Comparing Vision Transformers and Convolutional Nets for Safety Critical Systems

Michał Filipiuk, Vasu Singh
NVIDIA, Munich, Germany
{mfilipiuk,vasus}@nvidia.com

Abstract

Transformer based architectures like vision transformers (ViTs) are improving the state-of-the-art established by convolutional neural networks (CNNs) for computer vision tasks. Recent research shows that ViTs learn differently than CNNs, that provides an appealing choice to developers of safety-critical applications for redundant design. Moreover, ViTs have been shown to be robust to image perturbations. In this position paper, we analyze the properties of ViTs and compare them to CNNs. We create an ensemble of a CNN and a ViT and compare its performance to individual models. On the ImageNet benchmark, the ensemble shows minor improvements in accuracy relative to individual models. On the image corruption benchmark ImageNet-C, the ensemble shows up to 10% improvement over the individual models, and generally performs as well as better of the two individual networks.

Introduction

Machine learning plays an important role in computer vision applications. Safety critical applications like autonomous vehicles and robotics increasingly depend on machine learning for computer vision. Deep neural networks (LeCun, Bengio, and Hinton 2015) based on Convolutional Neural Networks (CNN) are well-known and widely used for their powerful representation. For instance, in the field of autonomous driving, various CNN models have been used for object detection and image segmentation algorithms that serve as perception units to process camera (e.g., PilotNet (Bojarski et al. 2016), Fast RCNN (Wang, Shrivastava, and Gupta 2017)) and Lidar (e.g., VoxelNet (Zhou and Tuzel 2018)) data. However, the success of CNNs comes at the cost of restricting the computation to leverage data limited spatially by using convolutional layers. The performance of CNNs has gradually saturated, and the ML research community has been exploring alternative architectures for computer vision.

One of these alternative architectures that challenges the dominance of CNNs is based on the transformer model (Vaswani et al. 2017). First proposed for natural language processing tasks, transformers have been adapted for computer vision tasks by different approaches. For example, the vision transformer model (ViT) (Dosovitskiy et al. 2020) uses self-attention layer instead of convolution layers, effectively removing the spatial inductive bias introduced by the convolution operation and enables the network to use full image data to its advantage.

The development of safety critical systems relies on stringent safety methodologies, designs, and analyses to prevent hazards during operation. Automotive safety standards like ISO26262 (International Standards Organization 2018-12) and ISO/PAS 21448 (International Standards Organization 2019-01) mandate methodologies for system, hardware, and software development for automotive systems. Furthermore, these standards have been extended with best practices to use machine learning based components in safety critical systems. Ashmore et al. (Ashmore, Calinescu, and Paterson 2019) describe the ML safety lifecycle that establishes best practices across the ML development cycle from data management, model selection, training to deployment. Similarly, the industry whitepaper titled *Safety First for Automated Driving* (Aptiv et al. 2019) specifies techniques for developing, deploying, and monitoring neural networks for safety critical systems. In general, these guidelines recommend to identify common causes of failures, avoid overfitting to training data, quantify uncertainty in prediction, and make networks robust to natural perturbations.

Based on these proposals for using machine learning in safety critical systems, we investigate the behavior of ViTs in comparison and conjunction with CNNs. We explore whether CNNs and ViTs could be combined into an ensemble (Dietterich 2000) for better accuracy. The fact that ViTs are based on a different architecture than CNNs is valuable for developing safety-critical systems, as this allows to reason about two independent network models, one based on convolution and the other on self-attention. We also investigate the robustness of the ensemble compared to individual networks. We consider the ImageNet-C benchmark (Hendrycks and Dietterich 2019) to create perturbations like gaussian noise, defocus blur, artificially added fog, and lowered image contrast. The ensemble shows an improvement of up to 10% compared to individual networks on these robustness benchmarks.

The paper is organized as follows. Section 2 describes the properties of vision transformers, motivating their use

in safety-critical applications and comparing them to CNNs. Section 3 provides a quantitative analysis of CNNs and ViT (and their ensemble) on image classification. Section 4 discusses related work. Section 5 concludes the paper with a summary of our ongoing work and future directions.

