Are Vision Transformers Always More Robust Than Convolutional Neural Networks?

Francesco Pinto*
TVG, University of Oxford and FiveAI

Philip H.S. Torr
TVG, University of Oxford

Puneet K. Dokania
TVG, University of Oxford and FiveAI

Abstract

Since Transformer architectures have been popularised in Computer Vision, several papers started analysing their properties in terms of calibration, out-of-distribution detection and data-shift robustness. Most of these papers conclude that Transformers, due to some intrinsic properties (presumably the lack of restrictive inductive biases and the computationally intensive self-attention mechanism), outperform Convolutional Neural Networks (CNNs). In this paper we question this conclusion: we show that CNNs pre-trained on large amounts of data are expressive enough to produce superior robustness to current transformers performance. Also, in some relevant cases, CNNs, with a pre-training and fine-tuning procedure similar to the one used for transformers, exhibit competitive robustness. To fully understand this behaviour, our evidence suggests that researchers should focus on the interaction between pre-training, fine-tuning and the specific inductive biases of the considered architectures. For this reason, we present some preliminary analyses that shed some light on the impact of pre-training and fine-tuning on out-of-distribution detection and data-shift.

1 Introduction

Transformers revolutionised natural language processing [Vaswani et al., 2017], and a standard procedure has become to pre-train transformers on large scale corpora to then fine-tune them on smaller datasets [Kalyan et al., 2021]. Since Transformers were introduced in Computer Vision [Dosovitskiy et al., 2020], several transformer variants have been introduced [Yuan et al., Touvron et al., 2020, Liu et al., 2021].
The literature points out that since transformers do not rely on the restrictive inductive bias of convolutions, they can capture long-range correlations in the input from the first layers through the self-attention mechanism and hence can leverage the massive amounts of data used to train them to learn better input embeddings. Several papers analysing the robustness of Transformer-based classifiers, indeed, claim Transformers achieve superior robustness with respect to convolutional classifiers.

In this paper we: (1) empirically show that some of the figures indicating a remarkable superiority of Transformers under data-shift and for out-of-distribution detection are wrong or misleading; (2) argue that the parameter count should not be used as a proxy of the expressiveness of models (capacity) to pick pairs of models for comparisons, given such a strategy only measures efficiency (i.e. ability to provide better performance with less parameters) without accounting for the effectiveness of the training procedure on a certain dataset in finding good parameters; (3) propose a fine-tuning

*Corresponding author: francesco.pinto@eng.ox.ac.uk

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

scheme to CIFAR-10 and CIFAR-100 that is more efficient than BiT [Kolesnikov et al., 2019], producing state-of-the-art results uncertainty properties for both Transformers and CNNs; (4) suggest the interaction between data, pre-training, fine-tuning and network architecture should be analysed more in depth and that robustness differences cannot be solely attributed to architecture components like the self-attention mechanism. We also provide a preliminary analysis of the impact of fine-tuning in the Appendix.

2 Assessing the robustness of a classifier to data-shift and out-of-distribution detection

Neural classifiers typically consist in a feature extractor $h = \phi(x)$ (where $x \in \mathbb{R}^d$ is the input, $h \in \mathbb{R}^h$ is the network embedding) followed by a linear logistic regression layer implemented through a linear layer followed by a softmax layer. The output of the softmax layer represent a distribution over the labels, and is used to both select a predicted label (the fraction of correct results is measured via the accuracy metric) and to extract uncertainty scores about such prediction. However, in most application domains of interest, no ground-truth uncertainty is available to evaluate the quality of the uncertainty scores produced.

For this reason, several downstream tasks have been devised to assess the usefulness of the uncertainty scores that can be extracted from a classifier's output. In this paper, besides measuring the accuracy, we also consider the following two important evaluation settings to measure the reliability of a classifier: calibration and out-of-distribution detection. For calibration, we report the ECE [Naeini et al., 2015] and AdaECE [Mukhoti et al., 2020].
We consider the accuracy and calibration performance both when the test set is i.i.d. with respect to the training set (i.e. IND test distribution), or when the test set undergoes a very specific form of data-shift: covariate shift (i.e. the set of labels is the same of the IND distribution, but the inputs come from a different region of the input space). For OOD detection we report the AUROC and AUPR [Hendrycks and Gimpel, 2016]. Refer to Appendix A for a more extensive discussion of these metrics.

3 Discussion of existing literature

Transformers are often regarded as more robust than Convolutional classifiers In the literature analysing the robustness of Transformers, it is common to point out their superior performance with respect to convolutional classifiers and to attribute it to the lack of strong inductive biases, and the ability of the self-attention mechanism to learn better representations. As we will show, some of the figures leading to this conclusion are wrong or unreliable. For a discussion of the most common baselines, refer to Appendix B. In this paper we consider ViT [Dosovitskiy et al., 2020] variants as Transformers, and BiT [Kolesnikov et al., 2019] and ResNeXt, models [Mahajan et al., 2018, Yalniz et al., 2019] as CNN models.

