SIIM-ISIC Melanoma Classification

27th Place

1st With Context

Yuval Reina
Zahar Chikishev

Agenda

1. Background
2. Summary
3. Base Models
4. Transformer Models
5. Ensembling
6. Important findings
7. Real world model

Background

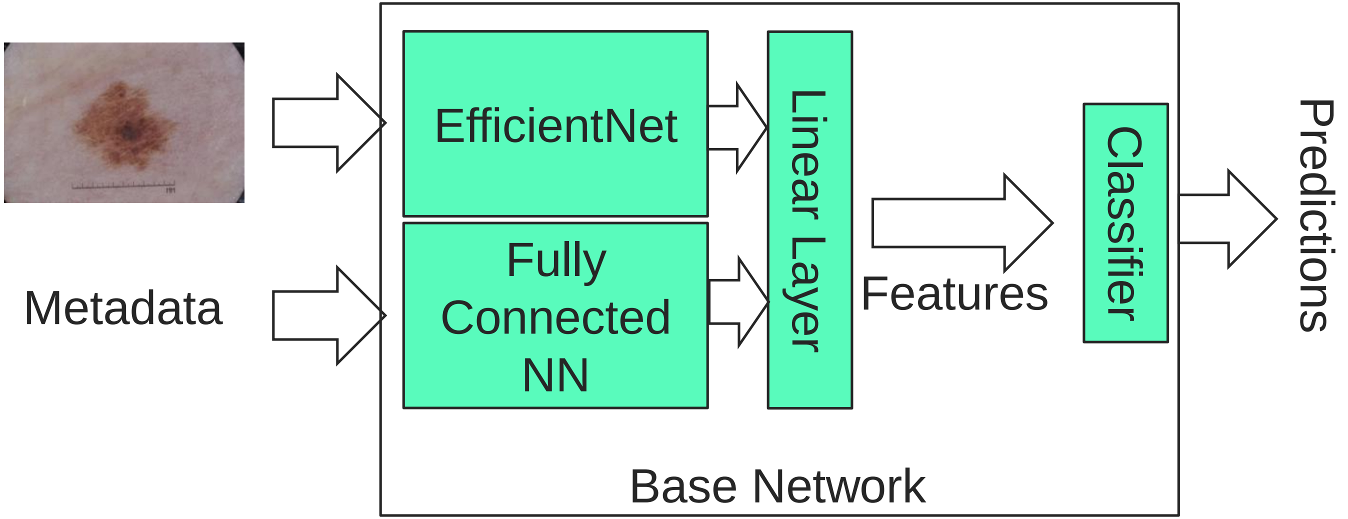
- Yuval Reina
    - BScEE, MBA, COO at Ceragon
    - Hobbyist, Self-Education

- Zahar Chikishev
    - M.Sc. in applied math
    - Kaggler

Summary

Two stage solution:

