Skin Cancer Classification : using CNN vs Transformer model

고려대학교 바이오의공학부 정은지

Abstract

이번 딥러닝 프로젝트에서는 지금까지 배운 여러 종류의 CNN 모델과 Vision Transformer 모델을 사용해 피부암 진단이라는 이미지 분류 task를 진행해보았다. Baseline CNN model, ResNet50, EfficientNet V2 B0, Vision Transformer(ViT) 모델을 구현했고 Accuracy, Loss, ROC curve, Inference rate, MCC 등의 성능 평가 지표를 사용하여 최종적으로 ViT 모델이 가장 뛰어난 성능을 보인다는 결론을 도출할 수 있었다.

Keywords: Image Classification, CNN, ViT

1. Introduction

1.1. Motivation

다양한 크기의 convolution layer를 사용해 이미지의 다양한 특징을 추출하는 방식을 사용하는 CNN 모델은 성공적으로 이미지 분류 문제를 해결할 수 있지만, context와 long-range dependency를 캡처하기 어렵다는 단점이 제기되기도 하였다.[1] Transformer model은 본래 자연어처리(NLP) 분야에서 사용되는 가장 대표적이고 성공적인 모델인데, 이를 이미지에 적용하는 Vision Transformer(ViT) 모델은 이러한 CNN의 문제를 해결하는 효과적인 대안이 될 수 있다는 의견이 제기되면서 이미지 분류 문제에서 각광받고 있다. 2021년에 발표된 "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE" [2] 논문의 저자는 CNN을 사용하지 않는 pure transformer 모델이 이미지 분류 task에서 매우 좋은 성능을 달성할 수 있다는 것을 주장하기도 했다.

이에 본 프로젝트는 실제로 이미지 분류 문제에서 Transformer 모델이 좋은 성능을 나타내는지에 대해 궁금증을 가지고, 이미지 분류 문제에서 SoTA를 달성한 여러 종류의 CNN 모델과 ViT를 피부암 진단이라는 이미지 분류 Task에 적용함으로써 ViT 모델의 성능을 관찰해보고자 한다.

1.2. Problem Definition

본 프로젝트에서는 의료데이터인 피부암 사진을 통해 피부암에 대해 양성인지 악성인지를 구분하는 task에 좋은 성능을 가지는 딥러닝 모델을 구현해보고자 한다. 이를 위해 이미 좋은 SoTA를 달성한 여러 CNN 모델(ImageNet으로 pretrain된 모델 포함)과 새롭게 각광 받고 있는 Vision Transformer 모델을 사용한 후 Loss, Accuracy, Confusion Matrix, Roc curve, MCC 등을 관찰하여 피부암 진단 task를 얼마나 성공적으로 수행하는지를 살펴보고 모델의 성능을 평가해볼 예정이다. 최종적으로는 이미지 분류 task에 사용되는 ViT 모델이 CNN 모델과 비교하여 비슷하거나 더 좋은 성능을 가질 수 있을지 탐구해보는 것을 본 프로젝트의 problem으로 정의한다.

1.3. Concise description of contribution

이미지 분류 task에 사용할 수 있는 CNN baseline 모델, ImageNet 데이터셋에 대한 pretrain된 ResNet model, EfficientNet model, Transformer(ViT) 모델을 구현하고 다양한 성능 지표를 이용해 성능을 비교 평가해볼 예정이다. 결론적으로 피부암 진단이라는 구체적인 이미지 분류 task의 수행에 있어서, convolution을 사용하지 않은 Vision Transformer 모델이 이미지 분류 측면에서 SoTA를 달성한 CNN 모델들 못지않게 좋은 성능을 가진다는 것을 결과로 확인해볼 수 있다는 점에서 본 프로젝트는 의의를 가진다. 또한 CNN 모델과 비교했을 때, Vision Transformer 모델이 가지는 장단점과 이미지 처리에 사용될 수 있는 Transformer 모델의 다양한 활용 방안에 대해 논의해보고자 한다.