The authors of [Zhang et al., 2021] define a taxonomy for some of the most popular data-shift datasets available for ImageNet and report the superior performance of DeiT[2] with respect to BiT under data-shift. However, they only consider the BiT-R50x1 architecture (the lowest capacity BiT variant) against several Transformer variants.
The authors of [Fort et al., 2021] only considers BiT-R50x1 and BiT-R103x3 against ViT-B/16 to conclude that Transformers achieve better out-of-distribution detection performance when fine-tuned on smaller datasets (CIFAR-10 and CIFAR-100 [Krizhevsky et al.]). We will show BiT classifiers can indeed achieve comparable OOD detection performance if fine-tuned with the same procedure used for transformers.
The authors of [Paul and Chen, 2021] report results for OOD detection on ImageNet-O and data-shift robustness on ImageNet-A and ImageNet-R, claiming transformers produce better performance. We will show these findings are inaccurate.
The authors of [Minderer et al., 2021], instead, perform a more extensive analysis that considers most of the relevant BiT/ViT variants. Their conclusion is that transformers are better calibrated than convolutional models, both on IND test sets and under data-shift. Our findings agree with their analysis.

[2]DeiT [Touvron et al., 2020] is a variant of ViT that can be trained directly on ImageNet-1K via distillation.

Methods	Params	Clean Data ImageNet-1K (Test) Accuracy (↑)	ECE (↓)	AdaECE (↓)	Domain-Shift ImageNet-R Accuracy (↑)	ECE (↓)	AdaECE (↓)	ImageNet-A Accuracy (↑)	ECE (↓)	AdaECE (↓)
BiT-R50x1	25M	73.98	3.54	3.52	39.75	15.55	15.55	10.97	42.91	42.91
BiT-R50x3	217M	77.95	6.55	6.50	46.38	14.69	14.69	24.01	34.50	34.50
BiT-R101x1	44M	75.87	5.10	5.10	41.74	12.23	12.23	16.52	36.52	36.52
BiT-R101x3	387M	78.14	7.67	7.67	47.02	15.77	15.77	27.24	32.86	32.85
BiT-R152x2	232M	77.86	6.59	6.54	48.01	15.41	15.41	27.25	31.95	31.95
BiT-R152x4	936M	78.24	9.28	9.26	47.62	15.29	15.29	31.16	29.52	29.52
ViTB-16	86M	78.09	1.59	1.57	43.22	5.09	5.09	24.12	23.12	23.12
IG-ResNeXt101-32x8d	88M	82.66	8.17	8.13	75.90	4.79	4.74	45.13	23.52	23.52
IG-ResNeXt101-32x16d	194M	84.17	7.55	7.52	78.96	4.51	4.42	53.47	18.86	18.94
IG-ResNeXt101-32x32d	468M	85.15	6.02	5.97	79.45	3.17	3.09	57.84	16.31	16.24
IG-ResNeXt101-32x48d	828M	85.50	6.29	6.26	79.74	3.89	3.85	60.80	14.84	14.84
SWSL-ResNet50	25M	81.12	5.46	5.43	68.22	7.53	7.53	29.88	32.07	32.07
SWSL-ResNeXt50-32x4d	25M	82.19	5.77	5.62	69.03	5.41	5.41	33.41	26.84	26.84
SWSL-ResNeXt101-32x4d	44M	83.25	5.43	5.33	72.68	4.77	4.73	41.27	24.00	24.00
SWSL-ResNeXt101-32x8d	88M	84.33	6.28	6.27	75.46	4.73	4.70	51.32	19.40	19.40
SWSL-ResNeXt101-32x16d	194M	83.46	7.45	7.42	76.25	5.25	5.25	46.11	21.87	21.87
ViTL-16	304M	84.40	1.92	1.86	61.57	3.02	3.02	47.50	11.6	11.6

Table 1: ImageNet-1K scale classifications and calibration experiments on clean data and under data-shift. In the upper part of the table, we report results for models with comparable levels of in-domain accuracy and similar training procedures. In the bottom part of the table, we report results of models regardless of their in-domain accuracy or training procedures.

The authors of [Bai et al., 2021] perform small scale experiments on ResNet-18/50 and DeiT-S/M. Their conclusion is that while CNNs and Transformers present similar adversarial robustness properties, Transformers still exhibit better distribution-shift properties and since they could not find a way to train CNNs to produce competitive results, they conjecture the performance gap is related to the self-attention mechanism.

The impact of pre-training is more important than the lack of self-attention Literature evidencing the usefulness of pre-training in order to improve out-of-distribution detection and data-shift generalization exists [Hendrycks et al., 2019a]. The authors of [Steiner et al., 2021] suggest some empirical methods to select the best pre-trained model for a downstream task (finding that high accuracy on pre-training dataset is a good predictor of fine-tuned accuracy). However, a proper understanding of how pre-training impacts on downstream task and how fine-tuning affects the performance is lacking. We provide some preliminary analyses about this topic in Appendix C.

