MEMO: Test Time Robustness via Adaptation and Augmentation - A Report

Cormac Cureton
ECSE 626 Course Project
McGill University
cormac.cureton@mail.mcgill.ca

Abstract

Test time robustification seeks to increase the performance of a pretrained model when confronted with challenging inputs such as domain shifts. Marginal entropy minimization with one test point (MEMO) improves image classification robustness with a full parameter update minimizing marginal entropy. The marginal entropy is computed as the entropy of the mean output distribution with respect to a group of randomly selected image augmentations. This report finds that MEMO improves classification performance on CIFAR-10 and ImageNet variant test sets with pretrained ResNet-26 and ResNet-50 base models. Further, analysis finds that the method most improves predictions for samples with high initial predictive entropy which also tend to have high marginal entropy. When the base model is uncertain, the method can nudge the predictive distribution in the right direction but it does not tend to improve samples where the model is confidently incorrect. Finally, improvements in classification performance are found to come at a cost of increased accuracy-confidence gap, meaning model calibration becomes worse. These trade-offs mean that MEMO is a practical option to improve predictions on individual unlabelled test points when calibration is not important and increased inference cost is acceptable. Code and replication instructions are made available at: https://github.com/Cormac-C/memo-proj.

1. Introduction

A common challenge across deep learning is that the performance of models which appear to be capable during development degrades when they encounter novel inputs outside of their training distribution [14, 19]. This challenge is relevant across domains including tabular deep learning [8, 21], natural language processing [6], and computer vision [10].

In the case of computer vision, these changes in distribution may stem from changes in factors such as object pose, noise, or lighting conditions. These challenges are well acknowledged and have motivated advances in vision model training through data augmentation [3, 11], novel model architectures [17], and different training strategies [25].

However, as models continue to grow in size and pretraining becomes a more expensive undertaking, it is often impractical to discard an existing model and retrain from scratch to increase robustness. Additionally, in domains with low data availability, it may not be possible to shift the training distribution to sufficiently capture less common modes which are nevertheless meaningful at inference time.

In some cases where the domain shift is known in advance and well captured in data, it can make sense to finetune a pretrained model. This approach is very common in language models [2] but also applicable in vision models [16]. However, in contexts where test-domain data is not available or the domain shift is not known in advance finetuning is not possible; in these cases test time adaptation can be an appropriate approach.

1.1. Test time adaptation

This report concerns **m**arginal **e**ntropy **m**inimization with **o**ne test point (MEMO) [26], however there have been many other approaches to test time adaptation of image models.

The *tent* method proposed in [24] adapts the model to minimize the test entropy across a small batch of inputs thereby increasing model confidence. In contrast, MEMO is concerned with minimizing the marginal entropy with respect to augmentations of a single test point. This method only updates normalization layers rather than the model's full parameters.

Schneider et al. [23] looked to update group statistics in batch normalization operations of pretrained models to adapt them to corrupted data. This adaptation method was shown to be effective even with a single test point, meaning it is compatible with MEMO. The MEMO paper explores how the two adaptation strategies can be effectively used together [26].

Even since MEMO's introduction, test time adaptation continues to be an area of interest. This is especially true in