2. Methods

2.1. Significance & novelty

NLP 분야에서 주로 사용되었던 Transformer 모델을 이미지 분류 분야에 새롭게 적용했을 때, 기존의 CNN 모델들과 비슷한 좋은 성능을 나타낸다는 것은 많은 엔지니어들에 의해 많이 증명되었다. 이러한 이론을 본 저자가 관심 있는 의료데이터를 사용한 이미

지 분류 task인 피부암 진단이라는 구체적인 상황에 적용해 결과를 관찰해봄으로써 실제로 ViT 모델이 이미지 분류를 잘 수행하는지를 관찰해보고 이미지 처리에서 중요한 개념이었던 convolution을 사용하지 않고도 이미지 분류 task를 잘 수행할 수 있는지를 판단해볼 수 있다는 점에서 중요성을 가진다.

2.2. Main figure

[Figure 1 about here.]

본 프로젝트에서 사용한 Main Figure는 아래의 "Figure 1. Vision Transformer(ViT)"[2] 이다. 그림에서 볼 수 있듯이, 입력 이미지는 Patch와 Position Embedding 과정을 거쳐 모델의 input으로 들어가게 된다. 단어 벡터가 input으로 들어가는 Transfomer 모델을 이미지 처리에 사용하기 위해서 이미지를 여러 patch로 나누어 patch들이 단어 벡터처럼 input으로 Encoder에 들어간다. 각 이미지 패치를 단어처럼 다루기 때문에 Flatten하는 과정도 포함하며 각 이미지의 위치 정보를 포함하는 Position Embedding 과정도 포함된다. Vision Transformer Encoder의 아키텍쳐는 기존의 Transformer Encoder 부분(=Bert)와 동일한데, normalization layer가 각각 Multi-Head Attention/MLP 전에 위치한다는 점과 활성화 함수로 GeLU를 사용한다는 점이 다르다. 최종적으로 나온 Encoder의 출력은 MLP Head를 거쳐 Classificaiton의 결과인 class를 도출하는 구조로 이루어져 있다.

2.3. Algorithm & Formulation

주로 살펴보고자 하는 ViT 모델은 위에서 언급한 Main figure의 구조를 가진다. 이 외에도 본 프로젝트는 피부암 진단 task를 진행하는 다양한 모델의 구현에 목표가 있으므로 Baseline이 될 수 있는 가장 기본적인 CNN 모델, 이미지 분류 문제에서 SoTA를 달성한 ResNet50, EfficientNetV2 B0 모델과 ViT 모델을 구현함으로써 총 3가지의 CNN 모델과 1가지의 Transformer 모델을 사용하여 성능을 평가해볼 예정이다. 사용한 모델의 구조는 Figure 2와 같으며, tensorflow에서 제공하는 ResNet50, EfficientV2 B0, ViT 모델을 사용하였고 ResNet50의 weight는 ImageNet 데이터셋으로 pretraining된 모델의 weight를 사용했다.

[Figure 2 about here.]

3. Experiments

3.1. Dataset & Computing resource

본 프로젝트에서 사용한 Dataset은 Kaggle의 Skin Cancer: Malignant vs. Benign 데이터셋[3]이다. Skin cancer에 대해 malignant/benign인 이미지가 포함되어 있으며, test와 train에 사용할 데이터가 나누어져 있다. 이미지의 size는 (224,224,3)이고 Train data에는 benign 이미지 1799개, malignant 이미지 1500개가 포함되며, test data에는 benign 1800개, malignant 1499개의 이미지가 포함된다. 앞서 언급한 논문의 ViT는 많은 양의 데이터를 사용해 구현했으므로 적지 않은 데이터셋을 사용하기 위해 이 데이터셋을 선택했다. 본 프로젝트에서는 Colab을 사용하여 코드를 구현했기 때문에 computing resource는 Colab에서 제공하는 GPU를 사용했다.

