EXAMINING TRENDS IN OUT-OF-DOMAIN CONFIDENCE

Hamza Qadeer, Michael Chau, Eric Zhu, Matthew A. Wright & Richard Liaw
RISELab @ UC Berkeley
Berkeley, CA 94720, USA
{hamza.qadeer,mchau16634,eric.zhu,mwright,rliaw}@berkeley.edu

Abstract

We closely investigate the relationship between a neural network's raw confidence and accuracy out-of-domain, asking if confidence can provide any insight into out-of-domain detection. We find that with few exceptions, modern neural networks are indeed less confident on out-of-domain predictions; they appear to have some inherent knowledge of what they do not know. This relationship does not appear to be consistent or guaranteed, however, and likely cannot be used as more than a rough heuristic for out-of-domain detection.

1 INTRODUCTION

In practical machine learning applications where reliability and model interpretability are important (i.e., healthcare, autonomous driving, security), developers need to understand the performance and correctness characteristics of their trained models (Guo et al., 2017). Recent work has focused on understanding the *calibration* of neural networks, drawing a relationship between confidence and accuracy for a given trained model. However, many modern neural network architectures have achieved excellent classification accuracies at the expense of poorly-calibrated probability distributions, leading to wildly overconfident predictions (Guo et al., 2017; Ovadia et al., 2019).

Completely out-of-domain data, which can arise in real-world scenarios due to sensor failures, adversarial attacks, or changing real-world conditions (Kumar et al., 2020; Vergara et al., 2012; Bobu et al., 2018; Farshchian et al., 2018), further reinforces the need for well-calibrated models. Ideally, models should detect these situations and adjust their predictions and confidence levels accordingly. In this work we explore the behavior of raw confidence under domain shift. We find that confidence usually decreases out-of-domain, though this relationship is neither consistent nor guaranteed. Our work suggests that raw confidence can serve as a helpful, albeit imperfect, heuristic for out-of-domain detection.

2 RELATED WORKS

Calibration captures the relationship between a model's confidence and its accuracy. In the case of perfect calibration for a k-class classification problem,

$$P(j|\mathbf{X}) = (\sigma(f(\mathbf{X}))_j \tag{1}$$

where $\mathbf{X}$ is an input tensor, σ is the softmax function, j is an integer-valued label, and $f(\mathbf{X}) \in \mathbb{R}^k$ is the model's pre-softmax output (class logits).

Much of the prior literature connecting calibration with raw (i.e., with no training-time augmentation or regularization) confidence and accuracy has focused on temperature scaling, a simple and effective way to improve calibration for neural models without affecting accuracy (Guo et al., 2017). Temperature scaling introduces a single scalar parameter τ. Using the same notation as before, the final prediction vector $\boldsymbol{y}$ is given by

$$\boldsymbol{y} = \sigma\left(\frac{f(\mathbf{X})}{\tau}\right) \tag{2}$$

Typically, $\tau > 1$, meaning that temperature scaling simply increases entropy across all predictions.

The high-confidence regions of a model can extend well outside the training distribution, as shown by the existence of out-of-distribution data points that fool models into very high confidence values

(Geng et al., 2018; Nguyen et al., 2014; Goodfellow et al., 2014). Unsurprisingly, temperature scaling fails to generalize well out-of-domain (Ovadia et al., 2019; Desai & Durrett, 2020).

3 Out-of-Domain Confidence

3.1 Discussion

At a fundamental level, temperature scaling works in-domain because models – even when they are overconfident in all cases – are less confident on low-accuracy data points. Given the opaque, uninterpretable nature of modern neural networks, the fact that this pattern holds across a diverse set of models and datasets is surprising and entirely non-trivial (Guo et al., 2017). We ask if, taking a longer-term view of confidence (i.e., a trailing average), this same trend generally extends out-of-domain.[1]

Temperature scaling's poor generalizability to new domains (Ovadia et al., 2019) implies that the drop in accuracy out-of-domain is not accompanied by a correspondingly large drop in confidence. This, however, begs the question of whether there is *any* accompanying drop in confidence. To the best of our knowledge, this question is currently unanswered. Just as a consistently overconfident model is likely to have lower confidence for low-accuracy predictions, a model with poor OOD calibration could still have lower average confidence out-of-domain, providing clues toward out-of-domain detection. Moreover, if this relationship is consistent across different out-of-domain sets, temperature scaling on the macro-level trends in confidence could provide a simple, on-the-fly solution to miscalibration under various types of domain shift.

