Summary

The paper provides an analysis toward the performance of Vision Transformer (ViT) on small datasets by comparing the performance of ViT trained on a small dataset from scratch with the performance of resnet50. It also visualizes the saliency map of the two models. In addition, the paper re-implemented Shifted Patch Tokenization and Locality Self-Attention, attempting to improve the performance of ViT on small datasets.

Pros:

This study compares the saliency map of ViT and CNN. The shape of the saliency map for ViT is a good support from the claim that how ViT treats everything in a patch as a whole.

This paper gives enough information to reproduce the results.

Cons:

This study does not include the comparison between CNN’s accuracy and ViT’s accuracy. I think it’s necessary to include this comparison to claim that CNN is better than ViT on small datasets.

The numerical citations are not in order. Figure 2 does not have a right boundary. [Clarity]

We can predict the change in training speed when changing the hyperparameter for the same model architecture, so it would also be better to include a training speed comparison between CNN and ViT.

Anonymous User , Apr 21 at 2:24am

A comparison of vision transformers, a recent extension of the state-of-the-art

in NLP to image classifiation, to convolutional neural networks. In particular,

a comparison of the two architectures' performance on small datasets, through

the use of saliency maps.

Quality

- The saliency maps were a good idea; they provided very interesting results

- The experiments seemed to me to be lacking, and not much data was provided.

For example, Figure 2 demonstrates the results of tweaking a few hyperparameters

in one direction. The results of a more exhaustive grid search could have been

more informative. Mention hyperparameter settings in discussion: the models are loaded from packages directly, thus can’t tune too many hyperparameters.

- The issue of why vision transformers perform worse than CNNs on small datasets

as mentioned in the abstract was only briefly addressed in the report.

Note: above fix is done by adding a paper reference with ideas mentioned in the paper for more supportive and persuasive conclusion.

- There was no data regarding the performance of the CNN, so it was not clear

how the two compared to each other in the experiment.

Clarity

- The concept of vision transformers, and how they differ

from CNNs, was well-explained

- I was not able to understand the shift-patch-tokenization

method from reading the report. It was not clear from the

pseudocode how the images were being concatenated (along which

dimension), why there was a kernel involved, or how the patches

were being extracted. As a result, I did not understand why the

method was meant to be helpful. More detail should be added to

make it understandable for the reader.

- It was not clear to me what the adjacency matrix being referred

to in regards to locality self-attention was, or why this masking

was helpful. I had to refer to the cited paper to understand.

- Exactly what occurred in the experiments was not clear from the

report. What is the difference between the VIT and MVIT? Was the VIT

a vanilla vision transformer while shifted-patch-tokenization and

locality self-attention was applied to the MVIT? Or was the difference

something else?

Originality

- There are plenty of citations in places where necessary

- The saliency maps were an interesting approach and very clearly demonstrated

one of the differences between CNNs and vision transformers.

- This was an analysis and understanding report rather than an area of new

research, so the criteria are a little difficult to apply. But I found that

the analysis did not go very deep, and the results provided, apart from the

saliency maps, do not shed much light on the performance of ViTs on small datasets,

or their weaknesses relative to CNNs

Significance

- The result that increasing the patch size substantially can harm performance was interesting,

and well-justified.

- For the most part, the results and experiments did not provide unique insight. The

conclusions regarding hyperparameter tuning, i.e. that increasing the number of

transformer layers can cause overfitting and poor generalization, or that a lower

batch size can improve quickly for a fixed number of epochs due to a

larger number of parameter updates, were not necessarily new.

Score 6/10

Anonymous User , Apr 21 at 2:54pm

Please ignore my previous comments.

Summary

The paper provides an analysis toward the performance of Vision Transformer (ViT) on small datasets by comparing the performance of ViT trained on a small dataset from scratch with the performance of resnet50. It also visualizes the saliency map of the two models. In addition, the paper re-implemented Shifted Patch Tokenization and Locality Self-Attention, attempting to improve the performance of ViT on small datasets.

Quality: [medium]

+ This study compares the saliency map of ViT and CNN. The shape of the saliency map for ViT is a good support from the claim that how ViT treats everything in a patch as a whole.

- This study does not include the comparison between CNN’s accuracy and ViT’s accuracy. I think it’s necessary to include this comparison to claim that CNN is better than ViT on small datasets.

- We can predict the change in training speed when changing the hyperparameter for the same model architecture, so it would also be better to include a training speed comparison between CNN and ViT.

Clarity: [medium]

+ This paper gives enough information to reproduce the results.

- Figure 2 does not have a right boundary.

Originality [medium-high]

+ This paper uses saliency map to evaluate the performance. I think it is novel to use it to analysis the reason why ViT performs poorly.

Significance [medium]

- Maybe some vanilla transformer needs to improve performance on small dataset, but I think that ViT has pre-trained weights and most researchers choose to use the pre-trained version, which is also true in NLP area.

Score: 7/10

Anonymous User , Apr 22 at 12:03am

This final report aims to compare the performance of a large vision transformer with that of a more conventional convolutional model and then attempt to explain the difference in performance. The authors’ strategy to achieve this was to use a saliency map to visualize which sections of images were most strongly influencing a model’s prediction. Additionally, they reimplement a vision transformer that is specially adapted to work on small datasets.

