


Team 11: Final Project Report

Team Name : E-CAT(Electronic Computer and Artificial Intelligence collaboration Team) 

Team Members : Yoon Byungjun, Kim ChoongHan, Kim Chiwon, Park Chanho

As our final project, we conducted image classification. Then, we got 1st place in the leaderboard. Our main topic was “**Can transformer and CNN capture different feature from image?** Through a total of 3 Gateway meetings, our team proceeded in order of : Data analysis and preprocessing, Data analysis and preprocessing, Hyperparameters tuning and model optimization

First we found that DL20 dataset is a subset of ImageNet, so we add images from ImageNet without duplication with DL20 test data. The total amount of images increased about 4 times, therefore we expected our model would be improved without overfitting. Initializing from a pretrained weight allows fast convergence and better performance, so we initialized our model from ImageNet pretrained weight. We hypothesized that pretrained weights would perform best in high resolution. GAN-based super-resolution techniques are employed, but super-resolution degrades data, which remains for further improvement.

EfficientNet is best among CNN which is standard for Image Classification. Recently, transformer based ViT model achieved renowned performance on image classification. So we decided to ensemble these two architectures to capture different features on data. Our team experimented with pretrained weights, RandAugment, external data. In addition, mixup, cutmix, color jitter, and other techniques which reported to perform well from previous works are used.

To analyze the results, we investigate the number of mispredictions of two models with respect to classes. Although ViT generally makes fewer errors, EfficientNet-v2 outperformed some classes. So, It seems to be reasonable to ensemble these two models according to the pipeline in the poster. However, ensembling failed to achieve better accuracy by. Hence, our team selected the ViT-L/16 model as the final model.

Possible Improvements

During the project presentation, questions were asked about the ensemble model. Among multiple models, if a model is convinced of the answer with a very high probability, the ensemble is organized in a way that follows the answer of the model. We haven't tried the ensemble in this way, so we can improve it by using this way. For example, using a confidence base ensemble is one possible option for reflecting different features of models. Regarding the ensemble method, to capture different features our model uses two models. Instead of this, MLP-Mixer, which we learned from other teams' presentations, can be added for a soft voting ensemble. Another team uses K-fold validation for their model. According to the training log, it is obvious that 18, 19 class is harder. If we can reconstruct the training loop with sufficient computation resources and time, the model can use all data as a test set and show more generalized performance.

While the number of parameters of EfficientNet-V2 was 121 million, and ViT-L/16 is 300 million. Due to insufficient time during the final project, we could not train many types of large models. If other CNN model has as many parameters as ViT will improve our ensemble performance. We also believe that if we could train other types of model, we can improve further by ensemble technique. Also, ViT-14/H has a greater number of parameters compared to ViT-L/16. During the project period, our

team wasn't able to converge some SOTA models, such as ViT-14/H and EfficientNet-L2 with noisy student and SAM optimizer. SAM optimizer needed to rewrite the training loop, which was not able to finish before the deadline. If we can converge those models, improvement of performance is guaranteed. One problem with ViT-14/H due to VRAM limit of GPU and small batch size. One problem regarding GPUs was that smaller VRAM memory enforce smaller batch size. This always leads to unstable slow training. Greater memory may be helpful for reproducing SOTA performance.

During the project, we failed to validate our hypothesis about super-resolutioning. This is because the superresolution model that we used is optimized for cleaning noisy images or enlarging a small portion of an image, rather than downsampled whole image. Our DL20 is downsampled from whole large image. If our team build a new super-resolution GAN model that optimized for our task, our model may fully utilize pretrained weight.

Impressions

Our Team Name is **E-CAT**, which was named after the first letters of **E**lectronic **C**omputer & **A**rtificial Intelligence collaboration **T**eam. The awkward first meeting quickly developed into teamwork that was organized and full of positive energy. Through regular Gateway Meetings, our team resolved issues by establishing a task execution process, confirming performance, discussing future schedules, and collecting ideas. As a result, our team was able to achieve the best results in the Kaggle Competition.

Each of our team members has faithfully performed their respective roles. Student Byungjun Yoon configured the overall architecture of the model and optimized the model through programming, Student Chiwon KIM made EfficientNet models and parameter tuning. Student Choonghan Kim made Vision Transformer models and compared them with model accuracy. Student Chan-Ho Park performs data preprocessing, and analyzes the latest thesis data on Super Resolution and applies it to the dataset. In this way, each of our team members did their best to create the top performance.

Listening to other teams' presentations, we felt the following. First, The other teams faced problems that we did. For example, DL20 does not have much data, so no matter how good the model is, it could not guarantee high performance. The lack of data was the first problem we were confronted with. Other teams use data augmentation to solve it. Second, not all the results of the experiment were ideal. Experimental results that were different from those suggested in the paper. For example, if we do an ensemble model, theoretically, the performance must be better than one model. However, in our case, just one model(ViT) shows better performance than the ensemble model. Similar situations could be observed in other groups.

Finally, We were so impressed how other teams performed great without employment of external data. Adding a lot of additional data always guarantees higher performance. This is the basic principle we learned in deep learning. However during poster presentation time, other teams' presentations give insight into how they construct the model without external data. Learning and Adapting their method gives direction on how we should improve our model.

Through this project, We came to realize that cooperation and voluntary participation ultimately produce good results, and that successive failures and successes are necessary to accurately define a problem and find a suitable solution. A precious relationship has been made, and we hope to carry out similar projects with enthusiasm once again in the future. We would like to express our gratitude once again to the professor and teaching assistants who passionately taught deep learning. Thank you all for your hard work. Again, Thank you very much.