Comparing Convolutional Neural Networks and Vision Transformers for Image Classification

Kristóf Kássa
Budapest University
of Technology and Economics
kassak@edu.bme.hu

Lilla Barancsuk
Budapest University
of Technology and Economics
barancsuk.lilla@vik.bme.hu

Abstract

Convolution Neural Network (CNN) algorithms have been prominent models for image classification, but in recent years Transformer based methods have also started to gain popularity. In an attempt to get a clear view and understanding of the two architectures for image classification tasks on a cloud dataset of approximately 2000 data points, a framework was designed to compare the characteristics of CNN and Vision Transformer (ViT) for image classification. For each model we provide a comprehensive review of architectural and functional differences. Then we compare their computing capacity requirements, validation accuracy and training time on an online image dataset with our own implementation of input pipeline from scratch.

1 Introduction

The Convolutional Neural Network (CNN) is a sub-type of Neural Networks. CNN [6; 7] models are exceedingly well suited for image recognition and classification applications due to their built-in convolutional layer(s) aiming to reduce the high dimensionality of images without any information loss. The architecture is based upon the idea that one pixel of the given picture is dependent on its nearby pixels (brightness, color or contrast). Therefore, a CNN based algorithm can extract important low/mid/high level features [12] and edges by applying filters on specific subsets of the given image in a convolutional manner.

The Vision Transformer (ViT) architecture however, adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data and used primarily in the field of natural language processing (NLP) [10]. The use of transformers in computer vision is still limited, although recent studies showed that transformers can be successfully applied for various image processing tasks.

The goal of this group work is to gain thorough understanding of each architecture to learn about their strengths and limitations.

2 Related work

Transformers were proposed by Dosovitskiy et al. in [3] for image classification with supervised machine learning. Comparison of results of ViT against CNN architectures on various datasets shows that despite of the fact that ViT requires considerably more computational resources, it can provide better performance for some datasets upon certain training conditions than CNN models. In fact, they showed that the chances of the model performing better are higher if filtering is not applied for extraction but the entire image is processed by the model (as in ViT). As a result, numerous studies were formed to replace CNN with ViT in real-world applications [1; 5].

Table 1: The analyzed Vision Transformer models.

Model name	Approximate parameter count	Patch size
vit_tiny_patch16_224	5×10^6	16
vit_tiny_r_s16_p8_224_in21k	10^7	88

3 Methodology

3.1 Dataset

For our study, we used the Swimcat-extend dataset [9] for cloud image classification. The dataset contains 2100 images of 6 different types of sky conditions, considering cloudiness: Clear Sky, Patterned Clouds, Thin White Clouds, Thick White Clouds, Thick Dark Clouds, Veil Clouds. Before streaming them into our models — 5 different configurations for both CNN and transformer architectures — all images are normalized and reshaped to a size of 224x224. Given the small size and class similarity of our cloud dataset we apply K-fold Cross Validation for each CNN and transformer model training to comprehensively evaluate their performance.

3.2 Data augmentation

To prevent overfitting, diversify the training set and artificially increase the amount of training data, we apply data augmentation. We define 3 distinct levels (none, low, high) of data augmentation to analyze the effect of augmentation on our model accuracy. While low augmentation layer performs random rotation, high augmentation layer performs rotation, flip, and zoom, as detailed below. For those models where we apply high data augmentation during the learning process, we expect greater robustness and generally improved accuracy and performance.

3.3 Neural network architectures

To compare the two network architectures, four pretrained CNN and ViT models were selected for our test framework. To analyze the effects of parameter count we have chosen 2–2 different models in complexity — number of parameters 5×10^6 and 10^7 — for the CNN and transformer models both.

3.3.1 CNN models

The first selected clonvolutional network model is *EfficientNetB1* which uses a technique called 'Compound Scaling' during its learning process, i.e., uniformly scales all dimensions of depth, width and resolution using a compound coefficient, unlike conventional practice that arbitrary scales these factors [8]. The parameter count of EfficientNet is approximately 10^7.

The second model is *ResNet18D*. This model utilizes an average pooling tweak for down sampling. Instead of hoping each few stacked layers directly fit a desired underlying mapping, residual nets let these layers fit a residual mapping. They stack residual blocks on top of each other to form the network [4]. The parameter count of this model is about 5×10^6.

