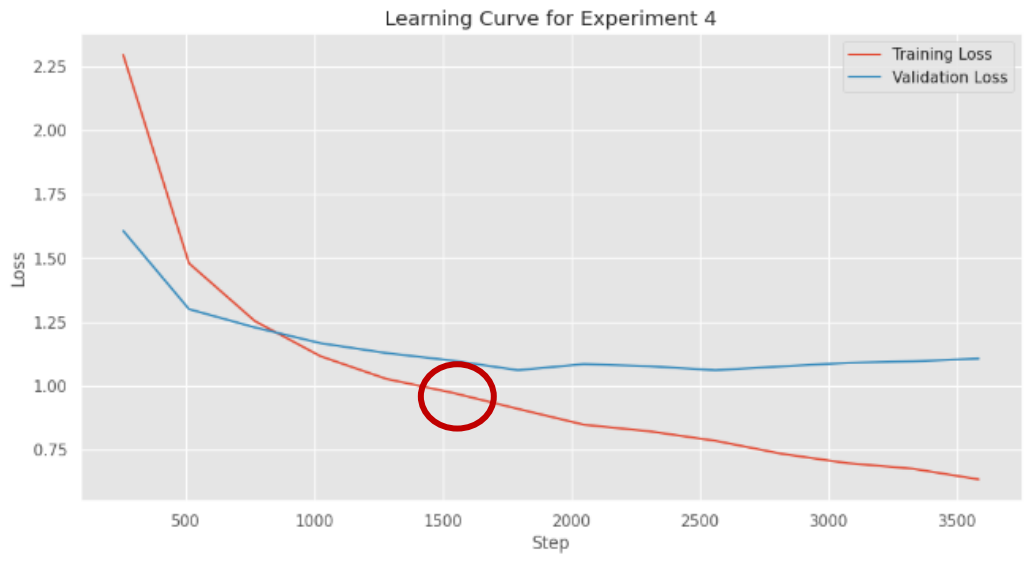
Experiments

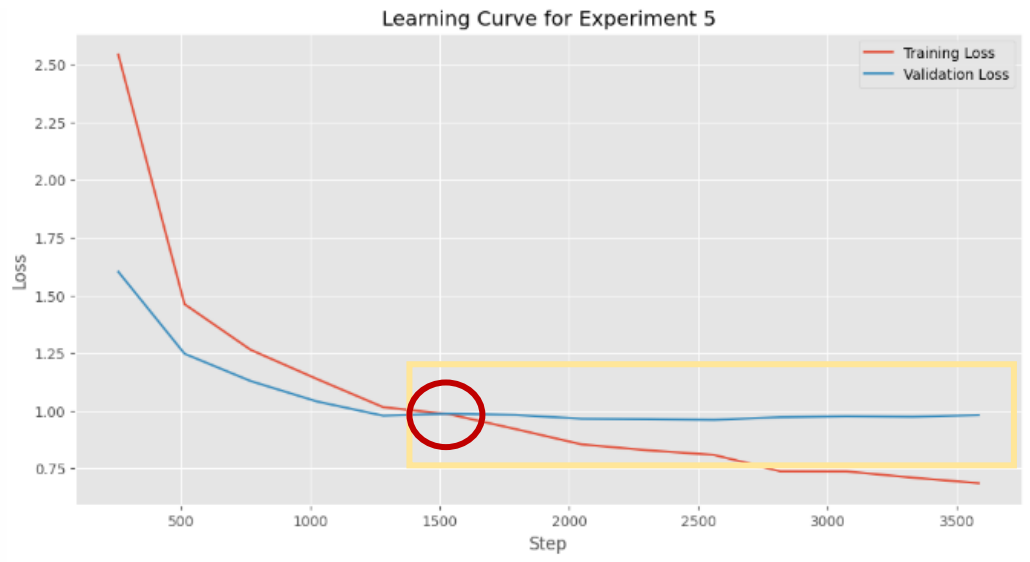
Exp.	Dataset (Size)	Encoder	Decoder	Epochs
1	Original (297)	ViT	GPT-2	10
2	Original (297)	MedViT + ViT	GPT-2	2
3	Original (297)	MedViT + ViT	GPT-2	10
Exp.	**Dataset (Size)**	**Encoder**	**Decoder**	**Epochs**
4	Original (3305)	MedViT + ViT	GPT-2	10
5	Equalized (3305)	MedViT + ViT	GPT-2	10
6	Equalized + Adaptative Masking (3305)	MedViT + ViT	GPT-2	10
7	Equalized + Adaptative Masking (3305)	MedViT + ViT	GPT-2 (PubMed)	10