In this paper, we find that adequate pre-training can dramatically affect robustness properties. In particular, CNNs pre-trained with (semi-)weakly supervised procedures on massive datasets [Mahajan et al., 2018, Yalniz et al., 2019] largely outperform Transformers in terms of robustness. While this comparison is unfair (due to the lack of results for transformers pre-trained with such procedures) this result is enough to conclude that CNNs are expressive enough to learn representations that are robust to data-shift. The problem is not the absence of the self-attention mechanism, but in the inefficiency of current training procedures to learn good representations with ImageNet-21K or ImageNet-1K scale datasets. This result also suggests that an understanding of why the self-attention mechanism behaves so favourably in these settings is lacking.

4 Experiments

Our experiments show that CNNs can achieve comparable reliability to Transformers for classification both at ImageNet-1K scale and on small-scale classification (CIFAR-10/CIFAR-100) when it is performed by fine-tuning a model that was pre-trained on ImageNet-1K. We do not compare models based on the number of parameters, as our experiments show this to be an unreliable proxy of the expressiveness of a model. Comparisons based on this metric can be misleading. Refer to Appendix D for further discussion of the topic. On the other hand, we believe comparisons of models that have similar in-domain accuracy to be more fair, as done in [Bai et al., 2021]. Indeed, for such models the training procedure considered has managed to find parameters that produce similar classification accuracy performance, and for similar in-domain accuracy we would like to understand which model exhibits better uncertainty and data-shift properties.

4.1 Imagenet-1K scale experiments

Accuracy and calibration on clean data and under data-shift In Table 1 we report the accuracy and calibration results for ImageNet-1K. All the inputs are first resized to have height 256 and then the central patch of size 224×224 is cropped and normalised with standard mean and variance ImageNet-1K values. We can observe that ViTB-16 has comparable in-domain accuracy performance to the BiT models, except R50x1, which shows inferior performance. ViT seems to be significantly more calibrated than neural networks on in-domain data.

As for the ImageNet-A [Hendrycks et al., 2019b] and Imagenet-R [Hendrycks et al., 2020] results, using the timm library [Wightman, 2019] checkpoints, we could not reproduce the results reported in [Paul and Chen, 2021]. Following the evaluation procedure proposed by the authors of such datasets, we find that comparing models with similar levels of in-domain accuracy, the performance of CNNs are either comparable or superior to transformers. However, similarly to what observed in [Minderer et al., 2021], ViT exhibits significantly better calibration also under data-shift.

If we compare models regardless of their in-domain accuracy and training procedure, we can observe that when CNNs are pre-trained in (semi-)weakly supervised ways CITE, they can exhibit superior robustness with respect to transformers. While this comparison is unfair, it clearly indicates that the self-attention mechanism is not a necessary component to achieve better robustness: CNNs inductive biases are sufficient. Hence, future research should focus why the self-attention mechanism seems to favour robustness under the current training procedures on ImageNet-21K and ImageNet-1K, while similar training settings yield worse robustness for CNNs. Another interesting research direction should try to understand why CNNs, even when more robust, tend to be more miscalibrated.

Out-of-distribution detection performance Firstly, we point out several evaluation choices that yield to the belief that ImageNet-O is a particularly hard dataset to discriminate from ImageNet-1K. Indeed, the authors of [Hendrycks et al., 2019b, Paul and Chen, 2021] report extremely bad OOD detection performance results using the AUPR metric. We observe that ImageNet-O contains only 2000 samples belonging to 200 classes, while the subset of ImageNet-1K taken as IND test set contains 10000 samples [3].
The AUPR is a metric that is extremely sensitive to: (1) class imbalance (the OOD dataset is 5x smaller than the IND dataset), (2) choice of the positive class. In particular, [Hendrycks et al., 2019b] arbitrarily choose to indicate OOD points as positive class. In Table 2 we observe that flipping the positive class yields dramatic changes in the AUPR. We also perform experiments rebalancing the proportion of OOD and IND samples by resampling each OOD sample 4 more times (so that both IND and OOD datasets have 10000 samples each)[4]. The application of this strategy yields to much less relevant changes when the choice of the positive class is flipped.
Across all the considered settings, the AUROC does not vary, and hence should be the privileged metric to describe OOD detection performance. These results allow us to conclude that ImageNet-O is evidently not as hard to distinguish from ImageNet as it is believed to be: the mentioned papers report solely results from the third column from the right of Table 2).
Comparing models with similar in-domain accuracy, we can observe that the performance gap between BiTs and ViTs is not dramatic, and they might be caused by slight differences in the training procedures and the inherent randomness of the process. Future research should investigate the statistical significance of this negligible difference.
It is interesting to observe that CNN models that generalise better under data-shift exhibit inferior performance in out-of-distribution detection with respect to ViT-L.

4.2 CIFAR scale experiments (results in Appendix E)

In our experiments, we fine-tune some of the BiT and ViT. We leave to future research the goal of extensively fine-tuning over all the BiT variants and the IG-ResNeXt and SWSL-ResNe(X)t models.

