Zero-Shot and Few-Shot Learning for Lung Cancer Multi-Label Classification using Vision Transformer

Fu-Ming Guo∗ Yingfang Fan†

Abstract

Lung cancer is the cancer leading cause of cancer-related death worldwide. Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are the most common histologic subtypes of NSCLC. Histology is an essential tool for lung cancer diagnosis. Pathologists make classifications according to the dominant subtypes. Although morphology remains the standard for diagnosis, significant tool needs to be developed to elucidate the diagnosis. In our study, we utilize the pre-trained Vision Transformer (ViT) model to classify multiple label lung cancer on histologic slices (from dataset LC25000), in both Zero-Shot and Few-Shot manners. Then we compare the performance of Zero-Shot and Few-Shot ViT on accuracy, precision, recall, sensitivity and specificity. Our study show that the pre-trained ViT model has a good performance in Zero-Shot setting, a competitive accuracy (99.87%) in Few-Shot setting (epoch = 1) and an optimal result (100.00%) in Few-Shot setting (epoch = 5).

1 Introduction

Lung cancer is the cancer leading cause of cancer-related death worldwide. It is not only because of smoking, but also exposure to toxic chemicals. Non-small-cell lung cancer (NSCLC) is any malignant epithelial lung tumor that lacks a small-cell component [1]. NSCLC represents approximately 85% of all new lung cancer diagnoses [2]. Lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are the most common histologic subtypes [3] of NSCLC. These subtypes are further subclassified into multiple subtypes according to WHO criteria [4]. Histology is an essential factor for individualizing treatment based on either safety or efficacy outcomes [5]. Adenocarcinomas are malignant epithelial tumors with glandular differentiation. It has clear morphologic patterns such as acinar, papillary, lepidic, micropapillary [4], although mix pattern adenocarcinomas are most common [6]. Squamous cell carcinomas are often centrally located and derived from bronchial epithelial cells. Unequivocal keratinization and well-formed classical bridges can be diagnosed as squamous cell carcinomas [4].

Histologic distinctions may be unclear due to poorly differentiated tumors and requires confirmatory immunohistochemical stains. The heterogeneous histology within the same lesion occurs in many NSCLC tumors. Pathologists make classifications according to the dominant subtypes. Although morphology remains the standard for diagnosis, significant tool needs to be developed to elucidate the diagnosis.

Using AI to analyze tissue sections is typically called computational pathology (Fuchs and Buhmann, 2011). Research in this area can trace back to the middle of the last century, with the seminal application of image analysis algorithms to medical images. Image analysis algorithms can classify cell images based on quantitative cell characteristics, e.g., size, shape, and chromatin distribution, and support the diagnosis of diseases (Mendelsohn et al., 1965). The early applications implemented

∗fmguo@acm.org, Association for Computing Machinery.
†Harvard Medical School.
computational features matched to a biological process, later replaced by radionics using generic features of texture descriptors (Zwanenburg et al., 2020).

Automated classification of abnormal lesions using images is a challenging task owing to the fine-grained variability in the appearance of abnormal lesions. Deep convolutional neural networks (CNN) (LeCun et al., 1999) show potential for general and highly variable tasks across many fine-grained object categories. In the recent booming (Dean, 2022) of Deep Learning (LeCun et al., 2015), a series of CNN-based models unceasingly refreshed the state-of-the-art performance on various computer vision benchmarks (Deng et al., 2009; Krizhevsky and Hinton, 2009). Nowadays, the Transformer (Vaswani et al., 2017) shows an advantage over computer vision tasks (Dosovitskiy et al., 2020b) after becoming the de-facto standard for natural language processing tasks (Devlin et al., 2019; Brown et al., 2020).

![Diagram of Vision Transformer (ViT)](image)

Figure 1: Model Overview - Vision Transformer for Lung Cancer Classification

The main advantage of the Deep Learning approach is automatically learning features from the data, instead of crafting meaningful features in feature engineering and conventional image analysis. The automated feature learning from the Deep Learning approach reduced the required domain knowledge and the implementation time. More importantly, the automated Deep Learning approach yields robust, hierarchical feature representations, which outperform traditional image analysis methods in most cases.