ViTs for Safety

The original transformer model (Vaswani et al. 2017) for natural language processing takes a sequence of one-dimensional token embeddings as input, and relies on self-attention to capture long-range data dependencies. Transformers have become the dominant network architecture for natural language tasks. A straightforward application of the transformer model to computer vision tasks would require attention between every pair of pixels - this does not scale to realistic image sizes due to quadratic cost in the number of pixels. The ViT model (Dosovitskiy et al. 2020) avoids this limitation by reshaping an image into a sequence of flattened patches of size $P \times P$, reducing the effective sequence input length P^2 times. Generally, the patch size P is chosen to be 16 or 32. We now describe desirable properties of neural network architectures for use in safety-critical applications.

Reusability. Transfer learning (Pan and Yang 2009) is a commonly used technique for training ML models, where a model is trained for a particular context, and then re-used in a different context with limited training data. It is a powerful technique that reduces computational effort and increases confidence in the trained model. CNNs are well suited for transfer learning since convolutional layers allow to encode features in the input space. It has been shown (Dosovitskiy et al. 2020) that ViTs also attain excellent results when they are trained at sufficient scale and then transferred to new tasks with relatively fewer datapoints. Especially when pre-training data is in abundance and transfer data is scarce (few-shot learning), ViTs outperform state-of-the-art CNNs.

Robustness. For use in safety-critical applications, it is important that the network is robust against image perturbations. For example, an automotive perception network trained under sunny weather conditions should perform well also in rainy and snowing situations. To simulate such effects, several image corruption benchmarks (Hendrycks and Dietterich 2019), (Michaelis et al. 2019) have been created and the performance of different network architecture studied. Bhojanapalli et al. (Bhojanapalli et al. 2021) investigate the performance of ViTs and CNNs in images with corruptions like noise and blur. They demonstrate that with a significant size of the pre-training dataset, ViTs are at least as robust as CNNs, and sometimes more robust on artificially corrupted data. Similarly, Naseer et al. (Naseer et al. 2021) show that ViTs are robust against severe occlusions of foreground objects and random patch locations compared to state-of-the-art CNNs.

Detection of Distribution Shift. In addition to robustness against image perturbations, it is also essential that the network can identify distribution shift during deployment, i.e. scenarios where the network is observing data that is different from its training data - this is because unseen data might result in incorrect predictions. For example, an automotive perception network trained to detect pedestrians should be able to distinguish cyclists as being different from pedestrians. Fort et al. (Fort, Ren, and Lakshminarayanan 2021) show that pre-trained transformers perform better in detecting out-of-distribution (OOD) samples than CNNs. Also, transformers are better suited to few-shot outlier exposure than CNNs, where a network is shown a few outlier samples in order to improve distribution shift.

Redundancy. Self-attention allows ViTs to integrate global information about the image even in the lower layers. ViTs have more uniform representation across layers, preserving input spatial information. This is contrary to CNNs where global information is available only in higher layers. Moreover, CNNs have an intrinstic local neighborhood structure in each layer. The translational invariance in CNNs also introduces an inductive bias. Raghu et al. (Raghu et al. 2021) show that ViTs and CNNs indeed learn differently, and this is reflected in their internal structures after training. This fundamental difference in how ViTs and CNNs learn provides a powerful tool for redundant design of safety critical applications. In addition, independent models can be pre-trained on different datasets and executed on different hardware platforms at runtime. An ensemble of CNN and ViT can argue the safety based on the fact that the two individual architectures differ in their detection mechanism (convolution and self-attention respectively) and thus the ensemble does not suffer from common-cause failures like an ensemble of multiple CNNs or multiple ViTs would. For example, Bhojanapalli et al. (Bhojanapalli et al. 2021) show that adversarial perturbations do not transfer across ViTs and CNNs.

