Team 4 - ViLT

Abstract

Visual Question Answering (VQA) has gained significant attention in recent years as it combines computer vision and natural language processing to enable machines to answer questions about images. In the medical field, VQA holds the potential to assist healthcare professionals in diagnosing and understanding radiology images.
This paper presents a detailed write-up of a Visual Question Answering model fine-tuned on the VQA-RAD dataset using the Vision-and-Language Transformer (ViLT). We explore this task's essential components and processes, including dataset preparation, model architecture, training procedures, and evaluation metrics. The proposed model demonstrates its potential to answer questions about medical radiology images with reasonable accuracy.

Implementation

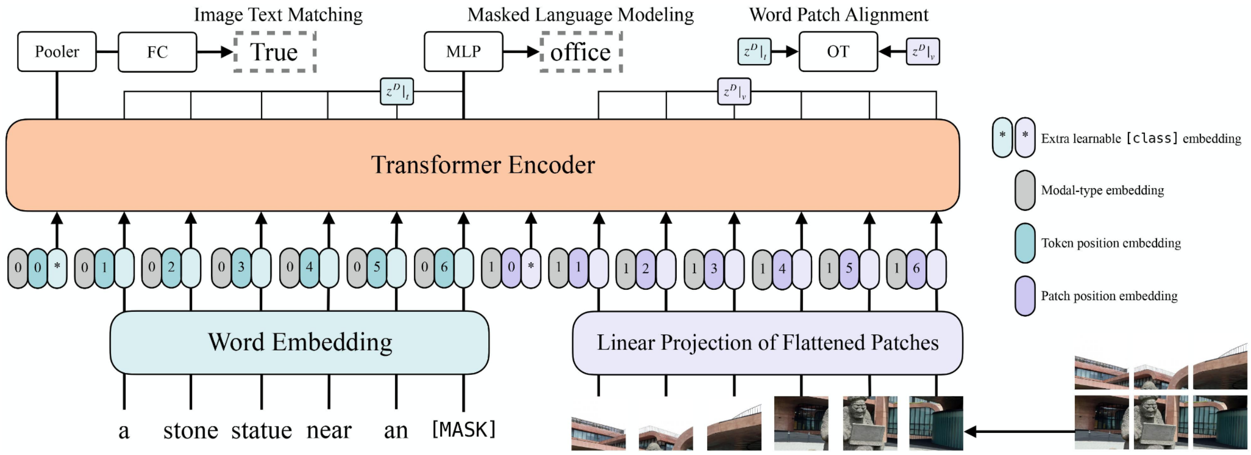
Approach 1

ViLT Transformer

The ViLT model was proposed in *ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision by Wonjae Kim, Bokyung Son, Ildoo Kim*. ViLT incorporates text embeddings into a Vision Transformer (ViT), allowing it to have a minimal design for Vision-and-Language Pre-training (VLP).

We chose ViLT over other transformers because ViLT, unlike other transformers, uses lightweight and fast embedding of visual inputs. Even in recent literature, most VQA studies have focused on heavy visual embedders.ViLT employs shallow, convolution-free embedding of pixel-level inputs. Removing deep embedders solely dedicated to visual inputs significantly cuts down the model size and running time by design. The merits of this choice can be summarized in three major points:

- ViLT is the simplest architecture by far for a vision and language model as it commissions the transformer module to extract and process visual features in place of a separate deep visual embedder. This design inherently leads to significant runtime and parameter efficiency.
- It performs competently on vision-and-language tasks without using region features or deep convolutional visual embedders.
- Also, for the first time, we empirically show that whole word masking and image augmentations that were unprecedented in VLP training schemes further drive downstream performance.



The model was trained on a **high-performance computer -** we trained for 100 epochs

Data Augmentation

The VQA-RAD dataset contains a limited number of data points,3515 to be precise, which is inadequate. Therefore, we applied various data augmentation techniques to boost the training set. In a dataset of such a small magnitude, synthetic sample generation becomes crucial to improving the results. We employed augmentation techniques for both the images as well as the questions.

To boost the image samples, we applied simple image augmentation techniques like flipping, rotation, adding Gaussian Blur to images, and RAND-Augment (as suggested by authors) so as to add noise and regularisation to the model. It significantly added to the robustness of the model.