3.3.2 Transformer models

Similarly to CNNs, two vision transformer architectures were selected. They are both a variation of the original Vision Transformer model proposed by Dosovitskiy et al. in [3]. Both models are trained on ImageNet-21k, at a resolution of 224x224 pixels. Model details are seen in Table 1.

3.3.3 Frozen vs. trainable weights in the pretrained models

To kick-start the training, we employed transfer learning [11]. In our study, using transfer learning is necessary, since we do not have a large enough dataset for achieving a good performance on a model trained from scratch. Prior to utilizing them in our study, both CNN and ViT architectures were pretrained on the ImageNet-21k dataset [2] for image classification. To analyze the effects of trainable parameter count, for some of our experiments, the pretrained base layer weights were frozen, while for other experiments, the base layers were were left trainable (Table 2.).

Table 2: Parameter configurations for the experiments performed.

	Parameter count	Trainability	Augmentation	CNN model name	ViT model name
1	5×10^6	frozen	none	cnn_5m_f_na	vit_5m_f_na
2			low	cnn_5m_f_la	vit_5m_f_la
2			high	cnn_5m_f_ha	vit_5m_f_ha
4	5×10^6	trainable	high	cnn_5m_t_ha	vit_5m_t_ha
5	10^7	trainable	high	cnn_10m_t_ha	vit_10m_t_ha

3.4 Model structure

The resulting models consist of the following layers:

- Input layer
- Optional augmentation layer
- Pretrained model (CNN or ViT, trainable or frozen)
- Dropout layer for increasing robustness
- Dense output layer

To increase robustness, an additional dropout layer is added, and for extracting features for the classification, a trainable dense layer was placed on top of the pretrained models in each case.

3.5 5-fold cross-validation

K-fold cross-validation is a technique used for evaluating machine learning models. The method splits the input dataset into distinct K sample sets, and trains the network separately for each set. This technique can increase the reliability of the performance evaluation and can predict the outcome of unseen data, thus offers a less biased evaluation. For our analysis, 5-fold cross-validation is used.

4 Experiments

Using the model, size and augmentation variations described in section 3., we compiled 10 experiments with distinct hyper-parameter configurations as seen in Table 2. We trained each model defined in Table 2., 5-fold on the entire cloud dataset, and saved categorical accuracy and loss scores throughout the training process.

5 Results

5.1 Loss and accuracy curve analysis

Figures 1,2 shows the validation accuracy and loss respectively of the ViT and CNN networks for the five configurations. Curves show an average and deviation across the 5 training runs. Solid lines shows the mean, and the shaded region is related to the empirical standard deviation. Trainable networks show a more rapid learning compared to frozen models. The main difference between the ViT and CNN result, is that the latter converges faster. On the other hand, CNNs show a higher level of uncertainty in terms of accuracy. Interestingly, models with low level of augmentation seem to perform better on the validation set. This is due to a strong similarity in the images of the validation set. To mitigate this problem, the models were evaluated using an unseen test dataset, as presented in section 5.3.

5.2 Comparing model performances

Figures 3.,4 compares the validation accuracy and loss respectively of the models on the training dataset. For this analysis the minimum loss and maximum accuracy were selected from each training run. Trainable models significantly outperform the frozen models. CNN models show a higher loss and lower accuracy compared to ViT, but the difference is less significant for the trainable models. There is no significant difference in terms of model size. The results show, that the most relevant parameter with respect to performance is trainability.

