The impact of domain shift on the calibration of fine-tuned models

Jay Mohta
Department of Computer Science
North Carolina State University
jtmohta@ncsu.edu

Colin Raffel
Department of Computer Science
University of North Carolina Chapel Hill
craffel@cs.unc.edu

Abstract

Transfer learning has become a standard technique in computer vision and natural language processing thanks to the fact that it often substantially improves performance on downstream tasks. Recent work by Hendrycks et al. [12] demonstrated that using a pre-trained model can also significantly improve a model's calibration, i.e. how well the model's confidence estimates correspond to the probability of its prediction being correct. In this paper, we provide some nuance to the claim that pre-training improves calibration by demonstrating that this beneficial effect diminishes when there is a domain shift between the pre-training and fine-tuning tasks.

1 Introduction

Transfer learning has become a central technique in the application of deep neural networks in both computer vision (CV) and natural language processing (NLP). Models like BiT [14], SimCLR [2], and ResNets [9] pre-trained on ImageNet [23] have achieved state-of-the-art performance on various vision benchmarks, while NLP models like BERT [7], RoBERTa [16], and T5 [21] have demonstrated excellent performance on the GLUE [26] and SuperGLUE [25] benchmarks and beyond. Previous works have also argued that using a pre-trained model can result in faster convergence and better generalization [22, 18].

Work by Hendrycks et al. [12] points out that pre-training can also help to improve the *calibration* of the models, i.e. how well the model's confidence estimates correspond to the probability of its prediction being correct. Quantifying the calibration of a model can be invaluable in determining how much to trust its predictions. Specifically, Hendrycks et al. [12] took models pre-trained on ImageNet and fine-tuned them on CIFAR-10, CIFAR-100, and Tiny Imagenet. These downstream tasks have a close resemblance to the pre-training task and are relatively large (with at least 60000 training samples). In practice, the performance of fine-tuned models can be greatly affected by the domain shift between the pre-training task and the downstream task. Recent work by Desai and Durrett [6] shows that calibration error can be greatly affected when there is a distribution shift between the training and testing data. In this paper, we focus on the question: How does domain shift between the pre-training task and downstream task change a model's calibration? Most past work on transfer learning has evaluated these factors in terms of their downstream task performance; to the best of our knowledge, there has been limited study of their impact on model calibration.

In this work, we, therefore, undertake an empirical study of how varying amounts of domain shift between the pre-training and downstream tasks can change the fine-tuned model's calibration. Our study includes experiments on NLP with BERT [7] and RoBERTa [16] in addition to CV experiments using BiT [14] and an ImageNet-pre-trained ResNet [23, 10]. We consider downstream tasks with varying amounts of domain shift: the QQP [3], PubMed-RCT [5], XNLI-Swahili [4] and XNLI-German [4] datasets for NLP, and the CIFAR-100 and CheXpert [13] datasets for CV Experiments.

Workshop on Distribution Shifts, 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Our results show that pre-training leads to lower calibration error in no or mild domain shift, but the beneficial effect of pre-training on calibration decreases in extreme domain shift cases. This insight will help practitioners make more informed design decisions when calibration of the fine-tuned model is a major concern.

The rest of the paper is structured as follows: section 2 provides background calibration error and domain shift. section 3 describes our experimental setup and results. We conclude in section 4 with a summary of our findings and suggestions for future work.

2 Background

Our focus in this paper is on measuring the impact of *transfer learning* (i.e. starting from a pre-trained model before fine-tuning on a downstream task) on a model's *calibration error* (i.e. how well a model's confidence correlates with the probability of it being correct). The main experimental factor we consider is the domain shift, which in the context of transfer learning refers to a mismatch between the type of data used for pre-training and fine-tuning. In the following subsections, we provide some background on calibration and domain shift before delving into our empirical study.

2.1 Calibration Error

A model is considered to be well-calibrated if the confidence estimates it gives its predictions are well-aligned with the actual probability of the predictions being correct. Given X (input to the model), ground-truth output Y, and the predicted probability of the model for input X denoted by $P_{out}(\hat{Y}|X)$, a perfectly calibrated model satisfies the following condition

$$P(\hat{Y} = Y|P_{out}(\hat{Y}|X) = p) = p, \forall p \in [0, 1]$$

In practice, the above condition is impossible to evaluate directly. Instead, calibration is typically approximated by discretizing probability intervals into a fixed number of bins. Then, the predictions are assigned to bins, and calibration error is approximated by computing the difference between the fraction of samples in the bin which are classified correctly (referred to as accuracy) and the mean of probabilities in the bin (referred to as confidence). ECE can be computed as follow

