Image Captions Generation for Medical Data

Ananya Devarakonda ad834
Mariia Dobko md927
Rhia Singh rts229
Patrick Mazza pmm259

June 3, 2023

Abstract

Image captioning is a popular task in machine learning that deals with generating text given images. Our project explores this field in the medical context. Our aim is to experiment with different datasets and models for image captioning in healthcare. In this report, we outline results from three models - Show, Attend, and Tell, a classification model, and Contrastive Language-Image Pre-training (CLIP), using different datasets - MedICaT, IU-Chest X-ray, and SIIM-ACR Pneumothorax. Our findings show that each method is promising with the Show, Attend, and Tell model having a BLEU score of 0.5957 on the test dataset.

1 Introduction and Background

The goal of our final project is to build a system for image caption generation using medical images. Image captioning in healthcare can benefit medical reporting and Electronic Health Records (EHR) standardization. The prospect of combining visual and textual data has shown great results in past research studies which leads us to believe that we can yield promising findings in our use case. We classify our project as an application in the field of healthcare. In this paper we focus on two approaches for image captioning: a popular baseline method called Show, Attend, and Tell(10) and a state-of-the-art model - CLIP(7) proposed by OpenAI. Our experiments include three publicly available datasets, namely IU-Chest X-Rays(6), MedICaT(9), and SIIM-ACR Pneumothorax(11). Through this work we aim to discover and report the challenges of training image captioning models on medical data. We support our findings with qualitative and quantitative results.

Magnetic Resonance Imaging (MRI) and Computerized Tomography (CT) Imaging techniques are commonly used in medicine. An MRI uses a magnetic field and computer-generated radio waves to create detailed images of organs and tissues (5). A CT scan creates cross-sectional images (slices) of the bones, blood vessels, and soft tissues through multi-angle X-Rays images (1). Common deep learning models for image captioning include non-attention-based models, like Convolutional Neural Network (CNN), and image-text early fusion (BERT-like) models. Most state-of-the-art image captioning architectures combine object detection techniques with modern text embeddings such as BERT(8). Using such methods, that merge Computer Vision with Natural Language Processing on MRI and CT scans, images can be captioned and synergy can happen between the two fields.

2 Datasets

IU-Chest X-Rays Dataset: The Indiana University (IU) dataset (6) consists of 7,470 image files and 3,955 medical notes containing findings, indications, and impressions on any given patient. Exploratory analysis shows that all images have a width of 512 pixels. The minimum height is 362 pixels, the average height of an image is 533 pixels and the maximum height is 873 pixels. There are 3,425 findings, so some images do not have a corresponding finding. An example of a finding from the IU-Chest X-ray data set is "Heart size normal Lungs are clear XXXX are normal No pneumonia effusions edema Pneumothorax adenopathy nodules or masses."

The Histogram Density Estimation of words distribution shows a right skewed distribution, further supported by the median being 29 words and the average being 55 words. The histogram does

not address two areas: not smooth and the shape depends on the bin position. The Kernel Density Estimation (KDE) results in a smoothed histogram with a Gaussian kernel and bandwidth of 0.8.

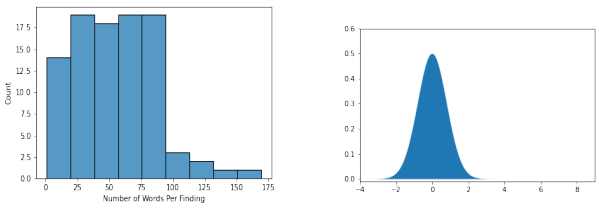

Figure 1: Distributions of Words in Findings a) Histogram Distribution, b) Kernel Density Estimation

MedICaT Dataset(9): Open-source data set with over 200,000 images, captions, inline references, and manually annotated sub figures. For images, the minimum shape is (144, 54, 3) and the maximum is (5950, 7012, 3). The captions in MedICaT have on average 59 words per sample, the median is 47 words, with a minimum of 6 and maximum of 839.

3 Analysis

Data exploration

We explore the captions from the IU Chest X-Rays dataset by using a wordcloud technique. Our observations show that such words as "lung", "pneumothorax", "pleural effusion" are the most repetitive, refer to the Figure 2 (a).