3.2. Experimental Design & Setup

앞서 언급했듯이 피부암 진단 task를 구현하는 baseline CNN 모델, ResNet50, EifficeintV2 B0, ViT 모델을 구현했다. 기본적으로 데이터 셋에서 나누어져 있던 test set은 test에 사용하고, train set은 train/validation으로 나누는 추가적인 전처리 과정을 거쳤다. 결과적으로 새롭게 구성된 Train, Validation data 개수는 Figure 3와 같으며 label 사이의 instance 수의 차이가 크지 않기 때문에 데이터 불균형 문제없이 잘 나누어짐을 관찰할 수 있다. Complexity가 높은 모델을 사용했으므로 overfitting을 방지할 수 있는 augmentation layer, dropout layer를 사용하였다. 모델 training 과정에서는 공통적으로 최대 epoch 수를 10으로 제한하였고, loss는 binary-cross-entropy, optimizer는 Adam, learning rate=0.001로 설정하였다. 마지막으로 모든 학습이 끝난 4종류의 모델의 loss, accuracy, ROC curve, Inference time을 관찰하면서 정량적, 정성적으로 각 모델의 성능을 평가, 분석해보았다.

[Figure 3 about here.]

3.3. Quantitative & Qualitative Results

Quantitative results를 나타내기 위해 다양한 성능 평가 지표를 사용해 보았다. 훈련을 모두 거친 4가지의 모델에 대해 test set을 사용해 일반화 성능을 관찰해보았을 때, 각 모델에 대해 epoch이 진행되면서 도출되는 Loss와 Accuracy는 Figure 4와 같다. 또한 Test 결과에 대해 표현한 각 모델의 ROC curve는 Figure 5와 같다. 최종적으로 각 모델에 대한 accuracy, precision recall, f1 score, MCC(Matthews Correlation Coefficient)를 Figure 6과 같이 표로 나타내며 한눈에 모델의 성능을 비교할 수 있게 표를 구성해보았다.

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

Qualitative results를 나타내기 위해 MCC와 Inference time, Trade-off를 나타낸 표를 Figure7과 같이 구성해

봄으로써 MCC와 Inference time의 trade-off를 관찰해보았다. 또한 report의 그림으로는 포함하지 않았지만, random으로 test data에 대한 결과(예측한 label이 옳은지, probability는 얼마인지)를 각 모델마다 코드를 구현하고 colab 결과로 살펴봄으로써 정량적인 분석을 해보고자 했다.

[Figure 7 about here.]

3.4. Results Analysis (with Figures & Tables)

Figure 4를 통해서 모든 모델에서 epoch이 진행될수록 loss는 낮아지고는 경향을 가지며 Accuracy는 높아지는 경향을 보이므로 학습이 잘 진행되었다는 결론을 도출할 수 있을 것이다. 하지만 다른 모델에 비해 ResNet50 model이 loss가 현저히 높고, accuracy가 꽤 큰 차이로 낮게 나왔다. 이 결과를 통해 보았을 때, weight를 모두 initialize한 후 학습을 시작한 다른 모델들과 달리, ResNet50 모델은 ImageNet 데이터를 기준으로 높은 정확도를 보이게끔 학습한 모델의 weight를 가지고 학습을 진행했기 때문에 local minima에 빠졌을 확률이 존재하며 Loss에서도 Validation Loss가 Training Loss와 꽤 큰 차이를 보이기 때문에 Class가 매우 많은 ImageNet에 학습된 ResNet model의 구조가 Binary Classification인 이번 프로젝트의 task를 수행하는 데 적합하지 않을 가능성도 존재할 수 있다고 생각했다.

Figure 5를 통해 나타낸 각 모델의 ROC Curve를 관찰해보았을 때, Figure 4와 마찬가지로 AUROC가 0.82인 ResNet model을 제외한 나머지 모델은 모두 AUROC가 0.9이상으로 이번 task를 수행함에 있어 좋은 성능을 가지고 있다고 해석할 수 있다. 더 나아가 EfficientNetV2 B0 model과 ViT model의 AUROC는 0.95로, Baseline CNN model보다 높은 성능을 가진 좋은 모델이라고 분석해볼 수 있다.

Figure 6과 7을 통해 정량적인 지표로 모델의 성능을 비교해볼 수 있는데, 표에서 볼 수 있듯이 모든 지표에서 ResNet50 < CNN baseline < EfficientNet, ViT 순으로 좋은 성능을 가지고 있다고 해석할 수 있다. 미묘한 값의 차이지만 EfficientNet보다 ViT가 조금 더 높은 score를 가지고 있다는 사실 또한 관찰할 수 있다.