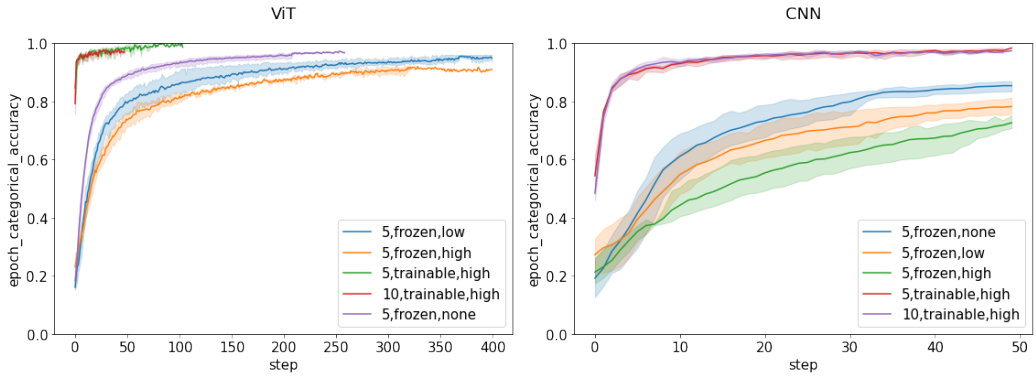

Figure 1: Categorical accuracy of the validation set during training.

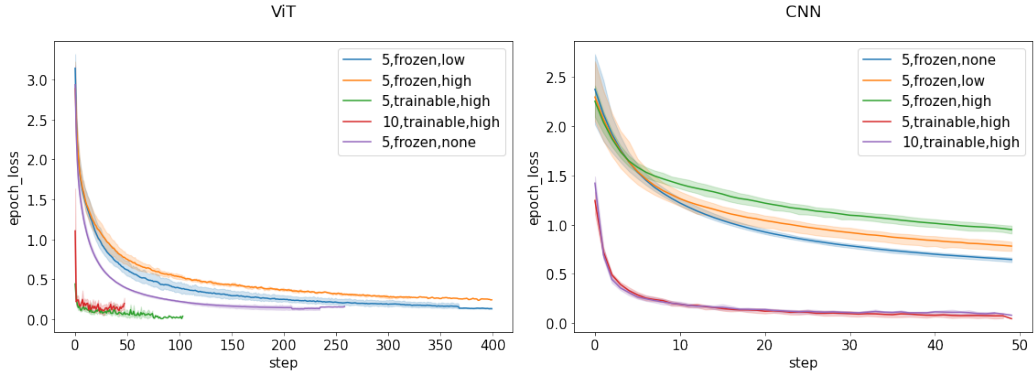

Figure 2: Categorical crossentropy of the validation set during training.

5.3 Classification accuracy using an independent test set

To test the accuracy of the models on an unseen set of images, independent from the training data, a collection of cloud images was compiled using online sources and our own photos. The test set contains 163 images manually sorted into the 6 classes. For each model, the iteration of the 5-fold cross validation with the best performance (minimum loss) was selected and evaluated. The resulting loss and accuracy scores are shown in Table 3.

The best ViT model ($77.3\,\%$) outperforms the best CNN model ($74.2\,\%$) in terms of test accuracy. According to our results considering categorical accuracy, loss and computational performance during training and validation, the best architecture for our purpose is $vit\_tiny\_r\_s16\_p8\_224\_in21$ with high level of augmentation.

On the test set, high augmentation models strongly outperform low augmentation ones, showing the importance of training data variability.

6 Conclusion and discussion

This group work proved the applicability and effectiveness of Transformer for image classification tasks. We compared CNN and ViT models with our cloud image dataset. Meanwhile, we have applied different levels of data augmentation to improve model accuracy and used K-fold cross validation to compensate the size of our dataset and to compare the performance of models at scale.

However, this study still has some limitations. As indicated, the training time for the ViT model is extremely long compared to the CNN models. Therefore, to implement in real-life applications its efficiency must be improved, which needs to be investigated. This work also only studies one type of dataset with limited number of CNN and Transformer models. Other CNN models such as VGGs and Inception-ResNets could be further studied to compare the performance between CNN and ViT models.