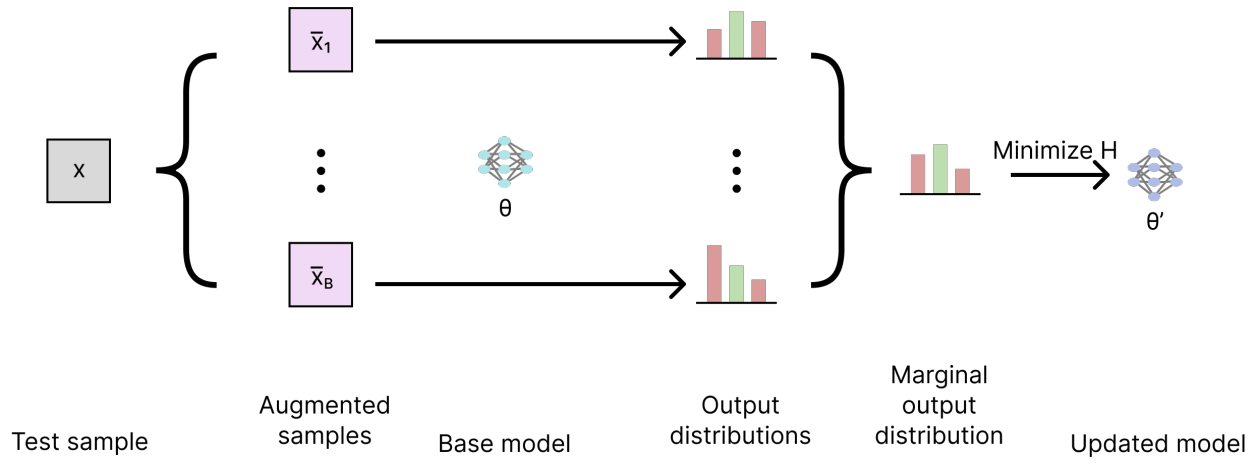


Figure 1. A diagram outlining the MEMO adaptation process. Flowing from left to right, a test sample is transformed with B augmentations, the base model makes B predictions which are combined to form the marginal output distribution. The model is then updated to minimize the marginal entropy.

the case of vision-language models (VLMs), which tend to be large and thus impractical to retrain. Recent works build on the marginal entropy minimization objective proposed in MEMO to create methods suited for VLM adaptation [4].

2. MEMO method

2.1. Test time robustification

The MEMO method addresses the problem of test time robustification, looking to improve the performance of a pretrained model using a single unlabelled test input. The robustification method replaces the normal model inference process, producing an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$. This differs from fine-tuning or many-shot methods which rely on access to input, output pairs $\{(x_i, y_i)\}_{i=1}^{N}$ to improve performance on a test input x.

The model outputs unnormalized logits $\mathbf{z}$ which are mapped via softmax to the distribution $p_\theta(y \mid x)$. In standard classification, this is treated as the conditional probability distribution of the label y given the input x and the prediction is selected via the maximum a posteriori (MAP) estimate $\hat{y} = \arg\max_y p_\theta(y \mid x)$. However, it is an important distinction that these values are often uncalibrated [5] and thus represent an estimate of confidence but not a true posterior uncertainty.

2.2. Marginal entropy minimization

Since the label y is unavailable for adaptation, the method instead looks to minimize the entropy of the model's output distribution marginalized across augmentations. The augmentations are drawn from a set $\mathcal{A}$ and transform the input x while preserving the relevant semantic content so the label is still y.

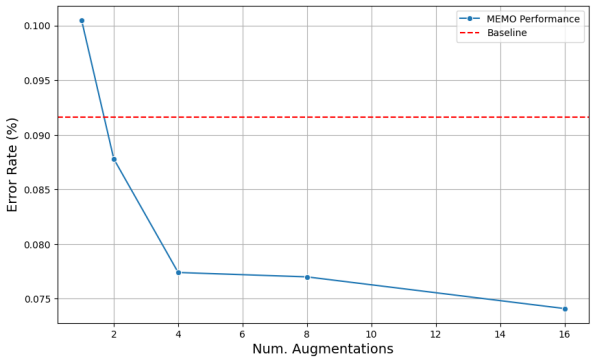
Since the full set of augmentations is impractically large, the marginal distribution is approximated via Monte Carlo sampling. B augmentations are selected from set $\mathcal{A}$ and applied to the test input x producing B augmented inputs $\{\tilde{x}_1, ..., \tilde{x}_B\}$. The model makes a prediction for each of the augmented inputs and the mean conditional distribution is then taken to be the marginal output distribution $\bar{p}_\theta(y \mid x)$, as seen in Eq. (1) from [26].