$$\text{ECE} = \sum_{i=1}^{b} \frac{n_i}{N} |\text{acc}(i) - \text{conf}(i)|$$

Here n_i refers to the number of data points in bin i, N is the total number of data points in the dataset, $\text{acc}(i)$ is the accuracy of bin i, and $\text{conf}(i)$ is confidence estimates of bin i. As noted by Nixon et al. [19], comparing the calibration of different models using the expected calibration error can be conflated by the models' accuracy. Nixon et al. [19] therefore introduced a new metric for measuring the calibration error called Adaptive Calibration Error (ACE). This metric uses an adaptive binning strategy where the bin interval are spaced such that each bin contains an equal number of predictions. ACE can be computed using the equation defined below

$$\text{ACE} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} |\text{acc}(r, k) - \text{conf}(r, k)|$$

Here $\text{acc}(r, k)$ and $\text{conf}(r, k)$ are the accuracy and confidence of the adaptive calibration range r for class k, K are the total number of classes in the classification problem and R is the total number of ranges.

2.2 Domain Shift

In the context of transfer learning, domain shift means there is a mismatch between the pre-training and downstream tasks. For example, if a model was pre-trained on ImageNet [23] (which contains a diverse set of natural images) and was fine-tuned on CheXpert (a dataset of chest x-ray images) [13], it is intuitively clear that the data seen during fine-tuning is substantially different from data used for pre-training. It follows that the features learned during pre-training might not provide substantial benefits for fine-tuning.

While domain shift is often defined intuitively, some methods have been proposed to attempt to measure it quantitatively (e.g. [8, 27, 1]). For example, a simple approach proposed by Ganin et al. [8] is to train a classifier to distinguish between data from each of the datasets. If the classifier can reach perfect accuracy, there might be a dramatic domain shift; if it can only attain chance accuracy the datasets might be highly similar. Instead of relying on any quantitative approximations of domain shift, in this work, we divide domain shifts into three intuitively-defined categories: minimal, mild, and extreme.

3 Experimental Setting and Results

We now present our main experimental results to measure the impact of domain shift on the calibration of fine-tuned models. To improve the robustness of our findings, we measure calibration in many settings (across different modalities, models, and datasets). In the following subsections, we describe our experimental setup and main findings. We report ACE in the main text of the paper because of its benefits over ECE but you can also find ECE values in the Appendix.

3.1 Setup

Natural Language Processing We experimented with the BERT [7] and RoBERTa [16] pre-trained models. BERT and RoBERTa are both pre-trained Transformer encoder [24] models that can be applied to text classification and span labeling problems. A full description of these architectures and their pre-training is out of the scope of this paper; we refer to the original sources of each model for additional details [7, 16]. To study the impact of domain shift, we consider the following downstream tasks: Quora Question Pairs (QQP) [3], PubMed-RCT [5], and the Swahili and German task from XNLI [4]. All downstream task datasets have at least 100,000 samples, we report the results for 1000 samples in the main text of the paper and also provide results for 100 samples in the Appendix. We specifically chose a low data regime (< 1000 samples for training) for our experiments because He et al. [11] pointed out that the performance of the pre-trained model is similar to the performance of models trained from scratch in the high data regime. When subsampling downstream tasks, we always choose data in such a way that the resulting dataset is class-balanced. On each downstream task, we fine-tuned BERT and RoBERTa for either 40 epochs or until it reached 100% train accuracy using a default learning rate of 2e-5.

Computer Vision We used BiT [14] and a ResNet [9] pre-trained on Imagenet [23]. BiT is a collection of pre-trained ResNet models that were designed to perform well on a wide variety of downstream tasks using the same fine-tuning formula. Our Imagenet-pre-trained ResNet corresponds to an incredibly common choice for vision-based transfer learning. To study the impact of domain shift, we used CIFAR-100 [15] and the Pleural Effusion detection task from CheXpert [13]. CIFAR-100 is essentially a subset of the classes from Imagenet, constituting almost no domain shift, whereas CheXpert focuses on the entirely different task of classifying chest x-rays. As in the NLP experiments, we report the results for 1000 samples per dataset in the main text of the paper and also provide results for 100 samples in the Appendix. For BiT experiments, we follow the recommended "BiT-Hyperrule", where we use an initial learning rate of 0.003 and decay the learning rate every 100 steps (For BiT CIFAR-100 Experiments we use a batch size of 256 while for CheXpert we use a batch size of 32). For ResNet experiments, we use an initial learning rate of 0.01 and use an exponential decay learning rate schedule with $\gamma = 0.975$ and batch size of 32 for both CIFAR-100 and CheXpert experiments. For BiT and Resnet, we train until 9,000 steps or 100% training accuracy is reached. We used A6000 GPU for all the experiments conducted in the paper.