최종적으로 Inference time과 MCC, Trade-off를 나타낸 Figure 8을 통해 앞에서 평가한 모델의 성능 분석에 대해 힘을 실어줄 수 있다. CNN 모델보다 parameter 수가 많고 time complexity가 높은 ViT 모델의 Inference rate이 다른 모델들보다 조금 더 큰 값을 가지지만 MCC 값은 제일 높은 값을 가지기 때문에 학습 시간은 상대적으로 조금 더 길지라도 task 수행에 대한 성능은 제일 좋다는 것을 알 수 있다. 두 값에 대한 Trade-off가 가장 낮은 것으로 보았을 때, 4개의 모델 중 피부암 진단이라는 이미지 분류 task를 가장 잘 수행하는 best model은 ViT 모델이라는 결론을 도출해낼 수 있다. 추가적으로 test samples 25개에 대해 모델이 평가한 label과 probability 값을 살펴보았을 때, EfficientNet, ViT 모델이 다른 모델에 비해 label 예측 정확도가 높으며 정확히 label을 맞춘 sample의 probability가 거의 99%가 되는 sample의 수가 많았으므로 높은 정확도를 보이는 좋은 성능의 모델임을 관찰해볼 수 있었다.

3.5. Discussion

이미지 분류 task에 Transformer가 사용된 ViT 모델이 CNN 모델에 비교하여 실제로 얼마나 좋은 성능을 가지는지에 대해 살펴보고자 함이 이번 프로젝트의 목표였다. 따라서 피부암 진단이라는 이미지 분류 task를 수행하는 다양한 CNN 모델과 ViT 모델을 구현해보고, 모델들에 대한 성능 평가를 진행해보았다. 최종적으로 다른 CNN 모델들에 비해 ViT 모델이 task를 수행함에 있어 가장 좋은 일반화 성능을 보였기 때문에 본 프로젝트의 목표는 성공적으로 달성했다고 생각할 수 있을 것이다. 또한 ViT 모델의 성능을 더 높이기 위해서 Classificatin(MLP) head의 width와 depth를 증가시키거나 fine-tuning 기술을 사용, ensembling method(blending, stacking, voting 등)를 사용하는 추가적인 프로젝트를 진행할 수 있다. 하지만 inference rate가 증가하거나 계산량이 많아질 수 있기 때문에 이를 고려해야 한다.

결론적으로 ViT 모델은 CNN 모델과 비교했을 때, convolution layer를 사용하지 않음에도 이미지 처리 task에서 비슷하거나 더 좋은 성능을 가질 수 있다. 또한 Transformer 구조를 거의 그대로 사용하기 때문에 확장성이 좋으며 큰 데이터의 학습에 좋은 성능을 나타내고 CNN보다 학습 과정에서 더 적은 계산 리소스를 사용한다는 장점이 있다. 하지만 Inductive bias의 부족으로 인해 CNN보다 더 많은 데이터가 요구되며 복잡한 구조를 가지기 때문에 inference rate가 더 길다는 단점이 존재한다.[4]

4. Future direction

앞의 discussion에서 언급했듯이, 이번 프로젝트의 연장선으로 ViT 모델의 성능을 더 향상시킬 수 있는 여러 가지 기법을 사용한 후 결과를 관찰하는 추가 프로젝트를 진행할 수 있다. 또한 이미지 분류 외에도 Image Captioning, Image Segmentation, Anomaly Detection 등의 이미지 처리 문제에서도 Transformer 기반 모델이 좋은 성능을 가진다는 연구 결과가 많이 존재한다.[1] 따라서 다양한 task에 적용한 Transformer 모델이 어떤 성능을 보이는지를 관찰하고 탐구해보는 것도 좋은 프로젝트 주제가 될 것이라고 생각하는 바이다.

References

[1] 문상선, "Vision transformer (vit) 란? - 정의, 원리, 구현, 응용분야," 2023.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] C. FANCONI, "Skin cancer: Malignant vs. benign," 2019.

[4] leehyuna, "Vision transformer(vit)," 2023.