$$\bar{p}_\theta(y \mid x) \triangleq \mathbb{E}_{U(\mathcal{A})}[p_\theta(y \mid a(x))] \approx \frac{1}{B}\sum_{i=1}^{B} p_\theta(y \mid \tilde{x}_i) \quad (1)$$

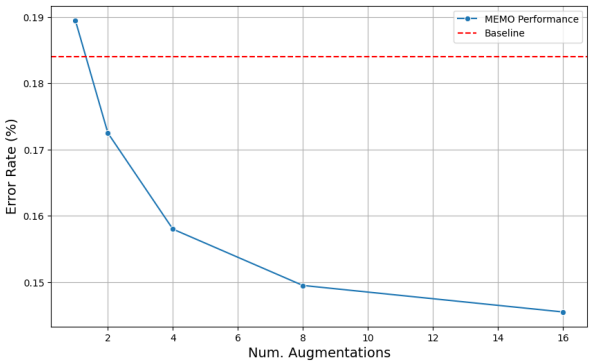
The marginal distribution's entropy is a measure of the model's uncertainty and inconsistency with respect to the selected augmentations. Since the augmentations were selected so as not to change the output label y, reducing the marginal entropy is desirable and can be interpreted as increasing the model's confidence and consistency across the augmentations. As such, MEMO robustifies the model by updating its parameters to minimize the loss function presented in Eq. (2) which is simply the marginal entropy [26].

$$l(\theta; x) \triangleq H(\bar{p}_\theta(\cdot \mid x)) = -\sum_{y \in Y} \bar{p}_\theta(y \mid x) \log \bar{p}_\theta(y \mid x) \quad (2)$$

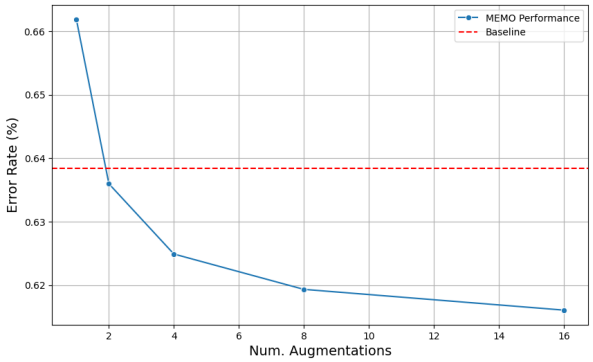
After calculation of the marginal entropy, the model's parameters are updated using a single step of stochastic gradient descent. In the original work [26], the authors explore using multiple gradient steps per test input but find that it hurts performance so all experiments in this report use a single gradient step. The update procedure is described visually in Fig. 1.



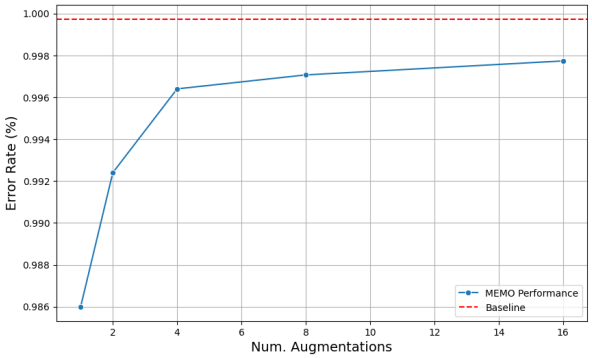
(a) CIFAR-10 error rate with different numbers of augmentations.



(b) CIFAR-10.1 error rate with different numbers of augmentations.



(c) ImageNet-R error rate with different numbers of augmentations.



(d) ImageNet-A error rate with different numbers of augmentations.

Figure 2. Comparison of the MEMO performance with different numbers of augmentations on four CIFAR-10 and ImageNet variants. In all cases, MEMO improves performance compared to the base model. Except for ImageNet-A, performance improves with more augmentations, though with diminishing returns.