3.2 Methods

To explore the viability of such a solution, we will compare performance in-domain and out-of-domain for several models across both textual and visual datasets, considering density distributions and trailing-average plots of confidence and accuracy. We will also take measurements over perturbed versions of the original data, allowing us to observe confidence and accuracy degradation over a smooth gradient and providing direct comparisons to the true OOD set.

The perturbed data will consist of several copies of the evaluation set stacked together with varying levels of perturbations, spanning a gradient from the original dataset to pure noise. In the text domain, noise will be simulated by replacing a word with another word from the dictionary at random. For images, we will employ perturbations both with Gaussian noise and with convergence to a fixed constant pixel value.[2]

More details on the experimental setup are available in Appendix A.

4 Results

We summarize our results here, but complete charts of all experiments are included in Appendix B.

4.1 Text Models

4.1.1 LSTM on Newsgroups-20

We trained a single-layer BiLSTM (Hochreiter & Schmidhuber, 1997) model via distillation (Tang et al., 2019) with RoBERTA (Liu et al., 2019) on the even genres of the Newsgroups20 dataset

[1] We focus on the general case instead of specially-crafted adversarial examples. Given the existence of in-domain, adversarial points (Stutz et al., 2018), this issue is not unique to out-of-domain data; temperature scaling can serve as an effective light-weight strategy in many use cases despite its vulnerability to adversarial attack.

[2] Recent works (Stutz et al., 2018; Kong et al., 2020) have shown more targeted methods of creating perturbed data outside the original manifold. However, the gradual accuracy decay in our experiments suggests that our coarser approach is effective as well. Moreover, our strategy results in a gradual descent off the manifold, rather than the immediate drop-off obtained by travelling even a small distance along an adversarial direction (Gilmer et al., 2018).

	Conf.	Accuracy Real	Noise
ID	0.948	0.755	0.765
OOD	0.868	**0.168**	**0.410**

Table 1: LSTM on NG-20

	Conf.	Accuracy Real	Noise
ID	0.884	0.781	0.760
OOD	0.792	**0.509**	**0.520**

Table 2: LSTM on SNLI

	Conf.	Accuracy Real	Noise
ID	0.962	0.896	0.902
OOD	0.873	**0.074**	**0.585**

Table 3: BERT on NG-20

The first row of the table shows confidence and accuracy on the in-domain set and the unperturbed portion of the noise set. The second row compares accuracies on the true OOD set and the perturbed at the same confidence level.

(Lang, 1995), using the odd genres as the OOD set. Even and odd pairs were collapsed onto the same label, so that the ID and OOD sets had the same sets of labels. (For example, the predicted label *0* corresponded to Newsgroups genres *0* and *1*; this made it possible for the model to "predict" the OOD set accurately.)

The ID and OOD confidences were clearly distinct, and confidence and accuracy both decayed slowly with noise on the perturbed set. However, the relationship between accuracy and confidence here was not consistent; for the same drop in confidence, the true out-of-domain set had a much larger drop in accuracy than the perturbed set (Fig. 1) (Table 1).

4.1.2 LSTM ON SNLI

We trained the same distilled BiLSTM model on SNLI (Bowman et al., 2015), using MNLI (Williams et al., 2017) as the out-of-domain set. (Note that MNLI, while distinct from and more difficult than the model's training domain, is still a related dataset and is not completely OOD data.)

Once more, out-of-domain confidence was markedly lower than in-domain, the ID and OOD sets had distinct confidence distributions, and confidence declined with noise. The accuracy drop-offs for the noise and the actual OOD set were well-aligned in this case, but confidence leveled off in the noise plot well before accuracy (Fig. 2) (Table 2).

4.1.3 BERT ON NEWSGROUPS-20

We fine-tuned a BERT (Devlin et al., 2018) model for classification on the Newsgroups20 dataset, with the same label coalescence as before.