How we change the BiT fine-tuning recipe Differently from the BiT [Kolesnikov et al., 2019] fine-tuning recipe, we only train for 60 epochs (against approximately 103 epochs of BiT), with

[3] Refer to their code-base:`https://github.com/hendrycks/natural-adv-examples`

[4] We could rebalance them by randomly sampling 2000 out of the 10000 IND points, but this induces randomness in the metrics; we also observed that the average of this strategy coincides with the other balancing strategy we propose.

Methods	IND=1,OOD=0				IND=0,OOD=1			
	Imbalanced		**Balanced**		**Imbalanced**		**Balanced**	
	AUROC (↑)	AUPR (↑)	AUROC (↑)	AUPR (↑)	AUROC (↑)	AUPR (↑)	AUROC (↑)	AUPR (↑)
BiT-R50x1	65.17	90.15	65.17	65.81	65.17	23.30	65.17	60.13
BiT-R50x3	74.56	92.30	74.56	71.28	74.56	36.26	74.56	72.49
BiT-R101x1	70.34	91.35	70.34	68.75	70.34	28.53	70.34	66.11
BiT-R101x3	77.32	93.40	77.32	74.84	77.32	38.74	77.32	74.66
BiT-R152x2	77.46	93.51	77.46	75.23	77.46	38.24	77.46	74.43
BiT-R152x4	80.07	94.39	80.07	78.10	80.07	44.25	80.07	78.17
ViTB-16	79.89	95.26	79.89	82.30	79.89	36.77	79.89	73.77
IG-ResNeXt101-32x8d	78.21	94.59	78.21	79.73	78.21	38.20	78.21	74.40
IG-ResNeXt101-32x16d	80.81	95.32	80.81	82.16	80.81	43.80	80.81	78.04
IG-ResNeXt101-32x32d	84.05	96.21	84.05	85.29	84.05	48.32	84.05	81.11
IG-ResNeXt101-32x48d	84.25	96.06	84.25	84.37	84.25	50.76	84.25	82.18
SWSL-ResNet50	71.29	93.08	71.29	75.71	71.29	25.51	71.29	62.73
SWSL-ResNeXt50-32x4d	72.64	93.14	72.64	75.29	72.64	27.82	72.64	65.60
SWSL-ResNeXt101-32x4d	76.08	93.95	76.08	77.58	76.08	32.55	76.08	70.25
SWSL-ResNeXt101-32x8d	78.64	94.51	78.64	79.16	78.64	38.79	78.64	74.99
SWSL-ResNeXt101-32x16d	79.65	94.49	79.65	78.68	79.65	41.92	79.65	76.95
ViTL-16	90.60	97.85	90.60	91.27	90.60	64.58	90.60	88.90

Table 2: ImageNet-O: Out-of-distribution experiments

learning rate 0.001 or 0.0001 and multiplying it by 0.1 at epochs 20 and 40. We report the best performing learning rate configuration, and observe that the downstream performance can significantly vary based on such hyperparameter (shuffling the ranking among methods). We also do not use Mixup, and directly upscale the 32x32 CIFAR-10/100 images to the resolution of 224x224. We apply random flipping as augmentation. No weight decay is used.

Accuracy and calibration on clean data and data-shifted inputs In Tables 4 and 5 we report the accuracy and calibration results for CIFAR-10 and CIFAR-100 respectively. Fine-tuning from models having similar in-domain accuracy on ImageNet-1K, we make the following observations. For CIFAR-10, the CNNs can be competitive with respect to transformers in accuracy and calibration. For instance, R101x3 is competitive with ViT-B/16. Pre-training on ImageNet-21K does not significantly improve the performance of transformers, but it improves for CNNs, allowing lower capacity models to be more competitive. Yet, observe how the performance of R101x3 pretrained on ImageNet-21K is inferior to when pretrained on ImageNet-1K. Similar observations hold under data-shift (we consider the CIFAR-10-C and CIFAR-100-C datasets [Hendrycks and Dietterich, 2019]), except that the gap between Transformers and CNNs is more remarkable and the Transformers are generally better calibrated. As already observed in [Steiner et al., 2021], better in-domain training accuracy on the pre-training dataset often correlates to better accuracy on the fine-tuning dataset.

Out-of-distribution detection performance In Tables 4 and 5 we report the results for out-of-distribution detection performance using CIFAR-10/100 and SVHN as out-of-distribution datasets. On far-OOD detection (CIFAR vs SVHN) some convolutional models achieve competitive performance. On near-OOD detection (CIFAR-10 vs CIFAR-100 or vice versa), although slightly inferior, in CIFAR-10 vs CIFAR-100 CNNs are competitive, while the gap increases for CIFAR-100 vs CIFAR-10. The difference with [Fort et al., 2021] suggests that choosing the same fine-tuning procedure for both Transformers and CNNs can significantly bridge the performance gap in OOD detection.