- Base model for feature extraction per image



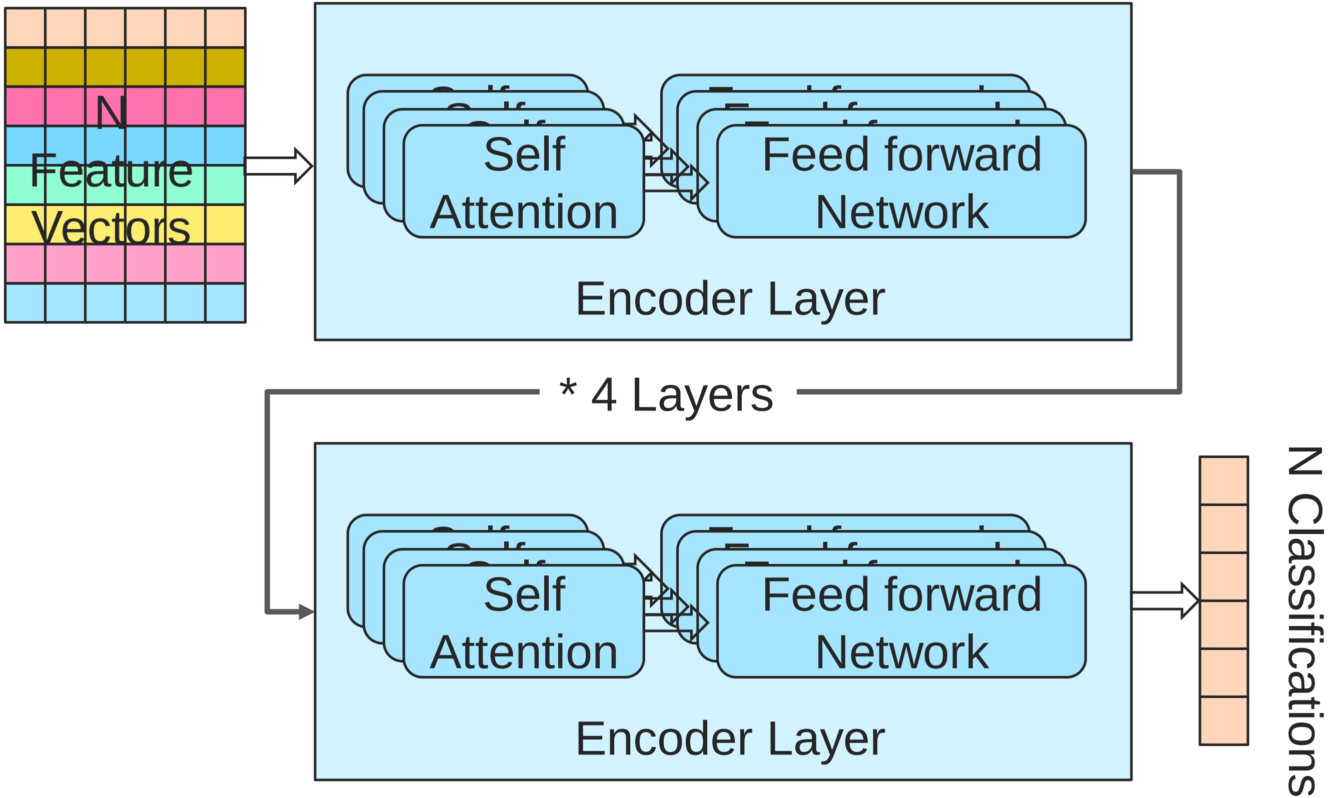
Pre – trained models:

- EfficientNet B3, B4, B5, B6, B7

Summary

Two stage solution:

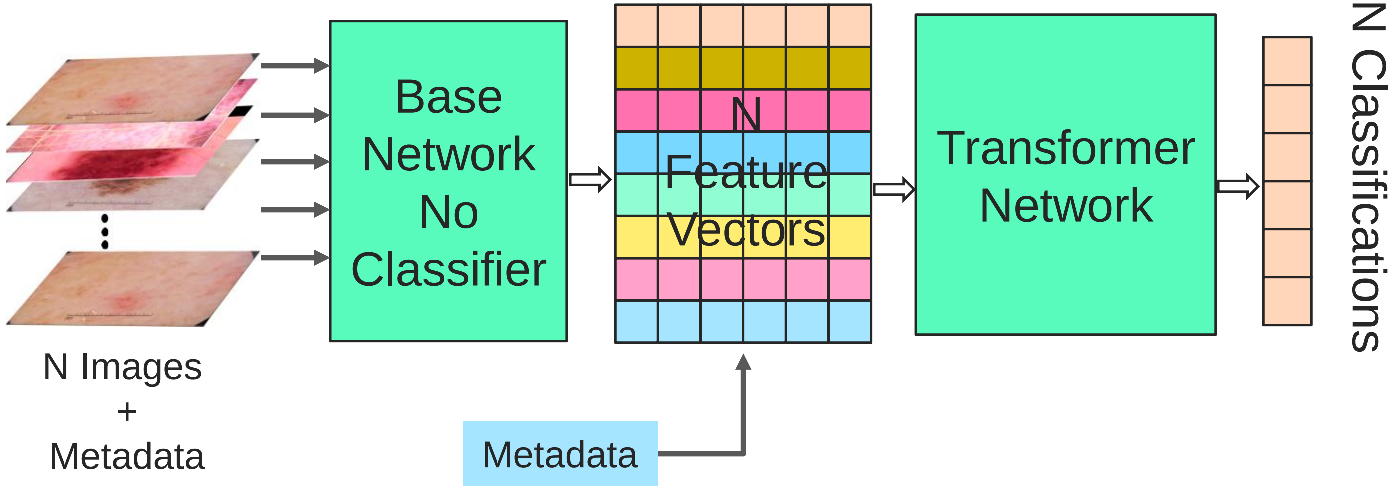
- Transformer model for context learning