3.2 Findings

First, we confirm one of the main results of Hendrycks et al. [12]: pre-training can significantly improve a fine-tuned model's calibration. This effect was apparent across all modalities and models we considered. For example, in table 1, pre-trained BERT-Base and RoBERTa-Base models always attained a lower calibration error on QQP than BERT-Base and RoBERTa-Base models trained from scratch. We also observe that the performance (accuracy) of pre-trained models is significantly better than the performance of models trained from scratch. Under mild domain shift (PubMed-RCT) the calibration error of the pretrained model is still significantly lower than the model that was trained

	⟶ More domain shift ⟶							
	QQP		PubMedRCT		XNLI-German		XNLI-Swahili	
	Acc	ACE	Acc	ACE	Acc	ACE	Acc	ACE
BERT-Base, pre-trained	75.2	0.22	75.3	0.065	44.8	0.338	39.9	0.34
BERT-Base, not pre-trained	61.5	0.37	56.1	0.147	41.1	0.4	37.1	0.36
RoBERTa-Base, pre-trained	77.2	0.23	75.4	0.087	40.0	0.37	43.9	0.31
RoBERTa-Base, not pre-trained	64.0	0.37	55.4	0.151	39.5	0.38	38.8	0.31

Table 1: NLP models performance (Test accuracy, ACE) for varying amount of domain shift and number of samples in each dataset fixed to 1000

	⟶ More domain shift ⟶			
	CIFAR-100		CheXpert	
	Acc	ACE	Acc	ACE
BiT-M-50x1, pre-trained	46.7	0.0039	76	0.617
BiT-M-50x1, not pre-trained	12.2	0.0101	69	0.285
Resnet50, pre-trained	47.6	0.0037	77.9	0.482
Resnet50, not pre-trained	10.2	0.012	65	0.58

Table 2: CV models performance (Test accuracy, ACE) for varying amount of domain shift and number of samples in each dataset fixed to 1000

from scratch. From table 1 and table 2 it is clear that under no domain shift or mild domain shift cases we typically observe 2-3$\times$ lower ACE than models trained from scratch. However, this effect does not necessarily hold when there is a significant domain mismatch. As seen e.g. table 1, when fine-tuning on XNLI-Swahili or German, using a pre-trained BERT-Base and RoBERTa-Base only slightly improves calibration. table 2 (right-most column) when fine-tuning on CheXpert, using pre-trained BiT-M-50x1 leads to higher calibration error than the model trained from scratch. It is tempting to explain this effect by conjecturing that extreme domain shift makes the solution found during pre-training about as useful as a random initialization, but the *accuracy* of pre-trained models is often better under extreme domain shift. This supports recent findings suggesting that pre-training can be beneficial for downstream performance even when there is an extreme domain shift between the pretraining and downstream task. [20, 17, 22].

4 Conclusion

This paper investigated how domain shifts between the pretraining and downstream tasks can affect the calibration error of models. In contrast to prior work, [12], we find that extreme domain shift can decrease pre-training's downstream benefits on calibration. Our findings provide significantly more nuances to the conventional wisdom that "pre-training helps calibration". We hope that these insights help guide practitioners towards making more informed decisions when the fine-tuned model's calibration is a major concern and that our work helps prompt future work on improving the calibration of fine-tuned models under domain shift.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6430–6439, 2019.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. ArXiv, abs/2002.05709, 2020.

[3] Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. Quora question pairs. 2017.

[4] Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In EMNLP, 2018.

[5] Franck Dernoncourt and J. Lee. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. ArXiv, abs/1710.06071, 2017.

[6] Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. 2020.

[7] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.

[8] Yaroslav Ganin, E. Ustinova, Hana Ajakan, Pascal Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. ArXiv, abs/1505.07818, 2016.

[9] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[11] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4918–4927, 2019.

[12] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In ICML, 2019.

[13] Jeremy A. Irvin, Pranav Rajpurkar, M. Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, H. Marklund, Behzad Haghgoo, Robyn L. Ball, K. Shpanskaya, J. Seekins, D. Mong, S. Halabi, J. Sandberg, R. Jones, D. Larson, C. Langlotz, B. Patel, M. Lungren, and A. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In AAAI, 2019.

[14] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, J. Puigcerver, Jessica Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning. In ECCV, 2020.

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[16] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. ArXiv, abs/1907.11692, 2019.

[17] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. arXiv preprint arXiv:2103.05247, 2021.