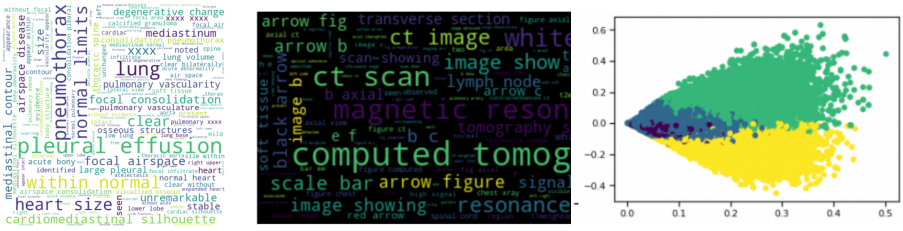

Figure 2: a) IU-Chest wordcloud b) MedICaT wordcloud c) K-means Clustering MedICaT

As we had hardware limitations, training on the entire dataset would have been a challenge. Therefore, we had to determine how to effectively sample the dataset so as to justify the dataset split. To explore patterns in the dataset, we used K-means clustering. The result of using PCA to visualize K-means clustering with K=4 is illustrated in Figure 2 (c). We then generated wordclouds for each cluster and found that one cluster predominantly contained words related to Computed Tomography (CT), as shown in Figure 2 (b), and another cluster contained words related to magnetic resonance imaging (MRI). This exploration informed us to partition the data into two subsets: CT and MRI.

Data Preprocessing

We preprocessed data from both the IU Chest X-Rays and MedICaT by first removing outliers. We excluded images smaller than 64 pixels and larger than 1024 pixels. For captions, we used the 1.5 interquartile rule to remove outliers. We then used standard Natural Language Processing (NLP) preprocessing techniques such as converting to lowercase, removing stop words, lemmatization, and removing punctuation.

For the preliminary experiments, we chose 5,000 random samples from each MedICaT subset and the IU Chest X-Rays dataset. We then split the data by keeping 70% of the total cases for the train set, 20% for validation, and 10% for test. The result was that the training data accounted for 3,500 samples, while 1,005 and 495 were used for the validation and the test accordingly.

4 Methods

Show, Attend, and Tell (10)- As the name suggests, the Show, Attend, and Tell model, is, in brief, a three-step methodology. The first step in the process is to encode the input image using a CNN. The next step is to generate a caption word by word given the encoded image using a Long Short-Term Memory (LSTM) model. The primary idea introduces a mechanism that generates weights that provides perspective into the relative importance of a location in the image using weighted averaging across pixels, "attention mechanism". We use this method as our baseline.

CLIP(7) - Contrastive Language - Image Pre-training that provides a zero-shot transfer to downstream datasets predicting the caption from the concepts in a given zero-shot classifier. This method uses the text paired with images to model the proxy training task: jointly training an image encoder and a text encoder to predict the correct pairings. To apply CLIP to a new task, we need to encode the names of the task's visual concepts, and CLIP will output a linear classifier of visual representations. For the model architecture, please refer to Figure A1 in the Appendix. In our experiments we use ResNet50 as an encoder within CLIP.

5 Results

5.1 Results of Show, Attend and Tell

Our preliminary experiment for the Show, Attend, and Tell model (10) was to check its performance on the well-known Flickr8k dataset as shown in Figure 3. We provide a visual illustration of some predictions in Appendix A.3. This experiment was to demonstrate that the Show, Attend, and Tell model was able to perform and for us to estimate the training and evaluation time. We referred to the official Tensorflow tutorial for the Show, Attend, and Tell model and tweaked it to train on the MedICaT dataset (9).

Initially, we trained the Show, Attend, and Tell model consisting of a MobileNetV3 (3) encoder and a transformer with attention decoder, on 5,000 samples of CT images in the MedICaT dataset for 100 epochs with early stopping. However, we found that the model was heavily overfitting. The training curves for this result are seen in Appendix A.4. In particular, the model overfit and over-predicted commonly occurring words such as "CT", "computed tomography", "figure." It can be hypothesized that this result is due to such words appearing several times in the sample of the dataset we chose. We observed a similar trend when training on 5,000 samples of MRI images.