To augment the questions, we added paraphrased versions of the questions. The dataset also came with paraphrased versions of the question, which were concatenated to the training set after some preprocessing. We also employed pre–trained transformers to generate paraphrased samples of questions as well as techniques discussed in various other papers such as TextAttack[2], EDA[3], NLPaug[4], and Backtranslation[5].

We introduce a novel approach for generating synthetic data by employing an Image Captioning system[6] to produce question-answer pairs through the creation of image descriptions. This innovative pipeline enhances dataset diversity and facilitates training for vision-language tasks. A Visual Question Generation pipeline can also be employed.

Model Architecture

Fine Tuning

In order to work with our limited dataset, we chose a ViLT model already finetuned on the VQAv2 dataset. We observed significantly better results with this approach.
We also aimed to use embeddings generated by sci-BERT for the model.

Rather than using the classifier head of ViLT [1], we tried using a decoder of a transformer based on sciBERT.

To generate the final answers, we experimented with an RNN network and an LSTM network to generate sentences based on the embeddings received from our transformer. However, the results were not semantically coherent, and we decided to move to an alternate appraoch.

We then tried to use the decoder of sciBERT to generate coherent answers, but sciBERT, being a huge pre-trained model, did not gain significant insights from our limited dataset, so we dropped it.

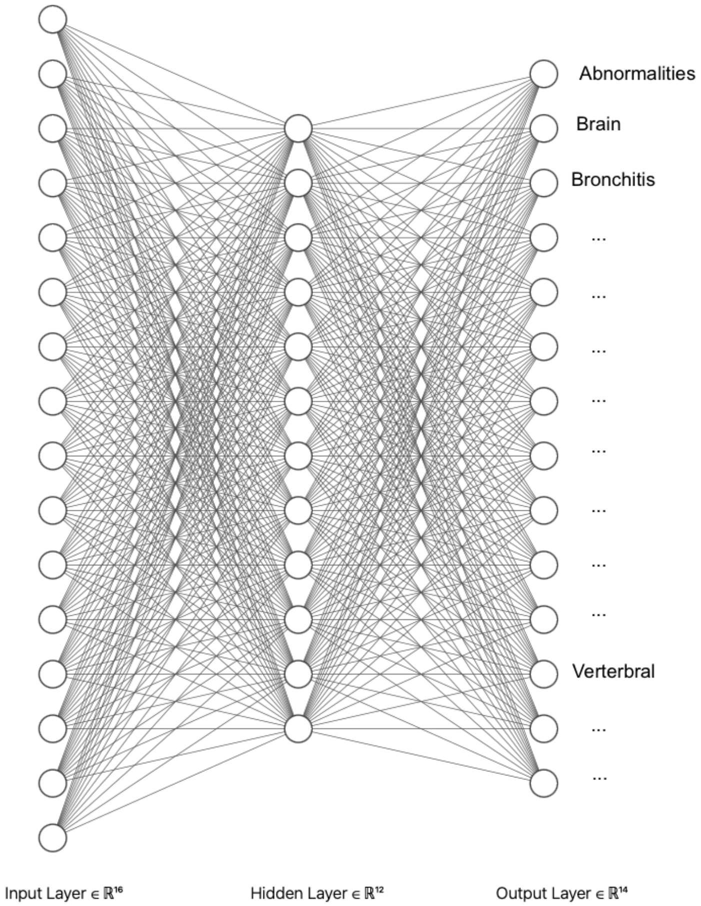
Soft Encoding

We realized that most of the answers were yes or no based. Very few unique words were encountered in the dataset. Therefore, we decided to create our own dictionary comprised of medical terms and radiology-related words. We did some comprehensive web scraping to generate a corpus of medical terms and radiology-related terms, which were then employed to create labels.

In a multi-word answer, all words are equally important. Therefore, we employed a soft encoding schema. If the answer is, say, “pleural effusion”, we assign equal weights (0.5 each) to both the words.
We employed the formula :

$$Score_{\text{each word}} = {}^{1}\!/_{\text{length of answer}}$$

to generate soft encodings.
This, along with using our own dictionary for tackling the question-answer pairs significantly improved the model’s ability to answer open-set questions.
As a result, despite being trained for multiclass-classification, the model will be able to handle open-set questions as well.