5 Discussion and Conclusion

Our experiments show that CNNs inductive biases are enough to achieve robust performance under data-shift at ImageNet-1K scale: the self-attention mechanism is not a necessary component for this purpose. Future research should however investigate why transformers seem to exhibit better calibration with respect to CNNs. Our experiments also indicate that ViTL-16 exhibits superior robustness although its training procedure is similar to BiT models. This indicates that the training procedures used to train these models cannot fully leverage the full expressiveness of CNNs. On the other hand, for ViT, similar training procedures seem to be easily able to find optimal solutions that are also more robust if the model number of parameters is big enough. However, the interaction between the self-attention mechanism and the training procedures is yet to be understood. We leave further experiments using other transformer architectures at ImageNet-1K scale for future research.

We will also consider other forms of data-shift (e.g. ImageNet-C [Hendrycks and Dietterich, 2019], ImageNet-9 [Xiao et al., 2020]).
As for the fine-tuning experiments, we propose to use the same fine-tuning procedure for both CNNs and transformers. We show that in this case, the performance gap between the two on CIFAR-10 is negligible, while it widens on CIFAR-100. Future research should try to understand the interaction between the fine-tuning procedure, the target fine-tuning dataset and the inductive biases of the considered models. In Appendix C we report a preliminary analysis of the impact of fine-tuning. We will also leave to future research the analysis of the variance of all the metrics considered using multiple seeds. Indeed, the small discrepancies might not be statistically different.

6 Acknowledgments

Francesco Pinto is funded by the European Space Agency (ESA) (contact point: Dr. Juan Delfa). This work is funded by FiveAI and the EPSRC grant: Turing AI Fellowship: EP/W002981/1, EPSRC/MURI grant EP/N019474/1. We would also like to thank the Royal Academy of Engineering.

References

Yutong Bai, Jieru Mei, Alan Yuille, and Cihang Xie. Are transformers more robust than cnns?, 2021.

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, 2009. doi: 10.1109/CVPR.2009.5206848.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model Complexity-Uncertainty Trade-Off. March 2019.

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. October 2020.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of Out-of-Distribution detection. June 2021.

K. He, X. Zhang, S. Ren, and J. 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.

D Hendrycks, K Lee, and M Mazeika. Using pre-training can improve model robustness and uncertainty. *Conference on Machine . . .* , 2019a.

D Hendrycks, S Basart, N Mu, S Kadavath, and others. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv*, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and Out-of-Distribution examples in neural networks. October 2016.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. July 2019b.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. AMMUS : A survey of transformer-based pretrained models in natural language processing. August 2021.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. December 2019.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. March 2021.

Dhruv Mahajan, Ross B. Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. *CoRR*, abs/1805.00932, 2018. URL `http://arxiv.org/abs/1805.00932`.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL `https://www.sciencedirect.com/science/article/pii/S0079742108605368`.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. June 2021.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. In *NeurIPS*, 2020.

Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proc. Conf. AAAI Artif. Intell.*, 2015:2901–2907, January 2015.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning. September 2019.

Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. May 2021.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael Curtis Mozer. Mitigating bias in calibration error estimation. September 2020.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? data, augmentation, and regularization in vision transformers. June 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. December 2020.

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 V 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.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *ArXiv preprint arXiv:2006.09994*, 2020.

I. Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *CoRR*, abs/1905.00546, 2019. URL `http://arxiv.org/abs/1905.00546`.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E H Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet.

Chongzhi Zhang, Mingyuan Zhang, Shanghang Zhang, Daisheng Jin, Qiang Zhou, Zhongang Cai, Haiyu Zhao, Shuai Yi, Xianglong Liu, and Ziwei Liu. Delving deep into the generalization of vision transformers under distribution shifts. June 2021.

A Evaluation Metrics Details

Calibration metrics Calibration is often defined via the Expected Calibration Error: $\mathbb{E}_{\hat{p}}[|P(\hat{y} = y|\hat{p} = p) - p|]$, i.e. the mismatch between probability that the model's predicted label $\hat{y}$ is actually y given that the model's confidence $\hat{p}$ (i.e. maximum probability of the softmax layer) takes the value of p and p itself. Informally, it's the expectation of the absolute difference between the model's accuracy and confidence. For instance, a classifier is calibrated if, whenever its prediction's confidence is 90%, it is correct 90% of the times. The theoretical Expected Calibration Error is often estimated by splitting the confidence range into M equally spaced bins and taking the weighted average of the bins' accuracy/confidence mismatch (i.e. $ECE \approx \sum_{m=1}^{M} |B_m||A(B_m) - C(B_m)|/n$, where n is the number of samples, $A(B_m)$ and $C(B_m)$ are the accuracy and the average confidence of the points in the bin B_m respectively [Naeini et al., 2015]. To alleviate the bias of this estimator [Ding et al., 2019], adaptive binning techniques that induce the same number of samples in each bin (reducing the impact of varying density of confidences on the estimator) have been proposed [Roelofs et al., 2020, Nixon et al., 2019, Mukhoti et al., 2020]. For this reason, we will also estimate this quantity via the Adaptive ECE (AdaECE) [Mukhoti et al., 2020].