Combatting Overfitting: As training on the same subset of MedICaT data (CT/MRI) resulted in overfitting, we used the following strategies to improve validation performance:

1. Training on a stratified random sample of 10,000 images containing an equal number of CT and MRI images. The idea was to not only add representation on the MRI subclass but also to increase the number of training samples to help prevent overfitting.
2. Training on a stratified random sample of 10,000 images on all 4 clusters found after K-means clustering. This data sample was not restricted to data from one or two particular classes but had equal representation from all four clusters.
3. Testing the EfficientNet as an encoder instead of MobileNetV3 as well as adding a dropout layer and setting dropout to 0.5.

Appendix A.2 provides a glimpse of the entire model architecture. We used Adam Optimization and `sparse_softmax_cross_entropy_with_logits` as loss for the final model. We trained for 100 epochs with early stopping and a learning rate of 1e-4. Figure 4 illustrates the loss and accuracy curves obtained while training the model on data from all four clusters. While the model still slightly overfits, we observe that the model is correctly able to distinguish between X-Rays, CT, MRI, and histopathology images.

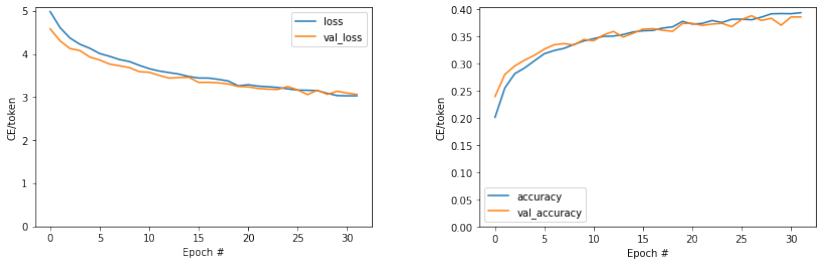

Figure 3: Flickr8k dataset training curves: masked Loss (a) and accuracy (b)

Figure 5 is an example test image result from the Show, Attend, and Tell model. Clearly, it recognizes that the image is of a brain MRI. In this case, the predicted caption is "figure 1 axial t2weighted mr image show multiple area high signal intensity lesion arrow," and the actual caption is "Fig. 2 T2-weighted axial images of the brain show mixed signal intensity lesion in right posterior parieto-occipital lobe with areas of hypointensity." The model captures the initial parts of the original caption and avoids punctuation and stop words as we preprocess the raw caption before training the model.

Data	BLEU Score
CT + MRI subsets	0.4797
MedICaT (all clusters)	0.5957

Table 1: Show, Attend, and Tell BLEU metrics on MedICaT data

To further illustrate how variations in the dataset help prevent extreme overfitting, Table 1. provides the BLEU score on the test set for a model trained on the CT and MRI cluster data and a model trained on all four clusters with stratified random sampling. As the BLEU score drastically increases, we may conclude that with more training examples, the better the model performance.

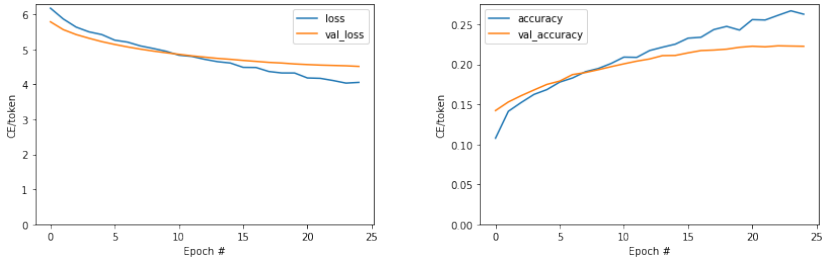

Figure 4: MedICaT dataset training curves: masked Loss (a) and accuracy (b)

5.2 Classification as an alternative

Based on a preliminary analysis of the content in medical notes from the IU dataset, we noted that 1,818 (73% of the findings) mention "Pneumothorax" as a medical professional's aim of identifying an ill patient. Recognizing the medical need, we aimed to classify the patient images from the IU dataset as either Pneumothorax positive or negative using a pre-trained ResNet50 (2) as demonstrated in other Pneumothorax classifications. This can be viewed as an alternative approach to image captioning. In this scenario, the model predicts a particular characteristic, which could be included in the medical report.