Modified Classifier Head

Our main goal was to predict the right words for the sentence. The model predicts the answers as individual words as its predictions for the multiclass-classification problem. As a result, the model suffers from grammatical inaccuracies, regardless, this may simply be fixed by passing the output to another transformer that can create coherent sentences.

Loss Function

We also tried to implement our own loss function. Integrating a supervised contrastive loss function into the training process of a Visual Question Answering (VQA) system holds several compelling advantages. At its core, this technique refines the model's ability to learn and represent the intricate relationships between images and textual questions. By explicitly defining positive and negative pairs, it fosters the creation of more discriminative embeddings, leading to an improved understanding of the visual and textual content.

This enhancement in feature learning is pivotal, allowing the VQA system to capture fine-grained details and intricate associations between different modalities. Moreover, this approach helps mitigate data imbalance concerns often encountered in VQA datasets. It promotes robustness by training the model to focus on relative similarities and dissimilarities, reducing its sensitivity to answer distribution disparities.

Furthermore, the acquired embeddings can be transferred to a variety of other vision-language tasks, amplifying the model's versatility and usefulness in a broader context. This capability to adapt to different tasks while maintaining performance demonstrates its transferability. This is particularly valuable in scenarios where vision and language understanding are integral, such as image-text retrieval, image captioning, and more.

Supervised contrastive loss functions also act as a form of regularization, reducing overfitting and enhancing the model's generalization abilities. This reduction in overfitting minimizes the model's propensity to memorize training data and allows it to adapt more effectively to novel, unseen data. This is especially valuable in scenarios where VQA models need to handle variations in question phrasings, image conditions, and other factors.

The advantages extend to fine-tuning and few-shot learning, enabling the model to adapt swiftly to new data or tasks with minimal examples. The learned embeddings and the associated understanding of semantic relationships facilitate a quicker adaptation process and improved performance in these scenarios.

Further Improvements

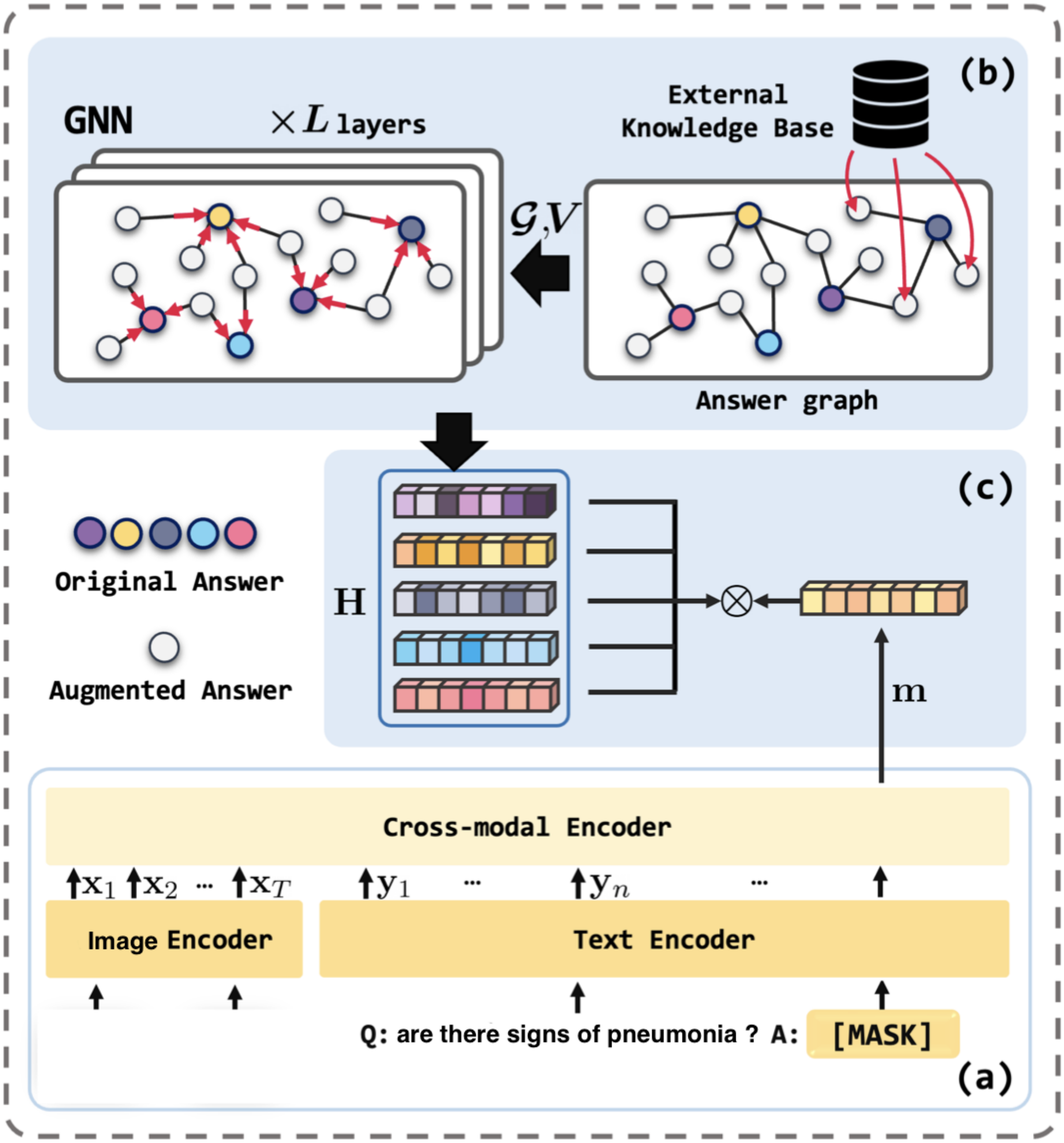
1. GNN based soft-verbalizer[7]:
   The soft verbalizer leverages a large language model's pre-trained knowledge and external resources like GloVe and ConceptNet to help the model better predict these answers. It does this by associating words in the answer with similar words from the external knowledge base, making it more versatile.