Model	**Top-1**	**Top-5**	**Top-10**
CNN	0.84788	0.97258	0.98678
ViT	0.85152	0.97412	0.98762
CNN + ViT	**0.86710**	**0.98128**	**0.99198**

Table 1: Comparison of accuracy for CNN and ViT on ImageNet benchmark

Quantitative Analysis

We start with an investigation whether CNNs and ViTs can be combined for more accurate detection. We compare the accuracy of the individual architecture with the ensemble model.

We choose the Vision Transformer(ViT) (Dosovitskiy et al. 2020) and Big Transfer(BiT) (Kolesnikov et al. 2019) models for our comparsion. The specific ViT that we picked for our experiments is the largest, publicly available ViT-L using patches of 16x16 pixels, pretrained at ImageNet21K and fine-tuned to ImageNet2012 images at resolution of 384x384[1]. It consists of 307M trainable parameters. Big Transfer model is a CNN architecture based on well-known

[1]available here: gs://vit_models/augreg/L_16-i21k-300ep-lr_0.001-aug_strong1-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_384.npz

(a) Correct label: Mountain bike

(b) Correct label: Remote control

Figure 1: Sample images from ImageNet where the ensemble provides the correct classification and the individual networks do not

ResNet networks with a few improvements, enabling it to be transferable between the datasets similarly to ViTs. Here, we also picked the biggest available checkpoint called BiT-M, based on ResNet152x4, pretrained on ImageNet21K and fine-tuned to ImageNet2012[2]. BiT consists of 937M parameters. Using these two models, we create an ensemble model. It combines the output of both CNN and ViT, and treats them as individual probabilities distribution over all 1000 ImageNet classes. The distributions are then multiplied element-wise to obtain values proportional to each class' likelihood. Such approach is just one of many possible techniques to combine the results of models' inferences. We plan to research other ensemble models in future work.

Table 1 shows the Top-1, Top-5, and Top-10 accuracy of the CNN, ViT, and the ensemble. CNN and ViT perform very similarly, with ViT being consistently slightly better, while an ensemble performs better across all metrics, significantly at Top-1.

Next, we investigate examples where the ensemble of CNN and ViT predicts the correct class (Top-1), while the individual models do not. Figure 1 shows two such examples. Figure 1a shows a bicycle handlebar where a large part of the bike is outside the image. This poses a challenge to correctly comprehend the image. Figure 1b shows an old remote control for an Apple device. We see its back, what is an unusual way of presenting the remote control. It's also not common to see such remotes as most of us associate Apple and its logo with iPhones and MacBooks.

Table 2 provides the softmax probabilities for the top-5 predictions per network for these examples. We observe that both networks predict different class (top-1) for both cases, while the correct prediction appears in the top-5 of both networks. We also observe that for the remote control, the CNN predicts it as a hard disc with significantly high probability (0.48), whereas the ViT does not predict it as a hard-disc in the top-5 predictions.

Robustness. Next, we ask the question: how robust are ViTs to perturbations for image classification in comparison to CNNs and can we somehow leverage their unique ways of comprehending the image to our advantage with the ensemble?

[2]available here: https://tfhub.dev/google/bit/m-r152x4/ilsvrc2012_classification/1

Correct label for image	**CNN**	**ViT**
Mountain bike	(joystick, 0.19682)	(microphone, 0.15267)
	(mountain bike, 0.18778)	(tripod, 0.05060)
	(disk brake, 0.11440)	(stopwatch, 0.03436)
	(tripod, 0.07354)	(mountain bike, 0.02978)
	(screw, 0.06855)	(joystick, 0.02686)
Remote control	(hard disc, 0.47829)	(iPod, 0.27300)
	(remote control, 0.28019)	(packet, 0.18125)
	(pencil box, 0.04867)	(remote control, 0.15923)
	(modem, 0.01540)	(modem, 0.14748)
	(cellular telephone, 0.01256)	(hand-held computer, 0.03882)

Table 2: Top-5 predictions and probabilities for the samples in Figure 1

We continue to work with the ViT and BiT models mentioned earlier, but we choose a different checkpoint for the ViT[3], as we change the images resolution from 384x384 to 224x224. The respective BiT model in our analysis has a degraded performance as there is no checkpoint available for smaller images.