3. Experiments

The following experiments examine MEMO's sensitivity to the number of augmentations used, a key hyperparameter influencing performance, as measured by test error (%). Additionally, this report characterizes the samples whose predictions are improved, worsened, or left unchanged by analyzing changes to the model's predictive entropy and the marginal entropy across augmentations. Finally, MEMO's effect on the model's calibration is measured via the expected calibration error (ECE).

3.1. Datasets

In this report, MEMO performance is evaluated using CIFAR-10 [1] and ImageNet [22] datasets and corrupted variants. In both cases, the base model is pretrained on the original dataset then the algorithm is evaluated on the corrupted variants which introduce some domain shift. These are the same datasets used for evaluation in the original work [26] although CIFAR-10-C and ImageNet-C are excluded because they contain 75 corruption-severity pairs [10] which were not feasible to evaluate within the scope of this project.

CIFAR-10. The base model used for the CIFAR-10 experiments was trained with the dataset's 50000 training samples. The method is evaluated with the original 10000 sample test set. Additionally, CIFAR-10.1 introduces 2000 new test images which are labeled according to the original CIFAR-10 definitions and thus serve as a slightly more difficult test set [20].

ImageNet. Similarly, the ImageNet base model is trained on the full training set. Evaluation uses ImageNet-Renditions (ImageNet-R), which contains 30000 test samples of the ImageNet classes depicted in different mediums (e.g. paintings), introducing several domain shifts [9]. ImageNet-A is also used, a dataset of 7500 images which are selected to be adversarial to the ResNet-50 model, though in theory they are images containing ImageNet classes [12].

3.2. Base model

All experiments use pretrained ResNet base models [7]. For the CIFAR-10 experiments, a ResNet-26 checkpoint released by the authors of the original MEMO paper [26] is

used. The ImageNet experiments use a pretrained checkpoint available from TorchVision[1].

Only ResNet models are investigated within the scope of this report due to resource and time constraints though the methodology is agnostic to the underlying base model. In the original work, the authors demonstrate that MEMO is also effective with a robust vision transformer (RVT) base model [15].

3.3. Implementation details

For model parameter adaption, this report uses basic stochastic gradient descent (SGD) as the optimizer. Within the scope of this report, the best performing learning rates and the number of parameter updates from the original work are used [26].

For all experiments a single parameter update step is used. For the ResNet-26 model, the learning rate is $5 \cdot 10^{-3}$ and for the ResNet-50 model, the learning rate is $2.5 \cdot 10^{-4}$. For image augmentations, AugMix [11] is used as it was in the original work.

I re-implement the core MEMO adaptation logic and evaluation logic in PyTorch [18] for this project. In the interest of accurate comparison, I reuse code for dataset loading, image augmentations, and loading of the pretrained base model from the MEMO repository[2]. In turn, the MEMO project used snippets from the official implementations of [9, 11, 12]. More granular attribution of copied and adapted code snippets are included inline in the repository.

3.4. Evaluation metrics

Test error (%). Classification performance is measured with test error (%), the percentage of samples where the model predicts an incorrect class, defined in Eq. (3).

$$\text{Error } (\%) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(\hat{y}_i \neq y_i) \tag{3}$$

Predictive entropy. The model's predictive uncertainty is characterized with the entropy of the model's output distribution, given in Eq. (4), where $p_\theta(y \mid x)$ are the post-softmax values. Although the output distribution is not calibrated and thus can't be interpreted as a true measure of epistemic uncertainty, changes in entropy can still be interpreted as changes in the model's confidence: lower entropy corresponds with higher-confidence, more peaked predictions and higher entropy corresponds with lower-confidence predictions.