Despite the change in model, there was again a clear but inconsistent relationship between confidence and accuracy (Fig. 3) (Table 3).

4.2 IMAGE MODELS

4.2.1 RESNET-18 ON CIFAR-10

We trained a ResNet18 (He et al., 2015) model on CIFAR-10 (Krizhevsky, 2012) with SVHN (Netzer et al., 2011) as the OOD set.

The relationship was identifiable but inconsistent, as accuracy leveled off before confidence in the noise chart, and the drop in confidence on the OOD set lead to a larger drop of accuracy than the same confidence drop over noise. Interestingly, confidence ticked upward as we converged to a constant value (Fig. 4) (Table 4).

	Conf.	Accuracy		
		Real	Noise	Const.
ID	0.959	0.850	0.860	0.839
OOD	0.863	**0.113**	**0.247**	**0.474**

Table 4: ResNet-18 on CIFAR-10

	Conf.	Accuracy		
		Real	Noise	Const.
ID	0.982	0.948	0.945	0.947
OOD	0.845	**0.128**	**0.106**	**0.516**

Table 5: DLA on CIFAR-10

4.2.2 DLA ON CIFAR-10

We trained a Deep Layer Aggregation (Yu et al., 2017) on CIFAR-10 and again used SVHN as the OOD set.

The results continued the general trend of declining confidence corresponding to declining accuracy in both the noise samples and the true OOD data. Yet the noise plot and the constant value plots both featured upticks; two data regions with the same average confidence correspond to dramatically different accuracies. Importantly, accuracy degradation was not consistent here either (Fig. 5) (Table 5).

4.2.3 RESNET-18 ON MNIST

We trained a ResNet18 (He et al., 2015) model on MNIST (Lecun et al., 1998), using SVHN (Netzer et al., 2011) (out-of-domain, but related), and Fashion MNIST (Xiao et al., 2017) (completely out-of-domain) as our OOD sets. As in the coalesced Newsgroups case, these datasets all shared the same labels.

The results for this setup were perhaps the most interesting of all our experiments. There was a noticeable difference between the out-of-domain and in-domain confidences; this is visible in both the density plot and the direct comparison. However, there was no difference in accuracy or confidence between the SVHN set (out-of-domain, but related; like MNIST, the dataset consists of digits) and Fashion MNIST (totally out-of-domain). In both cases, our predictions fared no better than random guesses.

Even more surprising was the trend in both noise charts: This model approached 100% confidence on both random noise and constant values. (We tried a range of values and saw the same result on all of them; this is not the result of a particularly hostile value.) Although the connected between confidence and accuracy discussed thus far may seem obvious, this anomalous result proves that it is not a foregone conclusion. Even in general use cases, a model's confidence can provide zero indication that it is out-of-domain (Fig. 6).[3]

5 CONCLUSIONS

Consistently, across a wide range of examples, we observe that there is a marked difference in confidence in-domain versus out-of-domain. However, this pattern is not guaranteed and is inconsistent across different OOD datasets. Our anomalous results with ResNet on the digit datasets seem particularly relevant to sensitive applications, as they suggest even general use cases – not just handcrafted adversarial examples – can expose the poor confidence alignment of neural networks.

It seems very unlikely that patterns in confidence could be used as part of a post-training scheme to tune temperature and recalibrate a model out-of-domain. Nonetheless, our results demonstrate that long-term confidence trends may serve as a useful, if highly imperfect, heuristic for out-of-domain detection. Future works may seek to investigate models' prediction zones and explain why certain OOD samples trigger high confidence values. This, in turn, might lead to more effective inference-time heuristics for OOD detection.

[3]No table is included for this experiment because the average confidence on the synthetic set never reached the level seen on the OOD data.

ACKNOWLEDGMENTS

In addition to NSF CISE Expeditions Award CCF-1730628, research at RISELab is supported by gifts from Amazon Web Services, Ant Group, Ericsson, Facebook, Futurewei, Google, Intel, Microsoft, Nvidia, Scotiabank, Splunk and VMware.

