Intro page

This page is notes & should be removed in the final copy

* Introduction - Clearly state the problem you wanted to solve.
  + Be persuasive, and convince your readers that this problem is important.
* Related Work - List 2 or 3 related projects and how they dealt with the same or a similar problem
  + It is even better to choose peer-reviewed articles from journals or conferences related to your topic.
* Methodology - Outline the methods you took to deal with the problem you laid out in the introduction section.
  + This section must include:
    - At least one figure that illustrates your workflow.
    - Your figure may detail how data was collected and processed or how your methods were implemented and applied.
* Result - Outline your major results.
  + Tell your readers about the impact or usefulness of what you discovered.
* Conclusion - Summarize your project.
  + Then, point out the future directions or things can be improved or expanded upon in the future.

Introduction

Eye-related diseases globally account for millions of affected patients and their treatment exceeds. $400 billion in the United States alone [4]. Many of these diseases are either preventable if detected early and can be detected through relatively inexpensive methods such as eye screenings [5]. While these diseases are easy to detect, and most are simple to treat and even prevent if detected early, the amount of well-trained ophthalmologists is experiencing shortages and the cost for these screenings is become too expensive for most people to consider [6]. In comparison, a computer trained eye screening model can be created at a relatively inexpensive price. This research paper builds on the learned knowledge from Burrola [7], where pre-trained CNNS were used to classify fundus images.

Vision transformers (ViT) have garnered an increase in popularity since the success of Dosovitsky et al. [1]. Traditional methods of image classification have relied mostly on convolutional neural networks (CNNs) as can be seen in most research papers regarding image classification. CNNs can perform optimally even on limited datasets, even with the need to use pre-trained models. These models however lack a stable way of decrypting what the model is thinking and how its classifying the images. Although that is not the focus of this research, it is a vast improvement that will be leveraged and used in the future of medical image research with artificial intelligence (AI).

A massive improvement of ViT over CNNs is they include an attention mechanism just like they do in Natural Language Processing (NLP). This attention mechanism allows the model to capture relationships in the images in short and long-range. Since the image is treated as tokens, self-attention can capture relationships between tokens which can be used for visualizing where the model is giving focus. This simple mechanism can allow a trained professional to look at an image and quickly focus their attention on aspects of the image that can lead to a diagnosis and even allow professional to notice something that they may otherwise miss.

Our goal is to create a model that can be used across medical facilities to quickly give a pre-diagnosis for patients in a matter of seconds at an inexpensive price. Early detection of diseases should be possible to do even at home, using the camera on a phone and a well-trained model. This goal is in progress and this research will not achieve that goal, though it will lead one step closer to making early detection from home a reality. For this paper, the goal will be to train a model on images of fundus known to have Glaucoma, Diabetic Retinopathy, or Cataracts and improve on previous accuracies done by CNNs [3][6].

Related Work:

* 1 article that I used before in CSML
* The github repo for the 89% accuracy using CNN
  + <https://github.com/bsdr18/Image-Classification-on-Eye-Disease-Dataset>
* Article on how transformers outperform CNN

* + <https://arxiv.org/pdf/2108.09038>
  + https://arxiv.org/pdf/2010.11929

The use of Vision transformers has been used some info about how its used in a bunch of other thigns here ViTs have been shown to work surprisingly well in comparison to CNNs [1] and have not been fully explored in the medical field [2]. ViTs require significantly more data than CNNs do to yield good results as compared to CNNs. CNNs can train on smaller data sets, but it is common for them to use transfer learning as a starting point. By that same methodology ViTs can take advantage of transfer learning and show that their improve their performance by doing so [2].

Work done by Sai Divya Battalapalli showed a CNN model trained on 7 epochs could yield results of approximately 80% accuracy [3]. This is the comparison benchmark this paper is attempting to improve upon by utilizing the same dataset, epoch count, and evaluation methods, where applicable.

Methodology

Due to ViTs requiring large amounts of data to be trained on, we will be using a process called transfer learning. Transfer learning is the process of taking the weights of an existing/pre-trained model and then training/fine-tuning from that starting point on a new set of data for faster learning and improved performance. This technique allows smaller datasets to be useful where normally a ViT would only become useful and efficient with datasets like Google’s JFT-300M, which contains over 300 million images [2]. Medical pre-trained models are extremely rare to find available publicly due to all the regulations around patient information, therefore we opted to use a generic image model. We chose the “vit\_base\_patch16\_224” PyTorch Image Model (timm). This model was trained on the ImageNet-21k dataset using approximately 14 million images of size 224x224 and used 16x16 size patches during its training.

We will be using the same retinal image dataset Battalapalli used when training their CNN model [3]. The dataset is called “Eye\_diseases\_classification” . it contains 2-dimensional colored retinal images taken from various datasets including IDRiD, Oculur recognition, HRF [8]. The is well-balanced with approximately 1,000 images of the following classes: “Normal”, “Cataracts”, “Glaucoma”, and “Diabetic Retinopathy”. The dataset on its own, though relatively small, is expected to be large enough for the ViT model to get performance equal to or greater than Battalapalli’s 80% accuracy. If need be, data augmentation is considered to increase the amount of training data.

Data

* + Dataset Used
    - <https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification?select=dataset>
    - Eye\_diseases\_classification
    - Contains (~1000 images of each):
      * Normal
      * Diabetic Retinopathy
      * Cataract
      * Glaucoma
  + Resizing + reason for resizing
    - The images were resized to (224x224) to accommodate the model being used.
  + Augmentation
  + Tensor conversion
  + RGB vs BW
    - RGB was used over black
* Training
  + Transformer architecture
    - ViT
  + Pre-trained model
    - Vit\_base\_patch16\_224
      * Patch 16
      * Sizes 224
  + Fine-tuning
    - Freezing layers
      * All = 48% accuracy
      * None = ??%
      * 50% of layers = ??%
* Testing
  + Base pretrained model results
  + Fine-tuned results

Results

Conclusion

* Use of transformers in medical image classification
* Speed
* Future:
  + Compare directly to a CNN using transfer learning
  + Larger datasets
  + Saliency maps and other tools to see where focus is given

References

1. Dosovitskiy: <https://arxiv.org/pdf/2010.11929>
2. Matsoukas: <https://arxiv.org/pdf/2108.09038>
3. Github compare repo
4. Deangelis 2011
5. Forouhi 2018
6. Cherlaramani 2021
7. Burrola 2024 (ODC CNN…)
8. Dataset used
9. Huggingface timm model: <https://huggingface.co/timm>