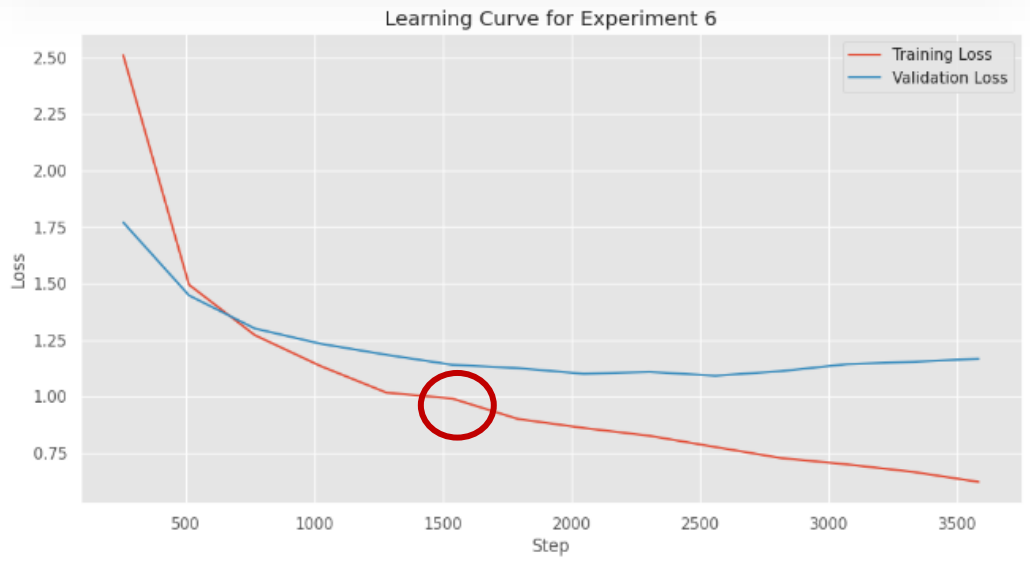
Preliminary Results

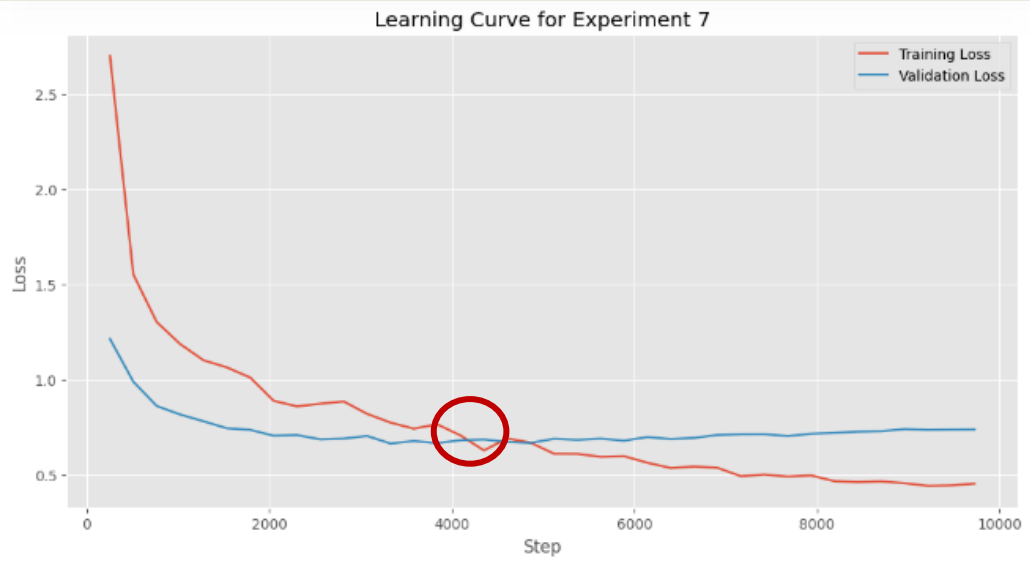
Exp.	Example Generated Caption
1	`I'm not going to lie to you," he said. "I'm just going to tell you what I think. I'm going to be honest with you. I don't think you're going to believe me." "I don't know what you're talking about," she replied. "You're just saying that you think I'm crazy. You don't believe me`
2	`"We are very pleased with the results of the study," said lead author Dr. Michael J. Schoenfeld, MD, professor of medicine at the University of California, San Francisco School of Medicine. "This is the first time that we have seen a significant increase in the incidence of acute myoca`
3	`The heart is normal in size and contour. The mediastinum is unremarkable. There is no pneumothorax or pleural effusion. No acute cardiopulmonary abnormality. No evidence of acute disease. 1. No focal airspace disease. 2. Mild pulmonary edema. 3. Mild nodule nodule disease.`