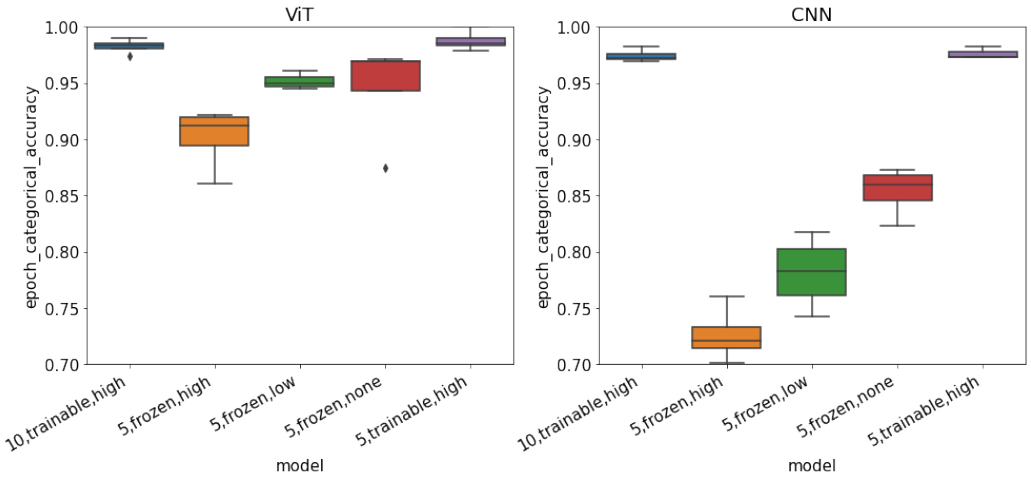

Figure 3: Maximum categorical accuracy of the validation set during training

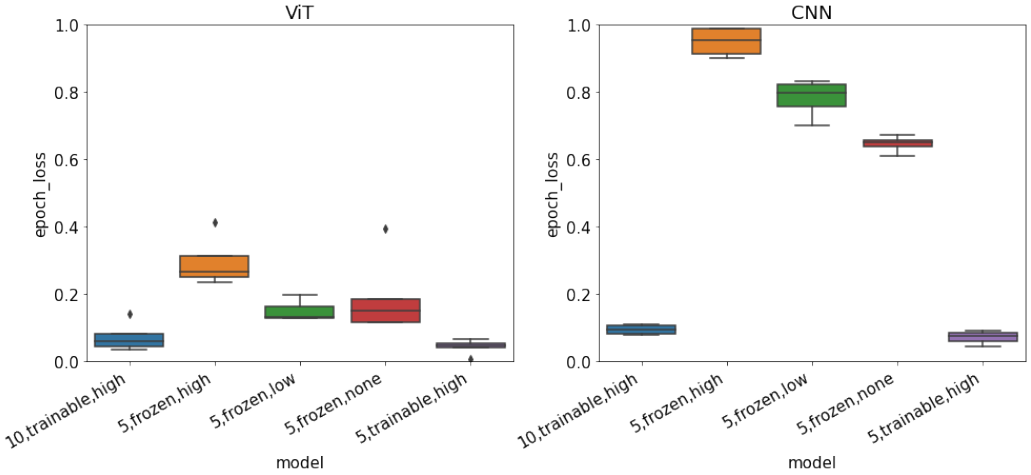

Figure 4: Minimum categorical crossentropy of the validation set during training

References

[1] Yakoub Bazi, Laila Bashmal, Mohamad M. Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3), 2021. ISSN 2072-4292. doi: 10.3390/rs13030516. URL `https://www.mdpi.com/2072-4292/13/3/516`. Accessed: 09/11/2022.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[5] Davood Karimi, Serge Didenko Vasylechko, and Ali Gholipour. Convolution-free medical image segmentation using transformers. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas

Table 3: Comparing the performance of the models using an independent test set.

model	accuracy [%]	loss
vit_10m_ha_t	77	1.52
cnn_5m_ha_t	74	1.12
vit_5m_ha_t	72	1.67
cnn_10m_ha_t	63	1.20
vit_5m_ha_f	56	1.60
vit_5m_la_f	50	2.56
vit_5m_na_f	43	3.42
cnn_5m_na_f	42	1.60
cnn_5m_ha_f	39	1.60
cnn_5m_la_f	37	1.61

Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 78–88, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87193-2.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[7] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

[8] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 05 2019. Accessed: 09/11/2022.

[9] Vinh Truong Hoang. Swimcat-ext, 2020. URL `https://data.mendeley.com/datasets/vwdd9grvdp/1`. Accessed: 10/12/2022.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

[11] Ricardo Vilalta, Christophe Giraud-Carrier, Pavel Brazdil, and Carlos Soares. *Inductive Transfer*, pages 545–548. 01 2011. doi: 10.1007/978-0-387-30164-8_401.

[12] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.