Out-of-distribution detection metrics Defined as the ability of the classifier to correctly detect samples whose class is not among the set of classes it has been trained on [Hendrycks and Gimpel, 2016]. The evaluation procedure takes the test set corresponding to the training set (considered as the in-didstribution (IND) set) and the test set of another dataset (the out-of-distribution (OOD) set) with no label overlapping with the in-domain test set. Both sets are fed to the classifier, for each point an uncertainty score is computed on top of the softmax output, and a binary thresholding classifier is used to distinguish between the IND and OOD set. Since the choice of the threshold depends on the risk exposure desired for a certain application, a general evaluation procedure considers all the risk thresholds by measuring the Area Under Receiver Operating Characteristic curve (AUROC) or the Area Under Precision-Recall curve (AUPR), which respectively plot the true positive and false positive rates or the precision and recall as the decision threshold is varied.

B Further literature discussion

Transformers and Convolutional Classifiers In this paper we consider a popular transformer architecture, whose robustness has been extensively analysed in the literature: ViT [Dosovitskiy et al., 2020]. ViT [Dosovitskiy et al., 2020] represents the first adaptation of the original Transformer architecture [Vaswani et al., 2017] to the task of Image Classification; to achieve ImageNet-1K [Deng et al., 2009] classification performance comparable with state-of-the-art convolutional classifiers, the model needs to be pre-trained on large-scale datasets first (for what we are concerned, ImageNet-21K [Ridnik et al., 2021]).

The most common convolutional baseline used in Transformer papers is BiT [Kolesnikov et al., 2019], a ResNet [He et al., 2016] variant that has been shown to achieve state-of-the-art accuracy on ImageNet classification and that, with an appropriate fine-tuning procedure, transfers well to many other datasets.

Several variants of both ViT and BiT exist. We will consider the following transformer variants ViT-B/16 and ViT-L/16[5], where B and L indicate the capacity (B = Base, L = Large), while 16 or 32 indicates the token patch size.The most popular BiT variants are BiT-R50x1, BiT-R50x3, BiT-R101x1, BiT-R101x3 and BiT-R152x2, BiT-R152x4 (where R50/101/152 indicates the ResNet variant, and the multiplicative factor scales the number of channels). For our experiments we use the checkpoints available in the timm library: `https://fastai.github.io/timmdocs/`

C A preliminary analysis of fine-tuning

Fine-tuning might focus on features that do not generalise Let us consider BiT-R101x3 and ViT-L/16 fine-tuned on CIFAR-10 as described before. We take snapshots before starting the fine-tuning (Epoch 0) and for the first 2 epochs of fine-tuning. We aim to understand how the features

[5]We omit ViT-B/32 ViT-L/32 since they always underperform with respect to ViT-B/16 and ViT-L/16 [Paul and Chen, 2021]

		ImageNet1K			ImageNet1K $\downarrow 32 \uparrow 224$			CIFAR10		
Epochs		0	1	2	0	1	2	0	1	2
ResNetv2-101x3	Acc	97.15	95.74	96.14	78.97	88.15	87.67	65.23	98.12	98.49
	ECE	9.85	1.57	1.27	14.67	5.16	7.23	10.92	0.57	0.67
	AdaECE	10.00	1.85	1.20	14.56	5.14	7.23	10.99	0.38	0.64
ViT-L/16	Acc	97.95	96.50	96.68	87.72	91.04	92.60	75.57	98.41	98.97
	ECE	10.90	1.31	1.44	28.10	4.18	4.00	24.36	0.58	0.46
	AdaECE	10.90	1.13	1.30	28.10	4.18	3.99	24.36	0.55	0.37

Table 3: Fine-tuning accuracy and calibration over the first two epochs of training for ImageNet1K and ImageNet1K $\downarrow$ 32 $\uparrow$ 224 (both restricted to classes that can be mapped to CIFAR-10) and CIFAR-10

of ImageNet-1K and CIFAR-10 change over the fine-tuning. For a fair comparison, we discard all examples in ImageNet-1K whose class does not map to a class contained in CIFAR-10 in the ImageNet hierarchy. We also discard the class "deer" in CIFAR-10, which is not present in ImageNet. We also consider the effect of the corruption induced by upscaling CIFAR-10 from 32x32 to 224x224, by simulating a similar form of corruption by downscaling ImageNet-1K images to 32x32 and then back to 224x224 (we call this evaluation setting ImageNet-1K $\downarrow$ 32 $\uparrow$ 224). For Epoch 0, we feed the three datasets to the pre-trained networks, and map the 1000-dimensional output to the CIFAR-10 labels set to measure Accuracy and Calibration. For Epochs 1 and 2, the head of the pre-trained networks has been replaced by a 10-dimensional output. We measure the Accuracy and Calibration on this output. As it can be seen from Table 3, both Transformers and CNNs start with a similarly high accuracy on ImageNet1K (97.95 and 97.15 respectively, making Transformers better), but CNNs exhibit better calibration. On ImageNet-1K $\downarrow$ 32 $\uparrow$ 224 the accuracy drops and the calibration gets worse for both, with Transformers having better accuracy, and CNNs having better calibration. On CIFAR-10, both models have worse accuracy compared to the one they have ImageNet1K$\downarrow$ 32 $\uparrow$ 224. Also in this case transformers have better accuracy and worse calibration. After the first epoch of fine-tuning, the ImageNet1K and ImageNet1K $\downarrow$ 32 $\uparrow$ 224 accuracies drop slightly (with transformers having better accuracy and calibration) while the CIFAR-10 accuracy and calibration improve dramatically. Given the performance on CIFAR-10 and ImageNet1K $\downarrow$ 32 $\uparrow$ 224 is first better on the latter and then on the second, with a significant gap, we can conjecture the fine-tuning procedure is specialising on features specific to CIFAR-10, that do not generalise well to ImageNet. Future analyses should understand whether the set of correct predictions at Epoch 0 is a subset of the set of correct predictions at Epoch 1 (hence indicating the network does not forget [McCloskey and Cohen, 1989] the pre-trained features). Otherwise, the improvements observed on ImageNet1K $\downarrow$ 32 $\uparrow$ 224 might be due to the forgetting of more general features and to the learning of CIFAR-10 specific feautures. We leave to future research a more extensive investigation of the phenomenon at a finer resolution, using different architectures and fine-tuning strategies.