[1] https://docs.pytorch.org/vision/0.8/models.html

[2] https://github.com/zhangmarvin/memo

	CIFAR-10		CIFAR-10.1	
	Err. %	ECE ↓	Err. %	ECE ↓
ResNet-26	9.16	**0.0442**	18.40	**0.104**
+ MEMO (B=2)	8.78	0.0799	17.25	0.158
+ MEMO (B=4)	7.74	0.0712	15.80	0.145
+ MEMO (B=16)	**7.41**	0.0690	**14.55**	0.137

	ImageNet-R		ImageNet-A	
	Err. %	ECE ↓	Err. %	ECE ↓
ResNet-50	63.84	**0.174**	99.97	**0.571**
+ MEMO (B=2)	63.60	0.448	**99.24**	0.807
+ MEMO (B=4)	62.49	0.432	99.64	0.809
+ MEMO (B=16)	**61.61**	0.420	99.77	0.812

Table 1. Error rate and expected calibration error for the base model and MEMO with varying numbers of augmentations evaluated across CIFAR-10 and ImageNet variants. Best performance metrics are indicated in **bold**.

$$H(p_\theta(y \mid x)) = -\sum_{k=1}^{K} p_\theta(y = k \mid x) \log p_\theta(y = k \mid x) \tag{4}$$

Expected calibration error (ECE). The model's calibration is quantified with expected calibration error (ECE), defined in Eq. (5) from [5]. ECE measures the gap between the model's confidence in an answer and its accuracy. The models predictions are grouped into M confidence bins and the ECE is the weighted average of the absolute accuracy-confidence difference. Lower ECE indicates better calibration, analysis looks at how MEMO adaptation impacts the alignment between confidence and accuracy. All ECE values reported in this report use $M = 15$ bins.

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{5}$$

3.5. Results

Classification performance. As seen in Tab. 1, adding MEMO adaptation does improve the classification performance across all datasets tested. For ImageNet-A, the improvement is negligible, as the base model and adapted models perform near random chance for the adversarial examples. However, for the other datasets, MEMO with two or more augmentations yields consistent improvements.

As shown in Fig. 2, increasing the number of augmentations generally improves performance. This is expected, as

a larger number of augmentations provides a better approximation of the marginal distribution. However, the improvement does plateau around eight augmentations. These observations of accuracy improvements and sensitivity to the number of augmentations are consistent with the findings of the original paper [26].

Success and failure cases. In Fig. 3, it is observed that the samples which the baseline models classify incorrectly tend to have higher predictive entropy (ie. lower confidence) than those that are correctly classified. Further, MEMO decreases the predictive entropy of all samples. This effect is consistent with the loss function defined in Eq. (2) because decreasing the predictive entropy helps minimize the marginal entropy. Tab. 2 shows that the decrease in predictive entropy is more pronounced for samples where the prediction was changed (worsened and improved) between the base model and MEMO. This aligns with the observation that samples whose predictions change tend to have higher marginal entropy which would lead to a stronger adaptation signal.

Effect on calibration. Tab. 1 also shows that the application of MEMO adaptation increases the expected calibration error of the predictions, meaning the calibration of the models becomes worse. Based on the predictive entropy decreases discussed above, it seems that the model is being adapted such that the confidence of its predictions increases faster than its accuracy. This effect is illustrated in the extreme with the ImageNet-A case. Although error remains above 99%, the ECE rises significantly, meaning the model becomes more confident despite continuing to perform poorly.