To validate the performance on the corrupted data, we have chosen ImageNet-C dataset (Hendrycks and Dietterich 2019) and selected a few corruptions: Gaussian noise, defocus blur, contrast and fog. For each corruption, we preprocessed the data with the highest level of corruption severity (sample image can be seen in Figure 2. The corruption was applied to original ImageNet images by TensorFlow dataset.[4]). We use first 10% of the ImageNet validation data (5000 images) with aforementioned corruptions added. We took pre-trained checkpoints mentioned in the section above and run inference using NVIDIA Quadro A6000 GPU.

Table 3 compares the Top-1, Top-5, and Top-10 accuracy of BiT, ViT, and their ensemble. For every corruption except Gaussian noise, the ensemble is superior to both CNN and ViT working separately, while in case of the Gaussian noise it is slightly worse. It is also interesting how the individual models perform on various corruptions: ViT is much better

[3]available here: gs://vit_models/augreg/L_16-i21k-300ep-lr_0.001-aug_strong1-wd_0.1-do_0.0-sd_0.0--imagenet2012-steps_20k-lr_0.01-res_224.npz

[4]https://www.tensorflow.org/datasets/catalog/imagenet2012_corrupted

at Gaussian noise and fog, while it seems to be on par with CNN on defocus blur, and performs worse in the contrast corruption.

Related work

Since the introduction of ViTs in 2020, their properties have been extensively studied. Naseer et al. (Naseer et al. 2021) observe that ViTs are resilient to domain shifts and occlusions. Bhojanapalli et al. (Bhojanapalli et al. 2021) show robustness of ViTs against adversarial and natural perturbations. Fort et al. (Fort, Ren, and Lakshminarayanan 2021) study the performance of vision transformers on out-of-distribution detection. Ranftl et al. (Ranftl, Bochkovskiy, and Koltun 2021) use vision transformers for monocular depth estimation and semantic segmentation. Raghu et al. (Raghu et al. 2021) investigate the difference between the learning representation of ViTs and CNNs. There have also been architectures that combine CNNs and ViTs. For example, CNNs meet Transformers (CMT) (Guo et al. 2021) is an architecture where the input image is fed into a sequence of convolutional blocks for fine-grained feature extraction, followed by CMT blocks (transformer with depth-wise convolution) for representation learning.

Conclusion

We investigated how the introduction of vision transformers as an alternative to CNNs impacts network design in safety critical systems for computer vision. We compare CNNs and ViTs as well as their ensemble on image classification tasks. We also show that for many common image corruptions, ViTs are relatively more resilient than CNNs. Moreover, an ensemble of a CNN and a ViT provides up to 10% higher accuracy than individual networks on the corruptions provided in the ImageNet-C benchmark.

Ongoing and future Work. Vision transformers are an exciting development for safety-critical applications of computer vision. Not only do they learn differently from CNNs, but also are more robust against natural and adversarial perturbations. We believe that vision transformers alone or in conjunction with CNNs need to be extensively investigated to develop stronger safety guarantees of ML based components. This is ongoing work, and we continue to investigate the following:

- How does the performance of transformers compare to CNNs on object detection benchmarks with and without corruption? We plan to use the corruption datasets (Michaelis et al. 2019) corresponding to common object detection benchmarks like Pascal, Coco, Cityscapes, and compare the performance of transformer based object detection networks like Swin (Liu et al. 2021) to CNNs. We also plan to combine Swin with CNN models and use ensemble techniques for object detection (Wei, Ball, and Anderson 2018).
- How good are vision transformers for detection distribution shift after deployment? We are investigating the performance of transformers on automotive benchmarks for OOD detection (Nitsch et al. 2021) with different metrics like maximum over softmax probabilities and Mahalanobis distance. We plan to compare zero-shot versus few-shot OOD detection for different architectures.
- How to implement redundant design in resource-constrained systems? The current performant ViTs are large, and challenges exist in scaling performance to smaller models (Liu et al. 2021) that are robust as well as suitable for resource-constrained domains like automotive and robotics. In future work, we plan to address these challenges.