REFERENCES

Andreea Bobu, E. Tzeng, Judy Hoffman, and Trevor Darrell. Adapting to continuously shifting domains. In *ICLR*, 2018.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. *arXiv e-prints*, art. arXiv:1508.05326, August 2015.

Shrey Desai and Greg Durrett. Calibration of Pre-trained Transformers. *arXiv e-prints*, art. arXiv:2003.07892, March 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, art. arXiv:1810.04805, October 2018.

Ali Farshchian, Juan A. Gallego, Joseph P. Cohen, Yoshua Bengio, Lee E. Miller, and Sara A. Solla. Adversarial Domain Adaptation for Stable Brain-Machine Interfaces. *arXiv e-prints*, art. arXiv:1810.00045, September 2018.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent Advances in Open Set Recognition: A Survey. *arXiv e-prints*, art. arXiv:1811.08581, November 2018.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S. Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial Spheres. *arXiv e-prints*, art. arXiv:1801.02774, January 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv e-prints*, art. arXiv:1412.6572, December 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. *arXiv e-prints*, art. arXiv:1706.04599, June 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv e-prints*, art. arXiv:1512.03385, December 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated Language Model Fine-Tuning for In- and Out-of-Distribution Data. *arXiv e-prints*, art. arXiv:2010.11506, October 2020.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, May 2012.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding Self-Training for Gradual Domain Adaptation. *arXiv e-prints*, art. arXiv:2002.11361, February 2020.

Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pp. 331–339. Elsevier, 1995.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, art. arXiv:1907.11692, July 2019.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NIPS*, January 2011.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014. URL `http://arxiv.org/abs/1412.1897`.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift, 2019.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. *CoRR*, abs/1812.00740, 2018. URL `http://arxiv.org/abs/1812.00740`.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling Adversarial Robustness and Generalization. *arXiv e-prints*, art. arXiv:1812.00740, December 2018.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from BERT into simple neural networks. *CoRR*, abs/1903.12136, 2019. URL `http://arxiv.org/abs/1903.12136`.

Alexander Vergara, Shankar Vembu, Tuba Ayhan, Margaret A Ryan, Margie L Homer, and Ramón Huerta. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166:320–329, 2012.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *arXiv e-prints*, art. arXiv:1704.05426, April 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv e-prints*, art. arXiv:1708.07747, August 2017.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep Layer Aggregation. *arXiv e-prints*, art. arXiv:1707.06484, July 2017.

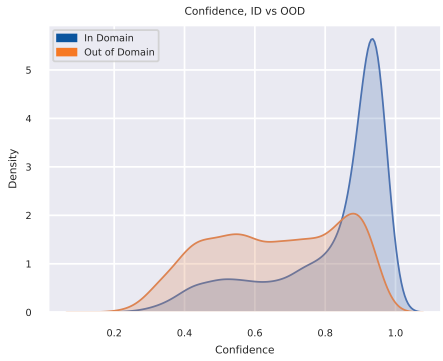
A Experimental Setup

We plotted three measurements for each model.

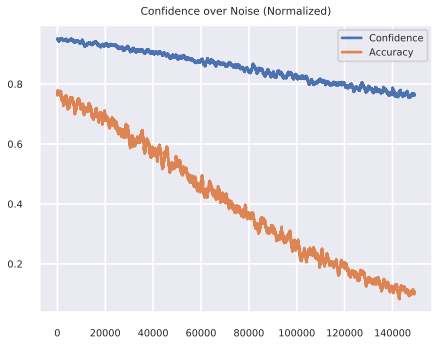
- A simple density plot to compare the distributions of confidence in-domain and out-of-domain. All confidences (ID and OOD) were temperature-scaled with $\tau = 2.5$ to enhance visibility.
- The trailing average (100) of confidence and accuracy over runs of in-domain and out-of-domain data. Our harness alternated between the in-domain and out-of-domain sets at random, using runs of length 500-1000 for each. The domain switches were not explicitly labeled but are visibly obvious in all of the plots.
- The trailing average (1000) of confidence and accuracy over introduced noise. (This serves as a gradual, controlled OOD set for the models, with which the actual OOD performance can be compared.) For these charts, 40-80 copies of the in-domain test set were stacked next to one another with increasing levels of noise; the first copy was the unaltered evaluation set, while the last copy was pure noise. For text models, noise was incorporated by swapping each word to a random word from the model's vocabulary with probability p, which increased in linear increments from 0 to 1. For image models, two separate forms of noise were plotted. In the first, zero-meaned Gaussian noise was introduced with linearly increasing variance. The image was then re-normalized to its original mean and variance to remove any distortions resulting from changed pixel magnitudes. For the second, each pixel x in the image was changed to $(1 - \alpha)x + \alpha c$, where c is a constant value close to the mean of the image. The parameter α was linearly increased from 0 to 1.

