Capstone Project Report

Executive Summary

**I want to apply and test the ability of state-of-art techniques connecting text with image, zero-shot classification models, to predict or annotate unlabeled medical images, MedMNIST dataset which is a benchmark dataset used to test new algorithms or new methods. The general ImageNet zero-shot models can annotate the high-level labels, like chest or breast, with a high accuracy on this medical benchmark dataset. However, it performs poorly on granular-level labels like malignant, normal, benign and so on. We then train some traditional CNN models on one of dataset by using labels annotated by zero-shot model and fine-tune it to have the best accuracy on its test dataset.**

Introduction/background

**Because there are millions of unlabeled medical images generated in hospitals waiting for annotations before they can be downstream analyzed, but manually labeling them is too time consuming. Also, sometimes, the labels or classes that we want to predict are not in the all pretrained models or some of them are not. At this time, we come across zero-shot, one-shot or few-shot transfer learning.**

Data

**MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. This benchmark dataset was collected and processed by 8 researchers from six institutions including universities and hospitals. They have published a paper about this dataset:** [**https://arxiv.org/abs/2110.14795**](https://arxiv.org/abs/2110.14795) **on the arxiv and they have a GitHub repository containing the guides and instruction of how to use this dataset by installing the package they wrote for this dataset, called medmnist. You can download each 2D and 3D respectively by clicking this website:** [**https://zenodo.org/record/5208230**](https://zenodo.org/record/5208230) **and also download all dataset automatically by setting download parameter in the package equal to True. The whole dataset is very easy to access. The researchers now are still actively supporting and developing this dataset and its GitHub Repo. Although dataset is not proposed for clinical use, it has many advantages of research purposes, such as diverse, standardize, lightweight, educational. The diverse dataset provides users with different organ images, different tasks, like binary/multi-class, regression problem.**

Methodology

**My goal is to automatically annotate the millions of unlabeled medical images without too much human involvement or manually labelling. I will use zero-shot clip model and convolutional neural network together to tackle this problem. The whole process involves two procedures. I first split data which we want to annotate into two parts. The first part of data will be annotated by zero-shot clip model choosing the highest probability label among all. Then I will use convolutional neural network to train the first part of data with their predictions from zero-shot clip model as their CNN ground truth labels. Eventually, I will evaluate this trained CNN model on the second part of the data with their truly ground labels. Once I have this best tuned CNN model, I am able to leverage this CNN to annotate new unlabeled medical images. In this way, I can show that using two method together can achieve a higher accuracy on annotation than only using zero-shot clip model.**

Zero-shot clip model

**I will first introduce the state-of-the-art model which connects text and images, clip model. The clip model efficiently learns visual concepts from natural language supervision. Clip model can be applied to any visual classification task by simply giving the visual labels to be recognized. The clip model is trained on a very wide range of natural language supervisions and images on the internet. The general clip model performs better than ImageNet Resnet101 on ImageNet, ImageNet v2, ImageNet Rendition, ObjectNet, ImageNet Sketch, ImageNet Adversarial.**

**Clip is the short name for Contrastive Language Image Pretraining. As the name suggests, Clip model is trained to pair the natural language concepts with the image concepts. Clip models want to realize the prediction and pair the image concepts with unseen labels. This is very impressive and inspirational words, because before this idea and clip model, in computer vision, people always train and predict labels on observed/seen labels. However, as demand increases, there are many unseen/unobserved categories needed to be recognized in images.**

**The overall architecture of clip models is below:**

**Diagram

Description automatically generated**

**The model contains three phases. The first phase is to train two encoders, the first is text encoder which extracts and learns the text representations in T1, T2, …, TN; the second is image encoder that learns visual representations in I1, I2, …, IN. Then the dot product or other similarity score between paired text representations and visual representations is maximized while that of non-paired text representation and visual representations is minimized. After this optimization, captions that warp the unseen labels are fed into text encoder to produce each text representation for each class, and the unrecognized image is fed into image encoder to produce its visual representation. Finally, computing cosine or other similarity score between this visual representation and each class text representation gives the highest score, and its corresponding class will be chosen as the prediction label for the new image. The above is overall process for training and prediction.**

**The trained clip model perform well on some common dataset, like 90.1% accuracy on FOOD101 data, 90.2% accuracy on SUN397, 89.0% accuracy on YOUTUBE-BB.**

**There are some advantages of this kind of zero-shot model, the first is CLIP is highly efficient.**

**Chart, line chart

Description automatically generated**

**The second one is CLIP is flexible and general, because the CLIP is trained on a wide range of visual and text dataset.**

Chart, line chart

Description automatically generated

Convolutional Neural Network

**The next architecture I apply is the convolutional neural network. This is very famous artificial neural network in image analysis like image classification, image segmentation, natural language processing. Recent year, there are too many research and applications in CNN because it is very powerful and useful in image analysis, including video analysis, signal analysis.**