Despite the powerful learning ability of the Deep Learning approach, there is still the central issue in the application in the medical domain, the domain shift. How do these state-of-the-art computer vision models perform on medical image analysis tasks? Specifically, the histological analysis of lung cancer in our case. The Zero-Shot Learning and Few-Shot Learning are the emerging paradigms to address this issue. Few-shot transfer learning, where a model is first pre-trained on a data-rich task before being fine-tuned on a wide range of downstream tasks, has emerged as a powerful technique in natural languages and computer vision tasks. Pre-trained representations have shown substantial performance improvements using self-supervised learning and transfer learning. There are also emerging zero-shot learning techniques showing outstanding performance, like prompt-tuning and instruction tuning. In our study, we utilize the pre-trained Vision Transformer (ViT) model to classify multiple label lung cancer on histologic slices (from dataset LC25000), in both Zero-Shot and Few-Shot manners. Then we compare the performance of Zero-Shot and Few-Shot ViT on accuracy, precision, recall, sensitivity and specificity. Our study show that the pre-trained ViT model has a good performance in Zero-Shot setting, a competitive accuracy in Few-Shot setting (epoch = 1) and an optimal result in Few-Shot setting (epoch = 5).

To illustrate the working mechanism behind these models, we utilized the cutting-edge model interpretation work Transformer models (Chefer et al., 2021)) compared with the analysis from the human expert. Our pathology expert provides the information about which regions help her decide whether the input image is a lung cancer lesion or not from the professional perspective. The
comparison between the pathology cognition and Grad-CAM visualization presents a high degree of consistency. This comparison helps us understand our deep learning model, and rationalizes our proposed method further.

2 Methods

The pre-train-finetune paradigm is the robust defacto pipeline in visual and language learning. In the finetune stage, a Multi-Layer Perceptron (MLP) (Rosenblatt, 1958) (Aizerman, 1964) often follows a thoroughly pre-trained backbone network to function as a projector and aid transfer learning (Wang et al., 2021) on new tasks (medical image classification in our case). Our customized models utilized the paradigm above: ViT as the pre-trained backbone, followed by an MLP projector, and finally, a cross-entropy loss to classify and diagnose the cancer lesion (Figure 1).

First-level headings should be in 12-point type.

2.1 Model Architectures

We utilize the ViT (Dosovitskiy et al., 2020b) as our Transformer based backbone. The ViT backbone is pretrained on ImageNet, ImageNet-21k, and JFT-300M. (Deng et al., 2009; Sun et al., 2017) (Ridnik et al., 2021). ViT backbones holds accuracy of 88.05% on Imagenet, 90.72% on ImageNet-Real, 94.55% on CIFAR-100 and 77.63% on the VTAB suite of 19 tasks. In Figure 1, we illustrate how the ViT backbone function in our customize d model. The multiple scale input image samples (768 * 768 pixels) are first normalized to 224 * 224 pixels. This follows the input image format in ImageNet, due to ViT is pre-trained on ImageNet. For the 224 x 224 pixels input, the ViT backbone first split the input image into 196 image patches, and each patch is 16 x 16 pixels. Image patches are treated the same way as tokens (words) in natural language processing applications. The image patches are linearly embedded, with positional embeddings. ViT backbone (Dosovitskiy et al., 2020a) then feeds the resulting sequence vectors into a standard Transformer encoder (Vaswani et al., 2017). The process is visually illustrated in Figure 1.

3 Experiments

In this section, we describe the dataset we used and our experiments in the Zero-Shot and Few-Shot manners.

3.1 Dataset

We utilized the Lung Cancer part of the dataset LC25000. LC25000 has 25,000 color images in 5 classes. Each class contains 5,000 images of the histologic categories: colon adenocarcinoma, benign colonic tissue, lung adenocarcinoma, lung squamous cell carcinoma and benign lung tissue. All images are de-identified, HIPPA compliant and validated. We utilize the lung cancer categories: lung adenocarcinoma, lung squamous cell carcinoma and benign lung tissue. We split the 15,000 images in lung cancer categories into training set ($D_{train}$), validation set ($D_{validation}$) and test set ($D_{test}$), by the ratios 60%, 20%, 20%, after random sampling.

3.2 Zero-Shot transfer learning

First, we conducted the experiments in a Zero-Shot manner. The well pre-trained ViT model functions as frozen wights in this setting. We directly use the frozen pre-trained ViT model to make the prediction on test set ($D_{test}$). We report the accuracy on validation set and test set in Table 1.