[18] Behnam Neyshabur, H. Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? ArXiv, abs/2008.11687, 2020.

[19] Jeremy Nixon, Michael W. Dusenberry, L. Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. ArXiv, abs/1904.01685, 2019.

[20] Isabel Papadimitriou and Dan Jurafsky. Learning music helps you read: Using transfer to study linguistic structure in language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[21] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. ArXiv, abs/1910.10683, 2020.

[22] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In NeurIPS, 2019.

[23] Olga Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115:211–252, 2015.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv preprint arXiv:1706.03762, 2017.

[25] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In NeurIPS, 2019.

[26] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In 7th International Conference on Learning Representations, 2019.

[27] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3712–3722, 2018.

Appendices

	⟶ More domain shift ⟶							
	QQP		PubMedRCT		XNLI-German		XNLI-Swahili	
	Acc	ACE	Acc	ACE	Acc	ACE	Acc	ACE
BERT-Base, pre-trained	70.6	0.11	58.9	0.086	36.5	0.317	35.8	0.12
BERT-Base, not pre-trained	60.3	0.24	41.4	0.104	35.5	0.346	35.7	0.21
RoBERTa-Base, pre-trained	65.6	0.32	55.2	0.12	35.8	0.36	35	0.17
RoBERTa-Base, not pre-trained	61.8	0.31	44.3	0.117	36.1	0.358	36.6	0.17

Table 3: NLP models performance (Test accuracy, ACE) for varying amount of domain shift and number of samples in each dataset fixed to 100

	⟶ More domain shift ⟶			
	CIFAR-100		CheXpert	
	Acc	ACE	Acc	ACE
BiT-M-50x1, pre-trained	19.6	0.011	65	0.22
BiT-M-50x1, not pre-trained	3.29	0.012	55	0.18
Resnet50, pre-trained	7.1	0.011	72.4	0.288
Resnet50, not pre-trained	3.79	0.017	63	0.282

Table 4: CV models performance (Test accuracy, ACE) for varying amount of domain shift and number of samples in each dataset fixed to 100

	⟶ More domain shift ⟶							
	QQP		PubMedRCT		XNLI-German		XNLI-Swahili	
	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
BERT-Base, pre-trained	75.2	0.23	75.3	0.16	44.8	0.50	39.9	0.52
BERT-Base, not pre-trained	61.5	0.37	56.1	0.36	41.1	0.60	37.1	0.55
RoBERTa-Base, pre-trained	77.2	0.24	75.4	0.23	40	0.56	43.9	0.47
RoBERTa-Base, not pre-trained	64.0	0.36	55.4	0.37	39.5	0.58	38.8	0.47

Table 5: NLP models performance (Test accuracy, ECE) for varying amount of domain shift and number of samples in each dataset fixed to 1000

	⟶ More domain shift ⟶			
	CIFAR-100		CheXpert	
	Acc	ECE	Acc	ECE
BiT-M-50x1, pre-trained	46.7	0.229	76	0.58
BiT-M-50x1, not pre-trained	12.2	0.347	69	0.32
Resnet50, pre-trained	47.6	0.1	77.9	0.56
Resnet50, not pre-trained	10.2	0.53	65	0.62

Table 6: CV models performance (Test accuracy, ECE) for varying amount of domain shift and number of samples in each dataset fixed to 1000

	⟶ More domain shift ⟶							
	QQP		PubMedRCT		XNLI-German		XNLI-Swahili	
	Acc	ECE	Acc	ECE	Acc	ECE	Acc	ECE
BERT-Base, pre-trained	70.6	0.11	58.9	0.085	36.5	0.46	35.8	0.15
BERT-Base, not pre-trained	60.3	0.24	41.4	0.173	35.5	0.51	35.7	0.27
RoBERTa-Base, pre-trained	65.6	0.32	55.2	0.3	35.8	0.55	35	0.24
RoBERTa-Base, not pre-trained	61.8	0.31	44.3	0.29	36.1	0.51	36.6	0.22

Table 7: NLP models performance (Test accuracy, ECE) for varying amount of domain shift and number of samples in each dataset fixed to 100

	⟶ More domain shift ⟶			
	CIFAR-100		CheXpert	
	Acc	ECE	Acc	ECE
BiT-M-50x1, pre-trained	19.6	0.24	65	0.3
BiT-M-50x1, not pre-trained	3.29	0.26	55	0.26
Resnet50, pre-trained	7.1	0.33	72.4	0.35
Resnet50, not pre-trained	3.79	0.87	63	0.38

Table 8: CV models performance (Test accuracy, ECE) for varying amount of domain shift and number of samples in each dataset fixed to 100