Summary

Two stage solution:

- Base model for feature extraction per image
- Transformer model for context learning



Base Models

As base model we used different types of EfficientNet
B3, B4, B5, B6, B7
Pre-trained on Imagenet using Noisy student algorithem

Loss Functions:

- Cross Entropy on 8 diagnosis classes

Metadata:

- Age
- Sex
- Anatomic site

Number of Features – 256

Base Models – Cont.

Augmentation

- Random resize + crop
- Random rotation
- Random flip
- Random color jitter (brightness, contrast, saturation, hue)
- Cutout - erasing a small rectangle in the image
- Hair - Randomly adding “hair like” lines to the image
- Metadata augmentation

Training

The data was form ISIC 2019 + 2020 competitions

The goal - minimize the loss after the classification layer.

Feature extraction

The feature where extracted by the output of the last layer before the classification layer.

Transformer Model

- The transformer is a stack of 4 transformer encoder as described in Attention Is All You Need with 4 attention heads
- The input to the transformer is N feature vectors from the same patient.
- For simplicity, 32>=N, if the patient has more then 32 images, they are randomly grouped to 32 images groups
- Metadata is embedded and added to the feature vectors

Transformer Training

Two stage training:

1. ISIC 2019 (only 1 image/patient) + ISIC 2020
2. Fine – tune on ISIC 2020 alone

Augmentation

- Feature extraction of base model on 12-16 differently augmented images
- Select different augmented image in every epoch

Loss Functions:

- Cross Entropy on 8 diagnosis classes
- Focal Loss in one model (B5) – similar results

Inference

1. Use the base model to extract 16 sets of features for each image using augmentation
2. Randomly Divide all the images of a patient to groups of size 32 (max)
3. Select one random example for each image in a series and construct to full series.
4. Inference using Transformer network.

TTA - Do the above 32 times (with random selections, and grouping) and average the results

Training And Inference Times

Setup:

- CPU – Intel i9-9920
- RAM – 64G
- GPU – Tesla V100 32G / Titan RTX (20% slower)

Training:

- Base Models: ~ 3 (B3) – 11 (B7) h/fold
- Feature extraction: 4 – 14 h (3 * 3 folds, 12 TTA train 16 Test)
- Transformer: 1h/model
- Total time for all models ~ 2.5W for 1 GPU

Inference:

- Inference base model: ~20 min - 1h /model (for 12xTTA)
- Transformer model: ~ 2min (for 32xTTA)

Ensembling

Ensembling was done by averaging the predictions before Softmax

Without Context Submission:

A. EfficientNet B3 noisy student image size 400*600
B. EfficientNet B4 noisy student image size 400*600
C. EfficientNet B5 noisy student image size 400*600
D. EfficientNet B6 noisy student image size 600*900
E. EfficientNet B7 noisy student image size 400*600

With Context

All the “without context” model +

1. Transformer on features from A.
2. Transformer on features from B.
3. Transformer on features from C using focal loss
4. Transformer on features from D.
5. Transformer on features from E.

Results

With Context Submission:

Private LB - 0.9441

Public LB - 0.9575

Without Context Submission:

Private LB - 0.9430

Public LB - 0.9582

Best Single model :

Transformer on EfficintNet B5

Private LB - 0.9397

Public LB - 0.9602

Simpler and Real World Model

A simpler model would be:

- One base model - EfficientNet B5 for feature extraction
- Transformer model
- Keep 3 * 3 folds
- Keep TTA

In real world, previous images of a patient are already tagged, using this information to predict the current tagging, makes the model a seq2seq model – which is what a full transformer (encoder +decoder) was invented for.

Important and Interesting Findings

- Due to small number of positive examples at the test set it was impossible to get reliable results.

kaggle™