3.3 Few-Shot transfer learning

Then we conducted the experiments in Few-Shot manners. We fine-tuned the pre-trained ViT model on training set ($D_{train}$) for 5 epochs, and validation the best model on validation set ($D_{validation}$) at each epoch. We find that the pre-trained ViT has prompt and strong learning ability, so that it quick achieves the optimal accuracy after fine-tuning of only 5 epochs (100.00% on both validation set ($D_{validation}$) and test set ($D_{test}$)). We report the accuracies on validation set and test set in Table 1.
Table 1: Zero-Shot and Few-Shot performance of ViT on LC 25000

| Part                      | Name       | Validation set (acc) | Test set (acc) |
|---------------------------|------------|----------------------|----------------|
| Zero-Shot                 |            | 33.77%               |                |
| Few-Shot (epoch = 1)      |            | 98.90%               | 98.87%         |
| Few-Shot (epoch = 2)      |            | 98.47%               | 99.50%         |
| Few-Shot (epoch = 3)      |            | 99.77%               | 99.70%         |
| Few-Shot (epoch = 4)      |            | 98.90%               | 99.87%         |
| Few-Shot (epoch = 5)      |            | 100.00%              | 100.00%        |

3.4 Receiver operating characteristic (ROC) curve

The Receiver operating characteristic (ROC) curve is used to evaluate the quality of a classifier (Green and Swets, 1966; Fawcett, 2006). The x axis of ROC curve is the ‘false positive rate’ (FPR), and the y axis means the ‘true positive rate’ (TPR). The point or line of a classifier locates more top-left on the ROC curve, the better this classifier is. Area under the Curve of ROC (AUC ROC), tests whether positives are ranked higher than negatives. We reported the AUC values of each model (of all the three classes, LUAD, BENIGN and LUSC) on Figure 3. The AUC values of ViT model (Few-Shot epoch = 5) is 1.00000000 for all the three classes (LUAD, BENIGN and LUSC), showing that the ViT model (Few-Shot epoch = 5) is an optimal classifier in our case.

![ROC CURVE](image2b)

Figure 2: Receiver operating characteristic (ROC) curves for Zero-Shot and Few-Shot ViT models

4 Discussion

A lot of work endeavors to make deep learning more sensible and explainable. In various deep learning applications especially into medical imaging, it is crucial to make the deep learning model more interpretative. Selvaraju et al. have introduced a Gradient Weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) technique which provides the interpretative view of deep learning models. Grad-CAM uses the gradients of any target concept, flowing into the final convolutional layer to produce the coarse localization map highlighting important regions in the image for predicting
the concept. In our application to classify the histology images, our visualizations (Figure 3) lend insights into failure modes of these models, showing that seemingly unreasonable predictions have reasonable explanations. Our visualization (Figure 3) is robust to adversarial perturbations, are more faithful to the underlying model and help achieve model generalizations by identifying dataset bias. The interpretation work of deep learning models is essential to the understanding of the mechanism behind the success of our model, and making the model more transparent. In our case, we utilize Grad-CAM to generate the visualized explanations via gradient-based localization. The localizations in Grad-CAM help explain which regions (confirmed salient features) in the image input contribute more to the model’s final decision and which regions are less (Figure 5). Furthermore, our pathology expert provides the information about which regions help him decide whether the input image is an infection or not from the professional perspective. The comparison between the pathology cognition and Grad-CAM visualization presents a high degree of consistency. This comparison is novel, helps us understand our deep learning model, and rationalizes our proposed method further.

Figure 3: Attention Visualization: upper side – Original histological image; lower side – Grad-CAM – localizes class-discriminative regions.

Acknowledgments and Disclosure of Funding

We thank the computing resource support from the TPU Research Cloud. Both Guo FM and Fan Y contributed to manuscript revision, read, and approved the submitted version.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), Advances in Neural Information Processing Systems 7, pp. 609–616. Cambridge, MA: MIT Press.
[2] Bower, J.M. & Beeman, D. (1995) The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural Simulation System. New York: TELOS/Springer–Verlag.
[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. Journal of Neuroscience 15(7):5249-5262.