List of Figures

1 Vision Transformer Architecture 6
2 Architecture of Models(baseline,ResNet,EfficientNet,ViT) 7
3 Train & Validation data distribution) 8
4 Loss, Accuracy of Models(baseline,ResNet,EfficientNet,ViT) 9
5 ROC curve of Models(baseline,ResNet,EfficientNet,ViT) 10
6 Evaluation Table of Models 11
7 Inference rate,MCC,Trade-off of Models 12

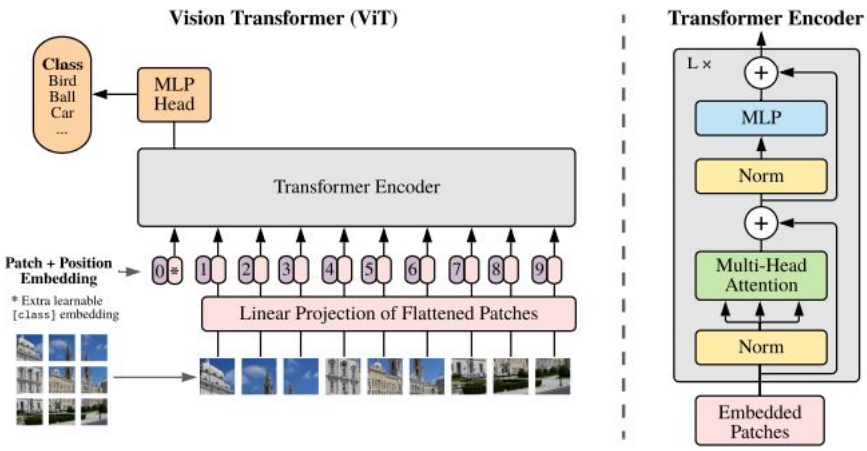

Figure 1. Vision Transformer Architecture

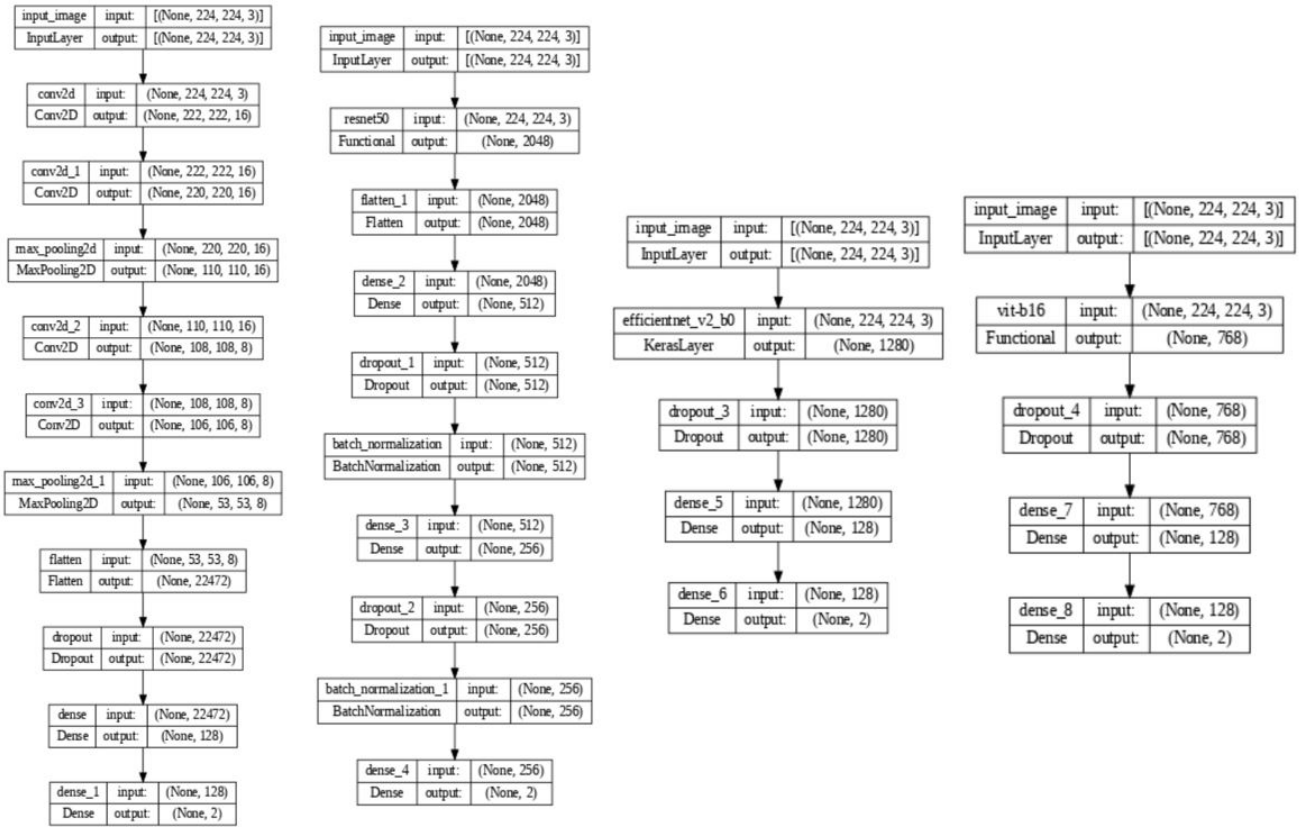

Figure 2. Architecture of Models(baseline,ResNet,EfficientNet,ViT)

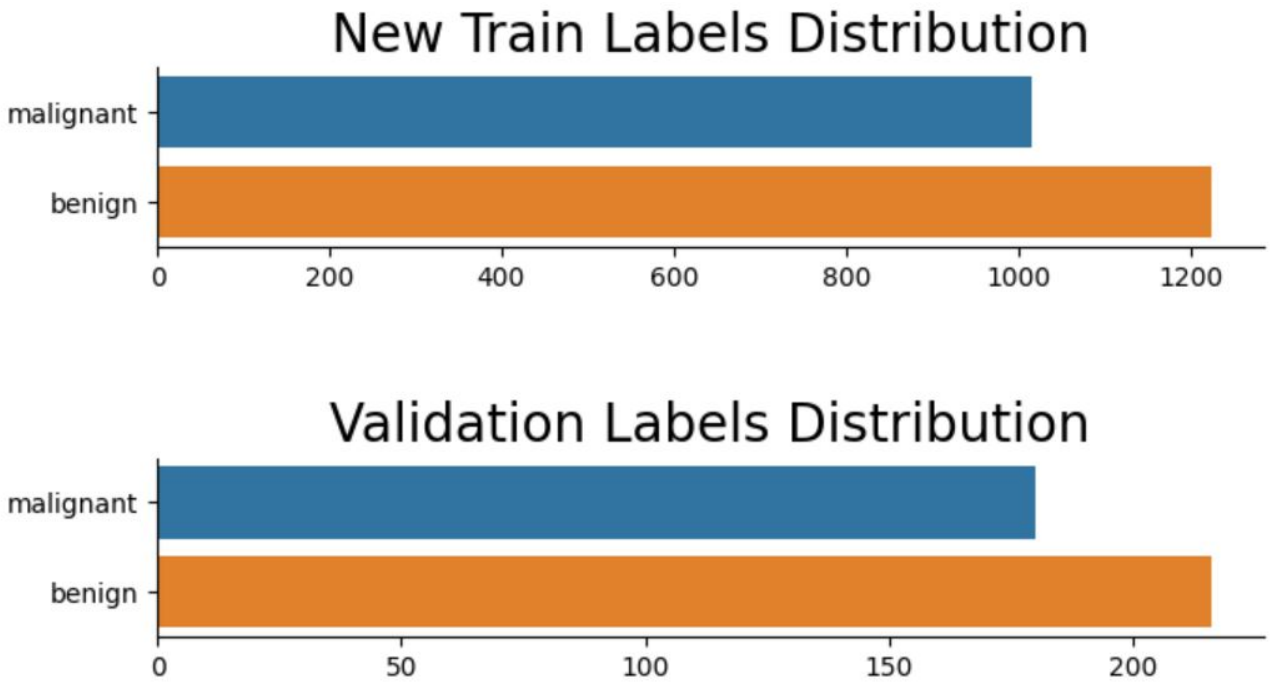

Figure 3. Train & Validation data distribution)