The majority of the findings (1,804) correspond to a negative class. To mitigate the class imbalance, the ResNet50 is trained on both IU and SIIM-ACR Pneumothorax chest X-Rays images. In particular, we use 2,500 positive samples from SIIM-ACR Pneumothorax to enrich our Pneumothorax samples from IU-Chest X-Ray that contains only 20 images with the disease. We use the ResNet50 architecture with weights pre-trained on ImageNet as a baseline for fine-tuning on medical task. The custom input

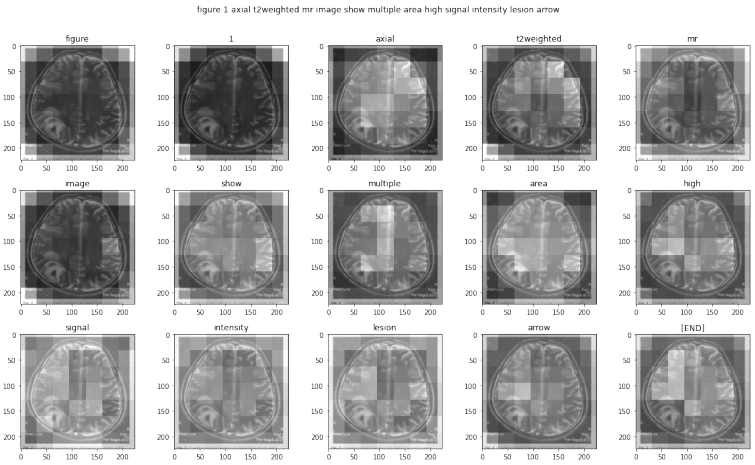

Figure 5: An example of an attention map result from the Show, Attend, and Tell model. The predicted caption is shown on top.

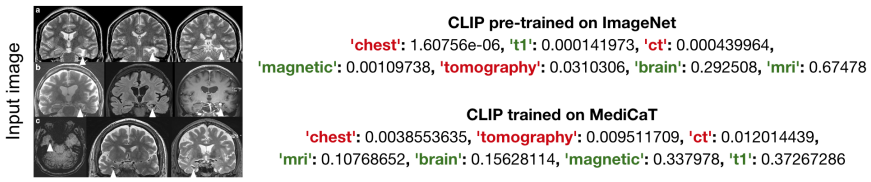

Figure 6: Comparison of probabilities for image-caption matching before fine-tuning on our data, and after. Captions are sorted in the increasing order of probabilities. Green color signifies that the caption relates to the image, while red text does not correspond to the visual input.

and output layers were to be developed according to our data. Specifically, 1 output neuron and a sigmoid activation function corresponding to the 2 classes in the data.
The model was then trained for 20 epochs with an image resolution size of 180 X 180. We tested various learning rates, settling on the rate 1e-5. The training set has 3,951 samples, and the validation set has 987 samples. The validation accuracy is 97%and the training accuracy is 99%, noting slight overfit. The precision 0.9816 , recall 0.9667 and f1 score 0.9732 resulting from the validation set show promising initial results. The test set containing 100 samples results in: precision 0.9655 , recall 0.9000 and f1 score 0.9241 and accuracy of 96%. Refer to the supplementary for training curves.

5.3 Results of CLIP

The pre-trained weights available for CLIP are received by training on large natural scenery datasets, so we fine-tune the pre-trained model by training on a subset of CT and MRI data from MedICaT. We utilize an open-source version of CLIP for training (4). Our data comprises of 5,000 samples: 3,500 pairs of images and text for training, and 1,005 for validation. We apply ResNet50 as an encoder, set learning rate to 1e-3, and batch size of 16. We observed a stable loss decrease for the first 15 epochs, but it started to fluctuate a few epochs after the warm-up. CLIP evaluates performance by looking at R-precision at 1, 5, and 10 top queries. On our validation set we received very low metrics of 'image to text' R@10: 0.0677, and 'text to image' R@10: 0.0707. We believe that there are three major reasons for a low performance. Firstly, we used only a subset of data when CLIP originally is trained on millions of samples varying up to 400M images. Secondly, our captions have low word variability. Lastly, a high semantic dependence between the words within each group may cause the model to learn to distinguish two subsets: CT and MRI, instead of focusing on the general understanding of the scene. Ultimately, the training of CLIP on MedICaT data would benefit from the ablation study on hyperparameters tuning which would require more resources.