**As the below figure suggests, the input digit image 2 is the input layer: (28\*28\*1). In this CNN architecture, it contains two convolutional layers, two max-pooling layers, two fully connected layers. The first convolutional layer has n1 filter/kernels, each of which is 5 by 5. Each filter is sliding over the width and height of the input layer and compute the dot product between the selected input layer entries and this filter map, once this filter slide over all possible width and height, it will produce a 2-dimensional activation map. If the first convolutional has n1 filters, then before the next layer, there are n1 activation maps. Each filter map will learn the different visual pattern of the input layer, such as edges of the orientation, the colors, and so on. In the below example, the input layer only has one channel (1-dimensional depth), however, in reality, most images have three channels (3-dimensional depth), in this case, the filter map should also be 3-dimensional depth to make each activation map only 2-dimensional space. In other words, the depth of filter map is always equal to the depth of the input layer. The most important property of CNN is the local connectivity and shared weights, because the image dimension is huge, making the fully connectivity impractical. The same filter will slide over all width and height of the input layer so weights from this filter map will be used to compute dot product with receptive field of the input layer for many times. The fully connectivity and shared weights can significantly reduce the computational cost and learn the local and different visual patterns. The filter map size usually is small, 3 by 3 and 5 by 5 are the most common one. The next most common layer is the Max-Pooling, here is 2 by 2 max-pooling map with 2 strides. Max-pooling layers can dramatically reduce the spatial size of the input layer so that the number of parameters is reduced as well to make computations efficient. The bonus advantage of it is to avoid the overfitting because overfitting often happens in the deep neural network due to too many parameters to fit the noise. By using 2 by 2 max-pooling layer with 2 strides, the 24 by 24 activation map is reduced to 12 by 12. At the end of this architecture, there are two fully connected layers companied with a non-linear ReLU activation layer. There is also dropout layer in the last one. The dropout layer is to mask out some neurons in this layer by a probability like 0.5 to overcome the common issue, overfitting in deep neural network. The final layer contains 10 neurons, each of which presents a score/probability for each class.**

Diagram

Description automatically generated

**The below picture is the typical local connectivity, shared weights in one convolutional layer. The input has three channels and one padding, there are two filters, each of which is 3 by 3 by 3, so the output volume has two activation maps.**

Diagram

Description automatically generated

**In practice, people don’t need to design a CNN architecture from scratch, they first should look for the current state-of-the-art pretrained models, like AlexNet, VGGNet, ResNet, and use these pretrained models to fine-tune on their own dataset. The best CNN model on ImageNet usually works better than the people’s own customized model.**

My workflow of combinational two above models

**Our problem is that we want to automatically annotate the medical images without manually labelling. So, I will apply one of best zero-shot clip models to predict the unseen image labels. However, these prediction labels only have 80% accuracy. Thus, the second stage is to train a custom CNN model to improve this prediction accuracy to 95% accuracy.**

**There are currently 8 pretrained zero-shot models:**

import clip

clip.available\_models()

['RN50',

'RN101',

'RN50x4',

'RN50x16',

'RN50x64',

'ViT-B/32',

'ViT-B/16',

'ViT-L/14']

**After some experiments and comparisons, the Vision Transformers-L/14 (ViT-L/14) performed best among others. ViT-L/14 was chosen to be used. It was pretrained on ImageNet containing around 14 million images with 21,843 classes, then fine-tuned on another ImageNet at resolution 224 by 224. Next, we can load this huge pretrained vision transformer model to have a look at its overall sketch.**

model, preprocess = clip.load("ViT-L/14")

model.cuda().eval()

input\_resolution = model.visual.input\_resolution

context\_length = model.context\_length

vocab\_size = model.vocab\_size

print("Model parameters:", f"{np.sum([int(np.prod(p.shape)) for p in model.parameters()]):,}")

print("Input resolution:", input\_resolution)

print("Context length:", context\_length)

print("Vocab size:", vocab\_size)

Model parameters: 427,616,513

Input resolution: 224

Context length: 77

Vocab size: 49408

**From the above code and results, we can observe that there are 427 million model parameters, the input resolution is 224 consistent with the pretrain ImageNet resolutions, the context length 77 refers to the maximum text size fed into the model. Training over 427 million model parameters requires a lot of computational resources and time, luckily, we don’t need to train this kind of huge model, all we need is just apply this model on medical images in the following.**

**I will use two benchmark datasets as illustration to my method. The first one is Chest MNIST which contains 78,468 training samples, 11,219 validation samples, 22,433 test samples. The second one is Breast MNIST which contains 546 training samples, 78 validation samples, 156 test samples. I chose all 546 breast training samples and randomly chose 546 chest training samples to form a mixture of two benchmark datasets. The following picture is 9 random images with its true labels selected from this merged dataset.**

**Graphical user interface, application

Description automatically generated**

**Next, I will create two captions encompassing these two labels, chest and breast. “This is a photo of chest, and it’s a type of organ” and “This is a photo of breast, and it’s a type of organ.” For CLIP model, it’s typically to write a short caption or sentence capturing the labels then feed these captions into text encoder model like below code:**

uni\_labels = ['breast', 'chest']

text\_descriptions = [f"This is a photo of {label}, and it's a type of organ." for label in uni\_labels]

text\_tokens = clip.tokenize(text\_descriptions).cuda()

with torch.no\_grad():

text\_features = model.encode\_text(text\_tokens).float()

text\_features /= text\_features.norm(dim=-1, keepdim=True)

Results

Conclusion/Next Steps

**There are some pretrained medical image deep learning models, like in the monai, medical open network for artificial Integillence which provides domain-optimized foundational capabilities for developing healthcare imaging training workflows in native PyTorch language. In the future, I may plan to embed them into current zero-shot, N-shot transfer learning architecture then apply them into MedMNIST image dataset because current N-shot learning is trained on ImgaeNet which is far off medical fields.**

**Once the open AI releases their codes which are used to trained the their zero-shot classification model, I plan to use those codes to train on all medical dataset, like all MedMNIST benchmark dataset and fine-tune it. Then I can have a medical domain adaptation zero-shot models. Then I test and apply them to some other medical dataset to annotate them and evaluate the results. In this case, the results should be much better than the general ImageNet -based zero-shot’s performance.**