Where OOD inputs are mapped while fine-tuning? In this paragraph we describe some insights we derived by measuring the distance of the embeddings of CIFAR-10, CIFAR-100 and SVHN datasets from 0. We consider pre-trained checkpoints of ViT-L/16 fine-tuned on ImageNet-1K and then fine-tuned on CIFAR-10 for 1 epoch. Consider Figure 1. At Epoch 0 (left-most) all three datasets are projected very close to 0 and the datasets are not distinguishable from the t-SNE plot. The CIFAR-10 classes are not distinguisheable from the t-SNE, given the high accuracy in Table 3 we can conjecture the class separation is disturbed by the presence of the other datasets somehow, but is visible to the classifier. At Epoch 1 (middle) all three datasets are now projected much further away from the 0, and the t-SNE plot shows the SVHN dataset is clearly distinguishable form the other two, while some overlapping exists between CIFAR10 and CIFAR100. In CIFAR100, class clusters are not visible, while they are in CIFAR10.

The fact CIFAR-10 embeddings lie much closer to 0 at Epoch 0 and the classification accuracy on it is still 97.15, suggests that the scale of the regions of the embedding space where IND/OOD points are mapped for a ViT checkpoint pre-trained on ImageNet is much smaller than the one for a fine-tuned checkpoint. This phenomenon requires further investigation also in the case of CIFAR-100 and CNNs, as this increased scale is a clear sign that the embedding space is undergoing a drastic change.

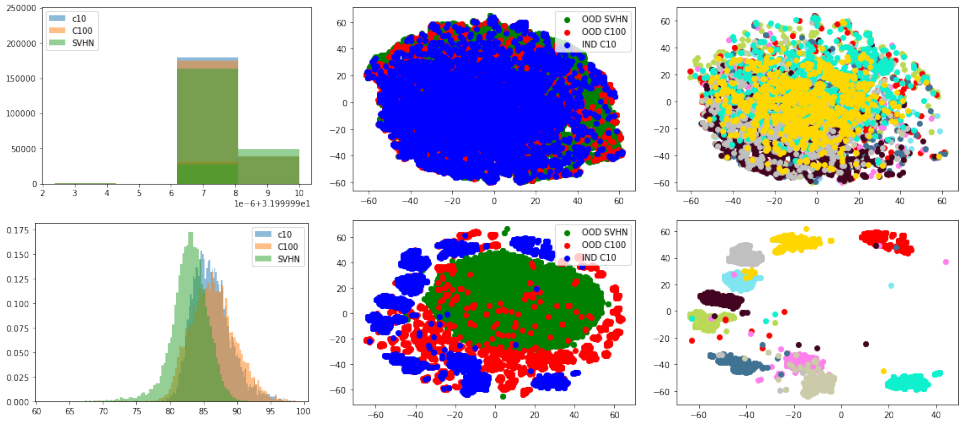

Figure 1: **First row: Epoch 0. Second row: Epoch 1**. From left to right, in the first column we report the histogram of the distances of the embeddings of CIFAR-10,CIFAR-100 and SVHN from the 0 of the embedding space. In the second column we report a t-SNE projection containing all three datasets. In the third column, we report the t-SNE projection reported in the second column, but removing the dots associated with CIFAR-100 and SVHN, and assigning to each CIFAR-10 class a different colour.

D Should the number of parameters of a model be considered when picking pairs of models to compare?

We would like to point out that, although assuming the number of parameters of a model to be a proxy of its capacity is a crude but often convenient approximation, it can be misleading. For this reason, the choice of performing a comparison between Transformer and Convolutional architectures (as done in [Paul and Chen, 2021]) solely based on the parameter count should be discouraged, as this comparison would only capture the efficiency of the model (i.e. providing better performance with less parameters)[6]. An obvious example of why we do not pick paris of models for comparison based on the parameter count can be given by comparing ViT-B/16 and ViT-L/16 to their respective ViT-B/32 and ViT-L/32 variants. The latter both contain almost 2 million parameters more than the former ones, yet they always dramatically underperform with respect to the former in all cases considered in [Paul and Chen, 2021]. The parameter count does not fully represent the expressiveness of the model and how easily the optimisation procedure can find good solutions for its parameters. For this reason, we do not perform a comparison between CNNs and Transformers based on the number of their parameters.