From our qualitative evaluation, we noticed a good trend. Our trained model managed to diversify the captions that relate to MRI from CT by giving higher probabilities to MRI-related text than CT when testing on an MRI image. This is not observed on the ImageNet pre-trained CLIP, refer to Figure 6.

6 Model Robustness and Unraveling the Black Box

In terms of evaluating the robustness for the model that we created, we decided to add different forms of noise to observe if this changes the accuracy of the image captioning. The first way we added noise to our images was through the salt and pepper method. It is only found in grayscale images and randomly selects pixels in the image and colors them either white or black. The second explored method was using a Poisson distribution by attenuating the images we would input. In comparing the original image to the image that was subjected to one of these methods, the results did not vary much. As the results were extremely similar, we can say the model was only slightly affected by the noise. We illustrate and compare the difference in predictions before and after adding the noise in the Appendix Figure A7.

The Show, Attend and Tell method has an attention mechanism embedded in the architecture since it uses a transformer for text embeddings. The transformer has a self-attention mechanism that allows us to propagate the attention to a visual image embedding and analyze the obtained results. After the model produces a predicted caption we are able to view the corresponding attention maps on top of the image and see which regions contributed to the prediction of the word. This step is essential to testing model interpretability. It also helps finding classes and words for which the method works poorly, so we can adjust our approach based on these observations. We show this in a Figure 5 on medical data, and in the Appendix in Figure A3 on Flickr8k example.

For CLIP model we could also look at the probabilities for each predicted word. These probabilities show the model's confidence and serve as an additional explainability technique, ref. to Figure 6.

7 Discussion and Conclusion

In this work, we explored a few approaches to image captioning for medical tasks, namely Show, Attend and Tell, and CLIP. We tested them on multiple datasets and subsets of medical data including computed tomography scans, magnetic resonance imaging, and X-Rays. We also showed that for a specific caption generation task where the model is predicting a particular characteristic, such as Pneumothorax presence, a classifier can benefit the task by predicting the most valuable information - disease diagnosis. Thus, we have trained a fully-supervised classifier on X-Ray images from IU-Chest X-Rays and SIIM-ACR Pneumothorax datasets.

We achieved meaningful results when training image captioning models on natural scene images such as Flickr8k data and showed the results with appropriate attention maps corresponding to the generated captions. While training the Show, Attend, and Tell, and CLIP models we found that they had a tendency to overfit on subsets of the MedICaT dataset. We observed improvements in results when we included more diverse examples from the data by training the model on data from all four clusters so that the model now sees as many different tokens as possible. For the CLIP model, while our results show promise, the next steps include hyperparameter tuning, which requires more computational resources. Additionally, there are several other restricted-access medical datasets to explore for image captioning that can help with model generalizability.

8 Members Contributions

The team discussed the methodology, experimental designs and results together. As a team we discussed challenges and obstacles to then explore methods for resolving the challenges. We split the data exploration according to - MedICaT captions, MedICaT image, IU-Chest X-Rays captions, and IU-Chest X-Rays image exploration. We split the work by methods: Ananya was working on 'Show, attend, and tell', Maria - 'CLIP', Rhia ran the 'ResNet50 classification,' and Patrick was working on model robustness. We wrote the report together, with each member leading the writing of their respective sections in the report.

References

[1] CT scan, https://www.mayoclinic.org/tests-procedures/ct-scan/about/pac-20393675

[2] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385

[3] Howard, A., Sandler, M., Chu, G., Chen, L., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q.V., Adam, H.: Searching for MobileNetV3. CoRR **abs/1905.02244** (2019), http://arxiv.org/abs/1905.02244

[4] Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773, if you use this software, please cite it as below.