Before reading my comments, I want to make clear that I understand computational resources may have been limited and may provide a valid rebuttal to my any of my critiques below.

Quality

The quality of this report could be significantly improved.

- No easily identifiable theorem was stated

- Discussion and Conclusion are condensed into one section.

- Citation is missing under “Re-Implement Modified Vision Transformer and …”

- It was unclear if VIT or Resnet had gone through any hyperparameter tuning to ensure they were performing well before their saliency maps were compared. If not, it makes the comparisons less valid and should be included as a limitation.

- However, using Saliency maps to give a visual explanation for differences of performance between CNN based models and vision transformers was interesting and insightful.

Clarity

Related work, methods, and abstract were fairly clear.

- Outline of experiments was fairly clear from abstract

- Related work provided necessary background and discussed relevant papers

- Methods described what was done over the course of this project.

Understanding the results was difficult.

- Results don’t state whether Vit or CNN perform better; making it difficult to interpret saliency maps.

- It is unclear why the hyperparamter tuning of the modified vision transformer is included, seeing as it doesn’t seem to add to the problem motivated in the abstract and introduction.

Explanation: the topic we chose is “analysis” of vision transformer on small dataset, thus including a modified vit and perform hyperparameter analysis is also part of analysis.

Limitation: the report didn’t apply modified Vision transformer model on Intel image dataset

Originality

This project had some original aspects.

- Using saliency maps to provide a qualitative comparison of why vision transformers perform worse than CNN based models has the potential to provide an interesting visual explanation for performance differences between the two models.

Significance

The significance of this work could have been more clearly motivated.

- Providing a qualitative, visual explanation for why vision transformers are outperformed by similarly overparametrized CNNs would have been relevant.

- This was not clearly motivated in the introduction however.

Score 6.5/10

Anonymous User , Apr 22 at 6:57pm

Analysis of Vision Transformer’s Performance on Small Datasets

In this paper, the authors used the "Intel Image Classification" dataset to compare the performance between the large vision transformer network, "vit\_base\_patch16\_224", and the traditional convolutional network, "wide\_Resetnet-50-2", using saliency map. The authors focused mostly on the vision transformer network by re-implementing the vision transformer network to boost its performance. The authors confirmed the original hypothesis that with a smaller dataset, the traditional convolutional network still outperforms the vision transformer network.

Quality

- The paper is technically sound, it has a clear goal and methodically arrived at the conclusion

- However, the paper failed to present enough data for CNN, it might be best to show more CNN results weather than just mentioning it.

Clarity

- An in-depth research of the vision transformer model is provided by the authors, the author presented many test results for its attempts to hyperparameter tunning

- Then again, it contains mostly the training results of the vision transformer but almost none of the CNN model, the authors should provide at least a training graph in the appendix

Originality

- There is some originality in the paper as it compares specifically the two models "vit\_base\_patch16\_24" and "wide\_Resetnet-50-2"

- Although the project is very similar to many existing papers that compare the performance of CNN and ViT

Significance

- The paper provides perhaps little significance to the community, since it just reinforces the idea that ViT performance is worst than CNN with smaller data set.

Score 7/10

Anonymous User , Apr 22 at 9:58pm

Summarized sentences:

The transformers gained great amount of attention for various deep learning tasks. However, the vision transformer is known to perform worse when trained on smaller dataset. Hence, this project compares the performance of a CNN and a vision transformer on a small size dataset. The insight this project provides is that the CNN model performed much better than the vision transformer on smaller dataset.

Quality

- The report includes a clear algorithm block to describe the shifted-patch-tokenization method used in a vision transformer. The report could improve by adding the significance of what they did and the insights they got clearly, especially in the abstract section. Additionally, because this report concerns the visual attention maps of a model, the report’s background section could mention or include figures that show what kind of output maps can be obtained at the beginning of the report.

Clarity

- The report could give more details on what sort of images(such as mentioning classes) the dataset they used contained, and give model architecture diagrams or figures for an easier understanding of the methods. Moreover, valuable insights and the major objectives of the project are often revealing in the later part of the report, so some of them could be highlighted or hinted at in the beginning of the report for better understanding by the reader. I also noticed that some references are missing in the report.

Originality

- The project introduced a comparison of CNN and vision transformer’s performances on smaller dataset and the saliency maps differences resulted in a greater insight into why the CNN performed better than the vision transformer on the smaller dataset.

This project could have done more analysis on if a certain kind of images resulted in the conclusions that the CNN performs better than the vision transformer on smaller dataset or a performance comparison using different size of the dataset with both CNN and the vision transformer to more firmly conclude if one or the other model is better for the smaller dataset.

Significance

- The conclusions are not particularly unique or ground-breaking, but it confirmed that some other models (in this case the CNN) can perform better than a vision transformer on smaller datasets. The report also identified various future research areas, which will be useful for the deep learning community.

Anonymous User , Apr 23 at 5:01pm

Dot product attention reference required!