Intelligent grading of cortical cataract

ZJU-UIUC INSTITUTE
浙江大学伊利诺伊大学
厄巴纳香槟校区联合学院

Author: Liheng Jing, Yuyang Wei
Advisor：Phd.Hongjin Wang, Prof.Jian Wu, Prof.Haochao Ying

Abstract

This paper demonstrates the potential of Vision Transformer (ViT model) to achieve multi-modal fusion.

Approach:

- Firstly, Image features and text features (sub-label features) are extracted based on residual network (ResNet) and deep-learning neural network (DNN), respectively.
- Secondly, The multi-modal fusion of image and text features is realized by using ViT model framework.
- Lastly, The trained high-precision fusion model is used to guide the learning of residual network through knowledge distillation to achieve no sublabel input and improve the accuracy of image unimodal.

Result:

Our high-precision fusion model achieves 93.02% accuracy on the cortical cataract dataset, and the residual network of knowledge distillation achieves 79.27% accuracy without sublabel input, which is nearly 5% higher than that of residual network (ResNet-18) alone.

Introduction

Problem:

The clinical diagnosis of cortical cataract has the problems of low efficiency and high error rate due to the complexity and variety of lesions. There is an urgent need for computer-assisted diagnostic methods to achieve intelligent classification of cortical cataract.

Current situation:

ResNet is less effective in processing cortical cataract images since the pathological features are not obvious and only concentrated in the small pupil part.

Solution:

Single mode information is not enough and it is necessary to introduce sub-label information for multi-modal fusion. Our innovation lies in replacing the original patches in ViT model with feature tokens extracted from images and text, through which achieves multi-modal fusion.

Approach

Extracting data features:

we used a series of methods of image enhancement on the image and extract the image features by ResNet-18. Similarly, we set up a simple deep-learning neutral network (DNN) independently to extract the text features. (Fig. 1)

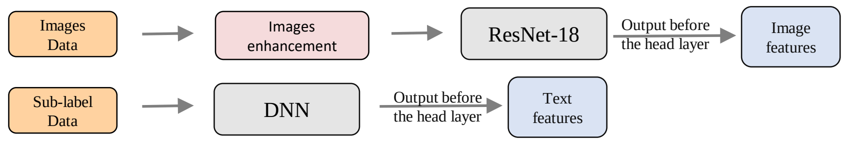

Fig. 1 Data-feature extractor frame work

Multimodal fusion:

Differing with ViT processing images, which divides an image into regular non-overlapping patches, our work is to fuse different features. So, our inputs are only two patches (i.e. vectors), image features and text features, rather than the origonal image and text data.

To satisfy the requirements of dimension, we use a linear layer to take place of a Conv2d in ViT PatchEmbed. The other parts are similar to the encoder of ViT.

A token synthesize the information and connection between the image features and text features via a series of Transformer blocks. Then the last MLP can predicate the label based on the information of the token. (Fig. 2)

Vision Transformer Fusion (ViT Fusion)

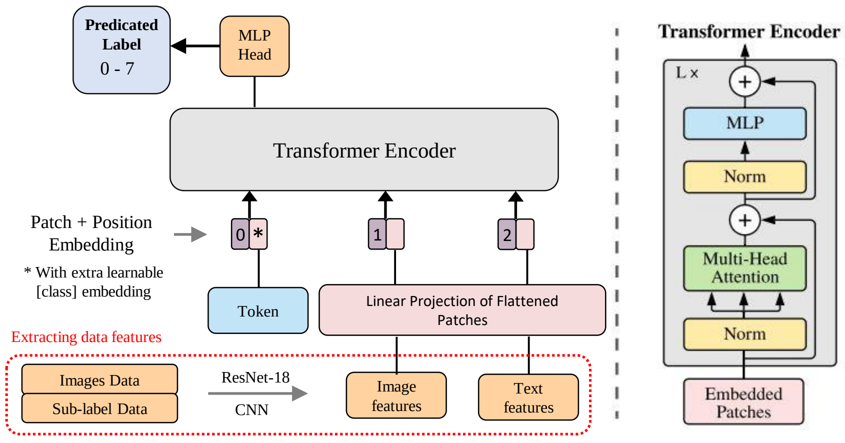

Fig. 2 ViT fusion frame work

Knowledge distillation

Combining the feature extractor model and fusion model, we get our high-accuracy teacher model. Here, we use the pre-trained ResNet18 as the student model. While giving one image data to the student, we give the corresonding image and sublabel to the teacher. Then we use the teacher's output as the soft-lablel to guide the student to learn. Comparing with only using the hard label to guide the student, with both the soft-label and hard-label guidance, the stident will get better performance.

Knowledge Distillation (KD)

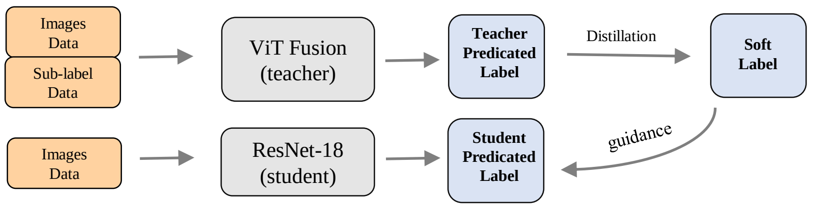

Fig. 3 Knowledge distillation frame work

Conclusion

In this study, we use ViT model to achieve multi-modal fusion and it turns out that it has a fairly good performance on our dataset. The detailed result comparison is in Table 1. Compared with 70%+ by ResNet-18 alone, with image enhancement, the test accuracy can be up to 73%+ by ResNet-18 alone. Our trained high-precision teacher model achieved an accuracy of 93% on the dataset. With teacher network's guidance, the student ResNet-18 improves the accuracy to 79%+, which is about 5% higher than that of ResNet-18 alone.

Model	Data	Pre-train data	Test Accuracy
ResNet-18	Images without enhancement	ImageNet	**70.45%**
ResNet-18	Images with enhancement	ImageNet	**73.95%**
DNN	Sub-labels	no pre-trained	84.65%
ViT fusion	Images and Sub-labels	no pre-trained	93.02%
Distilled ResNet-18	Images	ImageNet	**79.07%**

Table 1. Experimental results