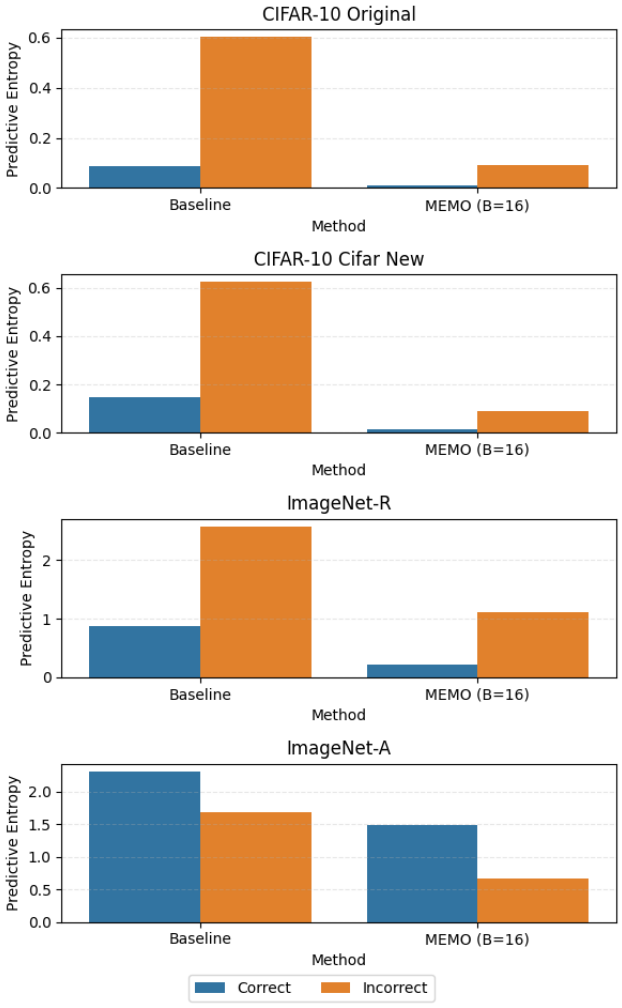


Figure 3. Changes in predictive entropy from the base model to MEMO across three datasets. MEMO decreases the entropy in all cases, correct predictions tend to have lower entropy.

4. Discussion & Conclusion

In this report, MEMO is found to reduce classification error rates on all four of the datasets tested. The method does offer a test time robustification strategy that can improve performance with only the unlabelled test point.

However, these gains come with a significant computational cost at inference. For a given sample, applying MEMO with B augmentations requires B additional forward passes and one backwards pass in addition to the cost of performing the augmentations. An additional restriction on the efficiency of this method is that the final prediction for each test sample uses a different set of adapted weights so there is not easy opportunity to batch the full pipeline. With these considerations it is clear that MEMO is specialized to the test time robustification process where only one test sample is available at a time.

Additionally, experiments suggest that MEMO is most effective in improving predictions on samples which have high initial predictive entropy (ie. low model confidence) and marginal entropy (ie. more variance across augmentations). In these cases, the MEMO adaptation succeeds by nudging the predictive distribution in a better direction. It should be noted that the failure cases of the approach have a similar character, the samples where predictions were worsened also tend to have low predictive confidence and higher variance across augmentations. However, the MEMO adaptation doesn't tend to change confident model predictions. This should inform in which domains this technique is applied.

As discussed in the original work, the high computational costs motivate the exploration of modified methods which selectively apply MEMO adaptation. In their Appendix A, they explore thresholding based on the marginal entropy but find it degrades performance [26]. Based on the

	Predictive Entropy Decrease			MEMO (B=16) Marginal Entropy		
	Improved	Worsened	Unchanged	Improved	Worsened	Unchanged
CIFAR-10 Original	0.643	0.514	0.096	0.833	0.941	0.220
CIFAR-10 New	0.697	0.551	0.166	0.963	1.070	0.360
ImageNet-R	1.669	1.223	1.173	3.061	3.014	2.817
ImageNet-A	0.861	N/A	1.021	3.315	N/A	2.900

Table 2. MEMO predictive entropy decrease and marginal entropy across augmentations stratified by samples where predictions were improved, worsened, or left unchanged. Note that there are no worsened samples for ImageNet-A because the baseline error rate is nearly 100%, thus those table entries are marked N/A.

results presented in Fig. 3 and Tab. 2, it seems that a predictive entropy threshold could help target samples likely to benefit from MEMO adaptation. The initial predictive entropy can be calculated with a single forward pass rather than requiring B augmentations and forward passes to compute marginal entropy. Thus predictive entropy could offer a more efficient thresholding method. This experimentation is left to future works.