References

Aptiv; Audi; Baidu; BMW; Continental; Daimler; FCA; Here; Infineon; Intel; and Volkswagen. 2019. Safety First For Automated Driving. In *Safety First For Automated Driving*, 116–132.

Ashmore, R.; Calinescu, R.; and Paterson, C. 2019. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *CoRR*, abs/1905.04223.

Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; and Veit, A. 2021. Understanding Robustness of Transformers for Image Classification. arXiv:2103.14586.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Dietterich, T. G. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, 1–15. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-45014-6.

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

Fort, S.; Ren, J.; and Lakshminarayanan, B. 2021. Exploring the Limits of Out-of-Distribution Detection. arXiv:2106.03004.

Guo, J.; Han, K.; Wu, H.; Xu, C.; Tang, Y.; Xu, C.; and Wang, Y. 2021. CMT: Convolutional Neural Networks Meet Vision Transformers. arXiv:2107.06263.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations*.

International Standards Organization. 2018-12. ISO 26262: Road Vehicles - Functional Safety, Parts 1 to 11. In *Road Vehicles - Functional Safety, Second Edition*.

International Standards Organization. 2019-01. ISO/PAS 21448: Road Vehicles - Safety of the intended functionality. In *Road Vehicles - Safety of the intended functionality*.

Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; and Houlsby, N. 2019. Large Scale Learning of General Visual Representations for Transfer. *CoRR*, abs/1912.11370.

(a) Original image (b) Gaussian noise (c) Fog (d) Defocus Blur (e) Contrast

Figure 2: Sample image and four different corruptions for robustness tests

Model	Original data		Gaussian noise		Defocus blur		Contrast		Fog	
	Top-1	**Top-5**	**Top-1**	**Top-5**	**Top-1**	**Top-5**	**Top-1**	**Top-5**	**Top-1**	**Top-5**
CNN	0.7826	0.9464	0.3780	0.5914	0.3536	0.5762	0.3792	0.6076	0.4700	0.7518
ViT	0.8326	0.9684	**0.5330**	**0.7532**	0.3966	0.5852	0.1980	0.3008	0.6036	0.7744
CNN + ViT	**0.8416**	**0.9726**	0.5130	0.7522	**0.4340**	**0.6634**	**0.4010**	**0.6210**	**0.6276**	**0.8646**

Table 3: Comparison of accuracy for CNN and Vision transformers for selected ImageNet-C corruptions

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv preprint arXiv:2103.14030*.

Michaelis, C.; Mitzkus, B.; Geirhos, R.; Rusak, E.; Bringmann, O.; Ecker, A. S.; Bethge, M.; and Brendel, W. 2019. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *arXiv preprint arXiv:1907.07484*.

Naseer, M.; Ranasinghe, K.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2021. Intriguing Properties of Vision Transformers. arXiv:2105.10497.

Nitsch, J.; Itkina, M.; Senanayake, R.; Nieto, J.; Schmidt, M.; Siegwart, R.; Kochenderfer, M. J.; and Cadena, C. 2021. Out-of-Distribution Detection for Automotive Perception. arXiv:2011.01413.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do Vision Transformers See Like Convolutional Neural Networks? arXiv:2108.08810.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision Transformers for Dense Prediction. *ArXiv preprint*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, X.; Shrivastava, A.; and Gupta, A. 2017. A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*.

Wei, P.; Ball, J. E.; and Anderson, D. T. 2018. Fusion of an Ensemble of Augmented Image Detectors for Robust Object Detection. *Sensors*, 18(3): 894.

Zhou, Y.; and Tuzel, O. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*.