   To sum up the two main concepts:

   - Verbalizer: In language models like BERT and GPT, they are often fine-tuned for specific tasks. The traditional approach discards some pre-trained information, but the soft verbalizer retains it by using a "fill-in-the-blank" approach, making it more effective for certain tasks.

- Graph Neural Networks (GNNs): These are used to understand relationships between words. If words are similar, they should be treated similarly. GNNs help the model learn from similar words and improve the handling of rare or unknown answers by considering their "neighborhood" in the language space.

Thus the soft verbalizer will be able to improve models capability to handle out of vocabulary answers even more.

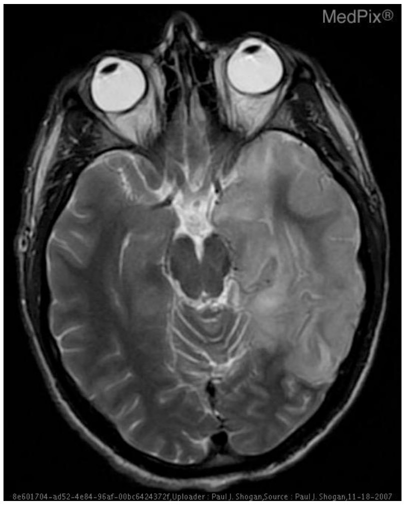


Overall architecture. (a) Image-question encoding: a image-question pair is first encoded through a backbone architecture and the output feature of [MASK] token, m ∈ R_D, is extracted. (b) GNN-based soft verbalizer: an answer graph is constructed with both original answers and their augmented words from an external knowledge base, and GNN aggregates their information. (c) Similarity calculation: we finally calculate the similarity (denoted as ⊗) between smoothed answer embeddings H_{train} (or H_{test}) and [MASK] token output feature m.