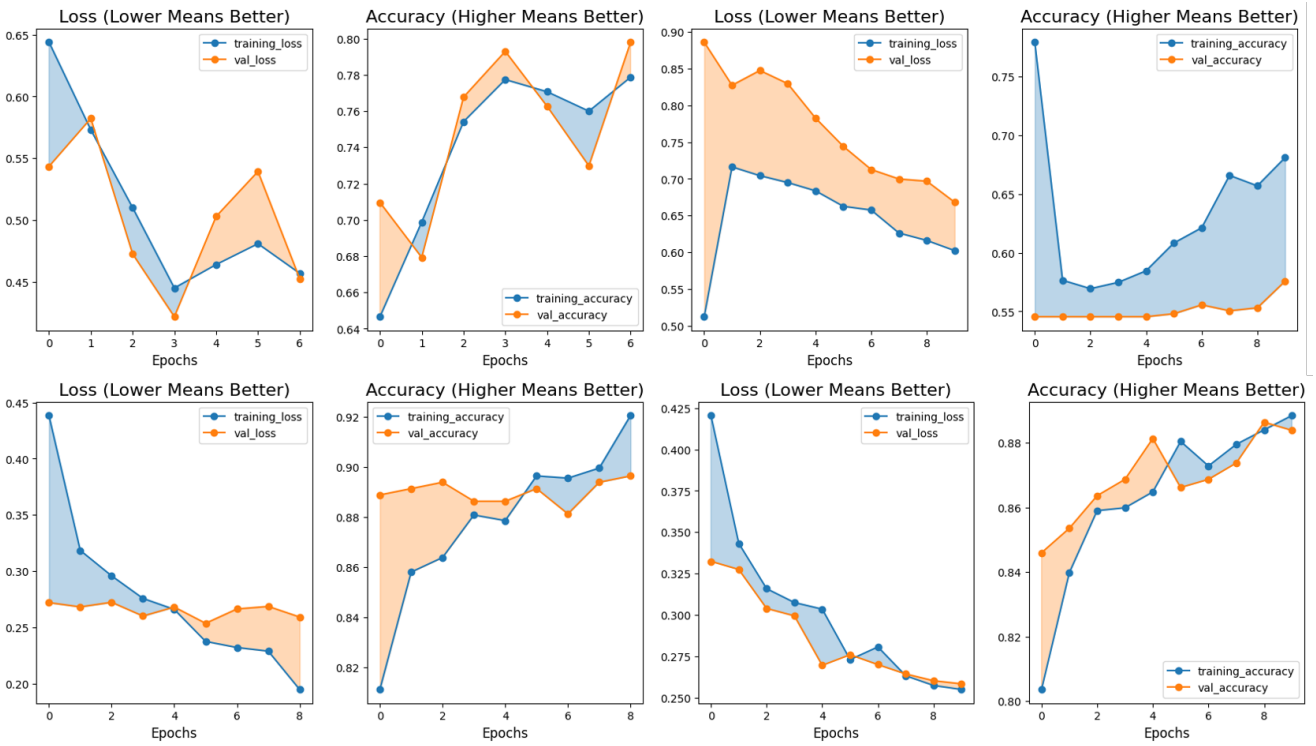

Figure 4. Loss, Accuracy of Models(baseline,ResNet,EfficientNet,ViT)