[5] MRI, https://www.mayoclinic.org/tests-procedures/mri/about/pac-20384768

[6] OpenI: Indiana university - chest x-rays (png images) https://openi.nlm.nih.gov/faq.php

[7] Radford, A., et al.: Learning transferable visual models from natural language supervision (2021). https://doi.org/10.48550/ARXIV.2103.00020

[8] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 539–559 (2023). https://doi.org/10.1109/TPAMI.2022.3148210

[9] Subramanian, S., Wang, L.L., Mehta, S., Bogin, B., van Zuylen, M., Parasa, S., Singh, S., Gardner, M., Hajishirzi, H.: Medicat: A dataset of medical images, captions, and textual references. CoRR **abs/2010.06000** (2020), https://arxiv.org/abs/2010.06000

[10] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)

[11] Zawacki, A., Wu, C., et al.: SIIM-ACR pneumothorax segmentation (2019), https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation

A Appendix

A.1 CLIP Model Architecture

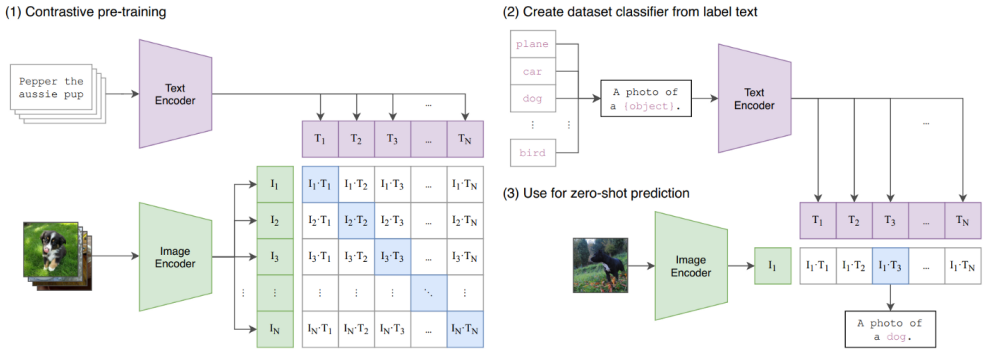

Figure A1: CLIP architecture, source: original paper (7)

A.2 Show, Attend, and Tell Model Architecture

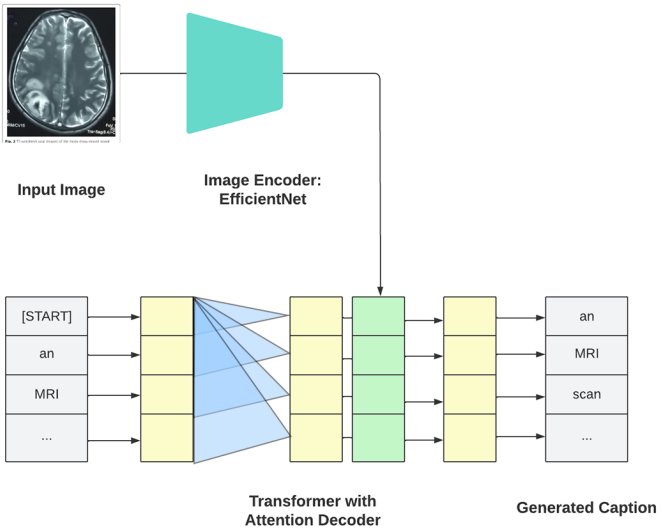

Figure A2: Show, Attend, and Tell architecture, modified from: Tensorflow tutorial (10)

A.3 Show, Attend and Tell Supplementary Results

Predicted captions: from the tutorial- "A man in a red shirt is surfing on a wave", random royalty-free image-"A man riding a bike on a dirt bike" as shown in the attention maps in Figures A3 and A4. The white parts of the image shown are where the model adds attention. While there are improvements to be made, the model focuses correctly on the wave and bike, respectively.