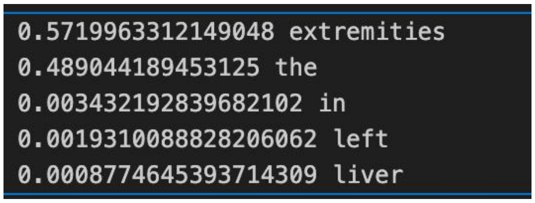
Results

We ran the model on some sample images, and the model performed remarkably well. Here are some examples

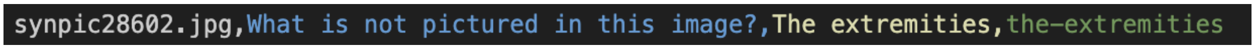
Closed Set Questions (From the test set)


MedPix®
8e601704-ad52-4e84-96af-00bc6424372f,Uploader : Paul J. Shogan,Source : Paul J. Shogan,11-18-2007

Question: what is not pictured in this image?

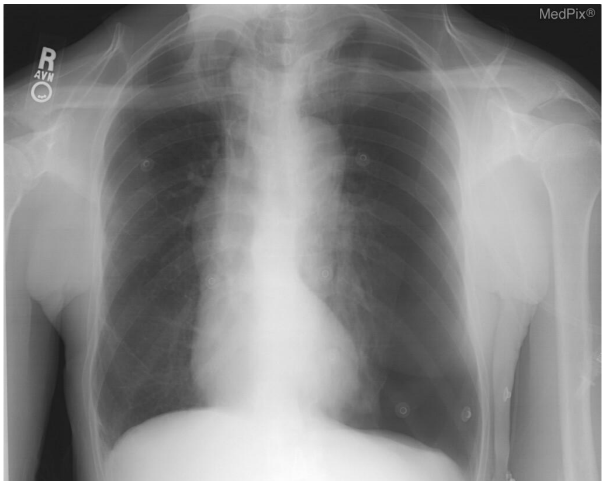


```
0.5719963312149048 extremities
0.489044189453125 the
0.003432192839682102 in
0.0019310088828206062 left
0.0008774645393714309 liver
```

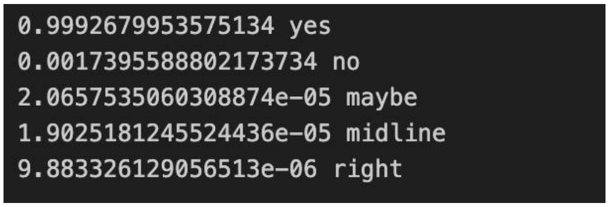


```
synpic28602.jpg,What is not pictured in this image?,The extremities,the-extremities
```

Open Ended Questions



Question: is this a male body?



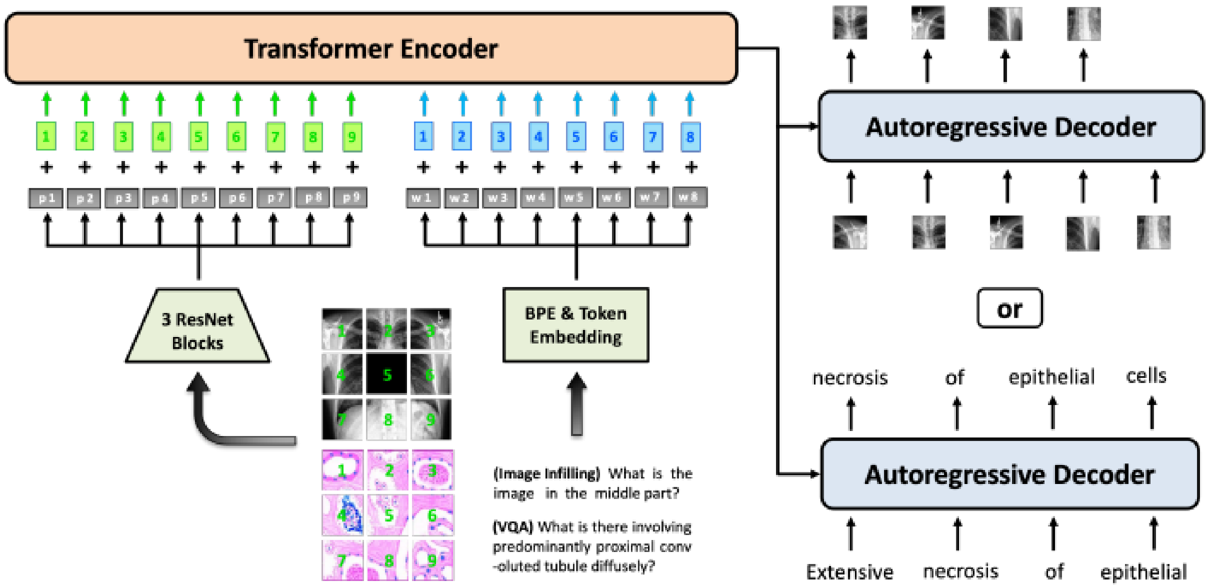
```
0.9992679953575134 yes
0.0017395588802173734 no
2.0657535060308874e-05 maybe
1.9025181245524436e-05 midline
9.883326129056513e-06 right
```