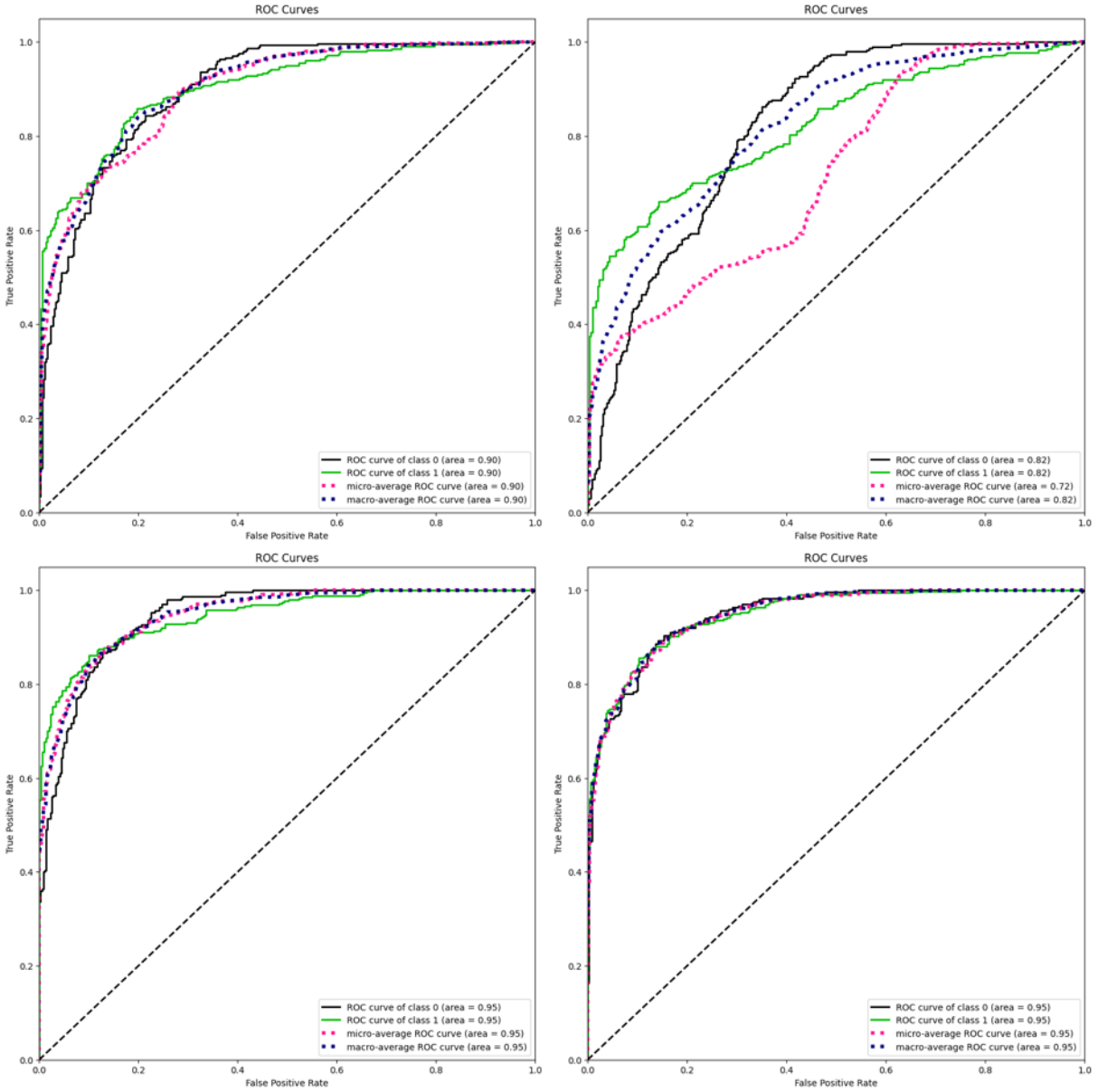

Figure 5. ROC curve of Models(baseline,ResNet,EfficientNet,ViT)

	accuracy_score	precision_score	recall_score	f1_score	matthews_corrcoef
model_cnn	0.787879	0.825786	0.787879	0.785493	0.615727
model_resnet50	0.580303	0.670484	0.580303	0.474570	0.171283
model_efficientnet_v2	0.865152	0.866466	0.865152	0.865361	0.729945
model_vit_b16	0.868182	0.868192	0.868182	0.867943	0.733769

Figure 6. Evaluation Table of Models

Model	Inference Rate	MCC	Trade-off
CNN(Baseline)	0.00120	0.6157	0.3843
ResNet50	0.00151	0.1713	0.8287
EfficientNet V2 B0	0.00144	0.7299	0.2701
ViT	0.00482	0.7338	0.2663

Figure 7. Inference rate,MCC,Trade-off of Models