Learning Curves









#	ROUGE-L Precision	ROUGE-L Recall	ROUGE-L F-Measure
4	**0.18930**	0.20868	0.19831
5	0.18625	0.21916	0.20123
6	0.18402	**0.23543**	**0.20646**
7	0.18140	0.23156	0.20334

#	BLEU-1	BLEU-2	BLEU-3	BLEU-4
4	**0.29448**	0.09582	0.04283	0.02194
5	0.29150	0.09761	**0.04890**	**0.02785**
6	0.28563	**0.09765**	0.04639	0.02455
7	0.28250	0.09556	0.04704	0.02682

Metrics

Exp.	Generated Caption
4	`The heart is normal in size. The mediastinum is unremarkable. The lungs are clear. No acute`
5	`The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence`
6	`The heart and lungs have in the interval. Both lungs are clear and expanded. Heart and mediastinum normal. No active disease`
7	`The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.. Card`

Real Caption: `Heart size normal. Lungs are clear. are normal. No pneumonia, effusions, edema, pneumothorax, adenopathy, nodules or masses. Normal chest`

Real	Generated (Experiment 7)
Stable calcified hilar and granulomas. Lungs are clear bilaterally. There is no focal consolidation, pleural effusion, or pneumothoraces. Cardiomediastinal silhouette is within normal limits. are unremarkable. No acute cardiopulmonary abnormality.	The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the
Heart size mildly enlarged, stable mediastinal and hilar contours. No focal alveolar consolidation, no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax. Mild cardiomegaly, no acute pulmonary findings	The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion.. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are

Repeated Captions

Conclusions

- (HE, AM): No improvement seen in metrics nor quality of caption.
- (PubMed): No improvement seen in metrics nor quality of caption.
- (MedViT): Notable improvement in quality of caption vs ViT.
- Repeated captions: sign of overfitting, lack of diversity in real captions
- Possible Future Steps:
  - Try with large version of PubMed model
  - Try other medical domain language models: Meditron, Med-PaLM
  - Experiment with different parameter configurations: batch size, maximum caption length, temperature
  - Apply regularization to tackle overfitting
- Weaknesses and Mistakes:
  - Lack of understanding of NLP fundamentals: tokenizer setup, text pre-processing, text representations, metrics calculation
  - Lack of expertise with HuggingFace library: deprecations, setups
  - Relatively small dataset: lack of data augmentation
  - Oversight in not using a fixed test dataset for the final experiments

References

[1] Beddiar, R., & Oussalah, M. (2023). Explainability in medical image captioning. In J. Benois-Pineau, R. Bourqui, D. Petkovic, & G. Quénot (Eds.), Explainable deep learning AI (pp. 239-261). Academic Press. https://doi.org/10.1016/B978-0-32-396098-4.00018-1

[2] Beddiar DR, Oussalah M, Seppänen T. Automatic captioning for medical imaging (MIC): a rapid review of literature. Artif Intell Rev. 2023;56(5):4019-4076. doi: 10.1007/s10462-022-10270-w. Epub 2022 Sep 17. PMID: 36160365; PMCID: PMC9483422.

[3] Giełczyk A, Marciniak A, Tarczewska M, Lutowski Z. Pre-processing methods in chest X-ray image classification. PLoS One. 2022 Apr 5;17(4):e0265949. doi: 10.1371/journal.pone.0265949. PMID: 35381050; PMCID: PMC8982897.

[4] Nejati Manzari, O., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., & Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. Computers in Biology and Medicine, 157, 106791. https://doi.org/10.1016/j.compbiomed.2023.106791.

[5] Monajatipoor M, Rouhsedaghat M, Li LH, Kuo CJ, Chien A, Chang KW. BERTHop: An Effective Vision-and-Language Model for Chest X-ray Disease Diagnosis. Med Image Comput Comput Assist Interv. 2022 Sep;13435:725-734. doi: 10.1007/978-3-031-16443-9_69. Epub 2022 Sep 16. PMID: 37093922; PMCID: PMC10120542.

[6] https://github.com/este6an13/transformers-image-captioning