The original work also explores the idea of continuous ad-hoc adaptation of the model. However, the adaptation only optimizes the marginal entropy without consideration for prediction accuracy, meaning that accumulating these updates can harm model accuracy. In fact, in their experimentation the authors report that the model collapses to always predict one class with full confidence [26]. This degenerate solution is clearly undesirable but does minimize the marginal entropy. This illustrates a key challenge for continuous test time adaptation, the model must stay accurate despite updating without access to ground truth labels. Until this challenge is overcome, these local adaptation methods are inherently wasteful, calculating input-specific model updates then discarding them.

Further, this report finds that MEMO harms the calibration of the underlying model. This should not come as a surprise because the marginal entropy minimization incentivizes the model to make more confident predictions without consideration for calibration. With these considerations in mind, this method is not appropriate for applications where model calibration is an important consideration. Exploring test time adaptation processes which maintain or improve calibration is an interesting but challenging area for future work, see [13] for one example.

Another direction for future analysis could look more extensively at which groups of augmentations to use. The original work does one ablation finding AugMix to outperform a set of 'standard augmentations' [26]. However, given the extensive number of augmentation techniques which have been developed in the field of computer vision it seems likely that there is more room for optimization here, different augmentation spaces may combine with varying degrees of success with base models that were trained differently.

In conclusion, while MEMO does offer improved robustness from a single test input, one should be mindful of the trade-offs in inference costs and model calibration when deciding if it is appropriate.

References

[1] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Apr. 2009. 3

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1

[3] Terrance DeVries and Graham W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout, Nov. 2017. arXiv:1708.04552 [cs]. 1

[4] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017. 2, 4

[6] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. 1

[7] 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, Las Vegas, NV, USA, June 2016. IEEE. 3

[8] Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-Resilient TabPFN: In-Context Learning Temporal Distribution Shifts on Tabular Data. In *Advances in Neural Information Processing Systems*, Nov. 2024. 1

[9] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, Oct. 2021. ISSN: 2380-7504. 3, 4

[10] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 1, 3

[11] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 1, 4

[12] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. *CVPR*, 2021. 3, 4

[13] Yong-Yeon Jo, Byeong Tak Lee, Jeong-Ho Hong, Hak Seung Lee, Joon-myoung Kwon, and Beom Joon Kim. Test-time calibration: A framework for personalized test-time adaptation in real-world biosignals. In Xuhai Orson Xu, Edward Choi, Pankhuri Singhal, Walter Gerych, Shengpu Tang, Monica Agrawal, Adarsh Subbaswamy, Elena Sizikova, Jessilyn Dunn, Roxana Daneshjou, Tasmie Sarker, Matthew McDermott, and Irene Chen, editors, *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pages 381–394. PMLR, 25–27 Jun 2025. 6

[14] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664. PMLR, July 2021. ISSN: 2640-3498. 1

[15] Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards Robust Vision Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12042–12051, June 2022. 4

[16] Imad Eddine Marouf, Enzo Tartaglione, and Stéphane Lathuilière. Mini but Mighty: Finetuning ViTs With Mini Adapters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1732–1741, Jan. 2024. 1

[17] A. Emin Orhan. Robustness properties of Facebook's ResNeXt WSL models, Dec. 2019. arXiv:1907.07640 [cs]. 1

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 4

[19] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. The MIT Press, Dec. 2008. 1

[20] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 Classifiers Generalize to CIFAR-10?, June 2018. arXiv:1806.00451 [cs]. 3

[21] Ivan Rubachev, Nikolay Kartashev, Yury Gorishniy, and Artem Babenko. TabReD: Analyzing Pitfalls and Filling the Gaps in Tabular Deep Learning Benchmarks. In *International Conference on Learning Representations*, Oct. 2024. 1

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 3

[23] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, pages 11539–11551. Curran Associates, Inc., 2020. 1

[24] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*, 2021. 1

[25] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, Apr. 2020. 1

[26] Marvin Mengxin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test Time Robustness via Adaptation and Augmentation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 4, 5, 6