The model correctly predicts that its a male body.

Approach 2

Biomed-GPT

Apart from working on our from-scratch approach, we also performed a thorough literature survey and examined various state-of-the-art models. The best we found so far was Biomed-GPT. It is a state-of-the-art model which leverages self-supervision on large and diverse datasets to accept multi-modal inputs and perform a range of downstream tasks. We used it to perform VQA.



(Image Infilling) What is the image in the middle part?

(VQA) What is there involving predominantly proximal conv -oluted tubule diffusely?

We fine-tuned BiomedGPT on the VQA-RAD dataset, however, we did not see any significant improvements since it has already been trained on a huge corpus of medical VQA. BiomedGPT is definitely the best-performing model so far, but it does not reflect any significant contribution from our end, therefore, we have not made it our primary approach.

Alternate Approaches

Here are some other transformers that we found and considered but ultimately, we decided to go ahead with ViLT due to a faster, and better performing architecture.
VisualBERT
Vision Text Dual Encoder

Literature Survey

In order to select the best possible architecture for a task, a comprehensive literature survey is warranted. We discovered multiple state-of-the-art models, the best ones we encountered so far are listed below;

- PMC-VQA

It is a generative-based model for medical visual understanding by aligning visual information from a pre-trained vision encoder with a large language model. Existing approaches have primarily treated medical VQA as a classification problem, with the goal of selecting the correct answer from a candidate. Consequently, this approach limits the system's utility to predefined outcomes, hampering its free-form user-machine interaction potential. PMC-VQA takes an alternative approach, with the goal to generate an open-ended answer in natural language.

Specifically, the system is trained by maximizing the probability of generating the ground-truth answer given the input image and question.

- PMC-CLIP

For the visual and text encoders, PMC-CLIP uses ResNet50 and PubmedBERT and 4 transformer layers for the fusion module. For input data, each image is resized to 224 × 224. During pre-training, the text encoder is initialized from PubmedBERT, while the vision encoder and fusion module are trained from scratch.

- Med-VQA

Med-VQA is a project that leverages Stacked Attention Networks[10] for Visual Question Answering. SANs use semantic representation of a question as query to search for the regions in an image that are related to the answer. The image is queried multiple times through the stacked layer to infer the answer progressively.

- LLaVA-Med

The key idea behind LLaVA-Med is to leverage a large-scale, broad-coverage biomedical figure-caption dataset extracted from PubMed Central. This dataset is used to instruct GPT-4 to generate open-ended instruction-following data from the captions.
The model is then fine-tuned using a novel curriculum learning method. Specifically, the model first learns to align biomedical vocabulary using the figure-caption pairs as is, then learns to master open-ended conversational semantics using GPT-4 generated instruction-following data. This broadly mimics how a layperson gradually acquires biomedical knowledge. This approach enables the training of LLaVA-Med in less than 15 hours

References

[1] *ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision Wonjae Kim, Bokyung Son, Ildoo Kim*

[2] *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, Yanjun Qi*

[3] *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks Jason Wei, Kai Zou*

[4] NLPAug

[5] *Back Translation Survey for Improving Text Augmentation Matthew Ciolino, David Noever, Josh Kalin*

[6] https://github.com/ashishthomaschempolil/Medical-Image-Captioning-on-Chest-X-rays/tree/main

[7] https://arxiv.org/pdf/2308.09363v1.pdf

[8] *BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Et. Al.*

[9] *Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

[10] *Stacked Attention Networks for Image Question Answering Zichao Yang, Xiaodong He , Jianfeng Gao , Li Deng , Alex Smola*