E Additional results

For space reasons, here we report Tables 4 and 5.

[6]It is evident that Transformers often are more efficient than CNNs.

	Clean Data			Domain-Shift			Out-of-Distribution			
	CIFAR10 (Test)			CIFAR10-C			CIFAR100		SVHN	
Methods	**Accuracy** (↑)	**ECE** (↓)	**AdaECE** (↓)(↑)	**Accuracy** (↑)	**ECE** (↓)	**AdaECE** (↓)(↑)	AUROC (↑)	AUPR (↑)	AUROC (↑)	AUPR (↑)
R50x1-1K-0.001	97.62	1.73	1.69	85.75	10.17	10.16	95.90	95.69	97.42	98.79
R50x3-1K-0.001	98.62	1.00	0.97	89.22	8.09	8.08	97.07	97.14	98.55	99.31
R101x1-1K-0.001	98.14	1.44	1.43	88.41	8.41	8.38	96.88	97.04	97.29	98.77
R101x3-1K-0.001	98.37	0.96	0.96	91.28	6.38	6.36	97.87	97.99	98.94	99.52
ViT-B/16-1K-0.001	98.91	0.77	0.75	92.48	5.23	5.21	97.92	97.90	99.35	99.71
ViT-L/16-1K-0.0001	99.14	0.50	0.50	94.64	2.93	2.90	98.84	98.88	99.80	99.91
R50x3-21k-0.001	98.59	1.04	1.01	87.86	8.70	8.68	97.44	97.51	99.19	99.64
R101x1-21k-0.001	98.45	1.16	1.13	88.97	7.74	7.72	97.22	97.32	98.90	99.51
R101x3-21k-0.001	97.91	1.51	1.45	84.55	10.86	10.84	96.19	95.94	98.41	99.27
R152x2-21k-0.001	98.74	0.91	0.91	90.90	5.88	5.86	97.97	98.03	99.03	99.52
ViT-B/16-21k-0.001	98.82	0.90	0.89	92.37	5.39	5.37	97.87	97.88	99.31	99.68
ViT-L/16-21k-0.001	99.24	0.51	0.51	95.07	3.31	3.29	98.73	98.87	99.76	99.90

Table 4: CIFAR-10 Experiments. To each method name we append whether the fine-tuned checkpoint starts from a ImageNet-1K (1K) or ImageNet-21K (21K). After we append the best learning rate. We observe that the other considered learning rate has often dramatically different performance. We will also consider more BiT variants in future research.

	Clean Data			Domain-Shift			Out-of-Distribution			
	CIFAR100 (Test)			CIFAR100-C			CIFAR10		SVHN	
Methods	**Accuracy** (↑)	**ECE** (↓)	**AdaECE** (↓)(↑)	**Accuracy** (↑)	**ECE** (↓)	**AdaECE** (↓)(↑)	AUROC (↑)	AUPR (↑)	AUROC (↑)	AUPR (↑)
R50x1-1K-0.0001	86.22	6.28	6.13	62.23	18.70	18.68	85.30	85.45	86.38	93.87
R50x3-1K-0.001	90.89	6.05	6.02	69.83	18.82	18.81	90.90	90.80	91.33	95.88
R101x1-1K-0.001	90.43	6.09	6.09	68.86	19.14	19.13	89.77	89.63	88.51	94.69
R101x3-1K-0.001	92.25	5.04	5.02	73.74	16.01	16.00	92.60	92.69	90.63	95.64
ViT-B/16-1K-0.001	93.32	4.17	4.17	78.32	13.19	13.17	94.78	94.79	94.02	97.29
ViT-L/16-1K-0.001	93.98	4.00	3.94	83.38	10.62	10.61	96.63	96.81	94.80	97.77
R50x3-21k-0.0001	90.15	5.22	5.17	67.50	15.64	15.63	90.43	90.63	94.79	97.86
R101x1-21k-0.001	90.95	5.74	5.74	68.80	18.47	18.46	90.61	90.48	85.86	93.06
R101x3-21k-0.0001	91.41	4.49	4.45	70.58	14.75	14.73	92.56	92.95	94.67	97.80
R152x2-21k-0.001	92.92	4.49	4.48	75.55	14.20	14.19	93.34	93.43	94.22	97.44
ViT-B/16-21k-0.001	93.17	4.45	4.41	78.27	13.26	13.24	94.16	94.09	94.48	97.56
ViT-L/16-21k-0.001	93.44	4.14	4.13	82.58	11.25	11.24	95.66	95.90	91.39	96.07

Table 5: CIFAR-100 Experiments. To each method name we append whether the fine-tuned checkpoint starts from a ImageNet-1K (1K) or ImageNet-21K (21K). After we append the best learning rate. We observe that the other considered learning rate has often dramatically different performance. We will also consider more BiT variants in future research.