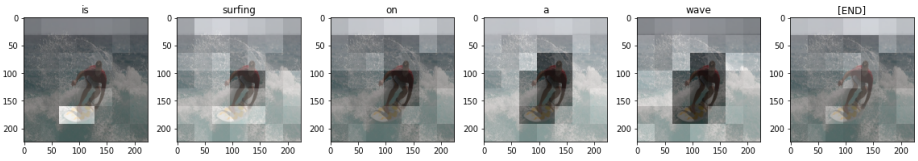

Figure A3: Generated caption for example from Flickr8k: "A man in a red shirt is surfing on a wave"

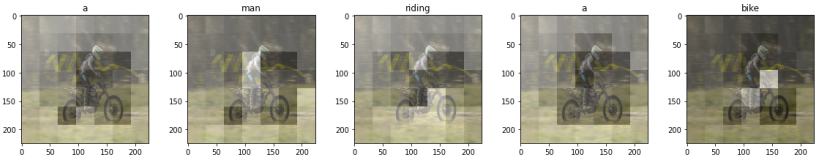

Figure A4: Partial attention map for generated caption for a random royalty-free image

A.4 Show, Attend and Tell Overfitting

When training on 5000 samples of CT images for 100 epochs with early stopping, we found the model overfitting dramatically as seen below. Specifically, the model over-predicted commonly occurring words such as “CT” and “Computed Tomography”, training curves are shown in the Figure A6.

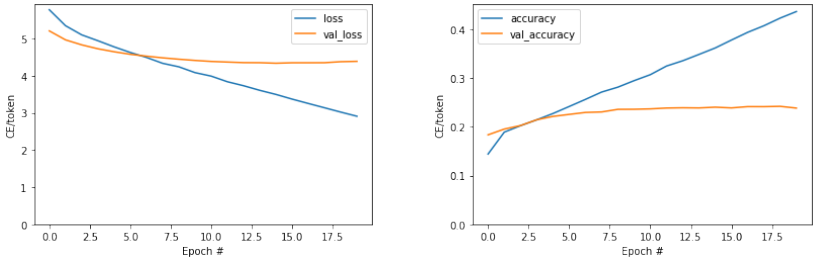

Figure A5: CT subset of MedICaT. Training curves: masked Loss and accuracy [major overfitting]

A.5 Classification on SIIM-ACR Pneumothorax Training Curves

After training on both SIIM-ACR Pneumothorax and IU samples for 20 epochs at a learning rate of 1e-5, the model steadily improves in accuracy as the loss stabilizes.

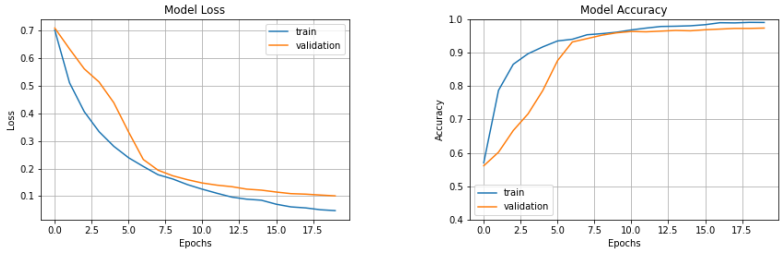

Figure A6: SIIM-ACR Pneumothorax Training Curves

A.6 Predictions of Show, Attend, and Tell on Noisy Images

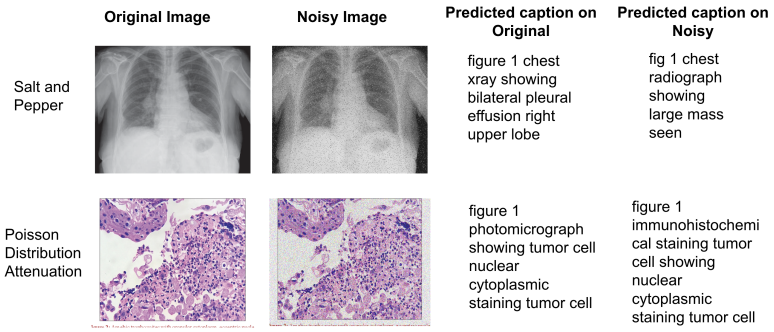

Figure A7: Comparison of Show, Attend, and Tell prediction before and after the noise perturbations