For the direct comparisons in the tables, we computed average confidence and average accuracy on the true in-domain and out-of-domain sets and then picked regions from the synthetic data corresponding to the same confidence levels. This allowed us to compare the accuracy-confidence relationship across two different OOD sets for the same model.

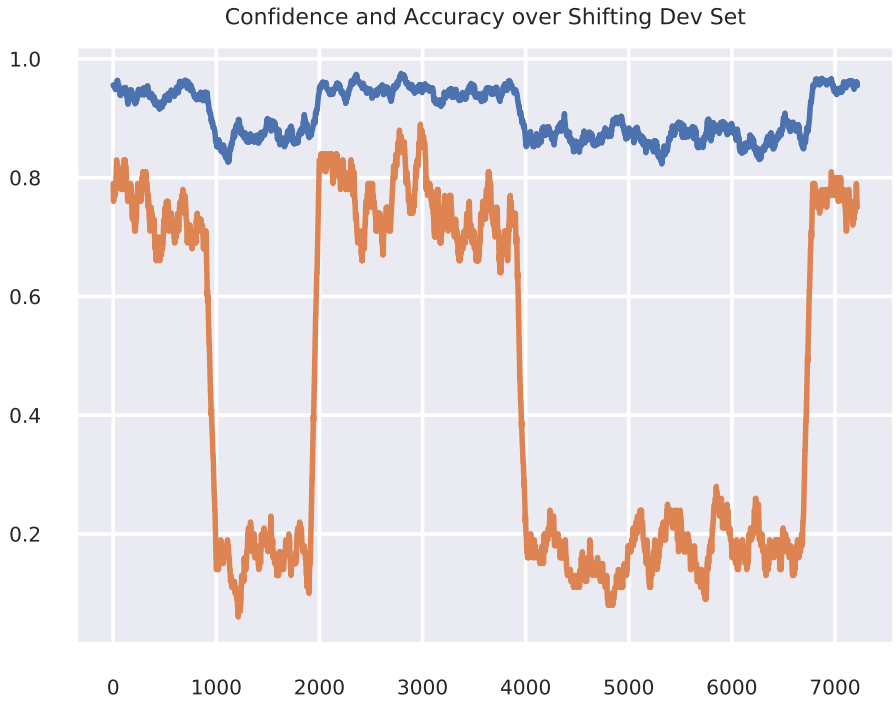
B CHARTS



(a) Density Plot

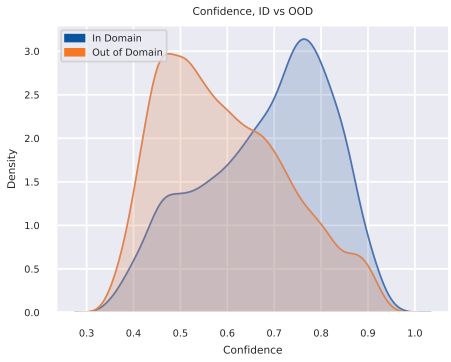


(b) Decay over Noise

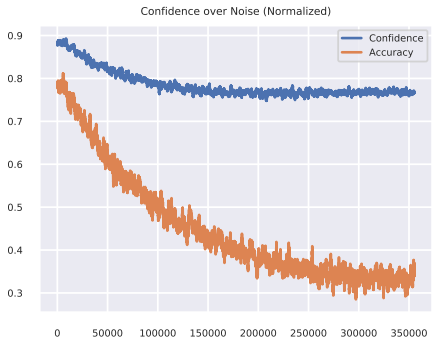


(c) Direct Comparison

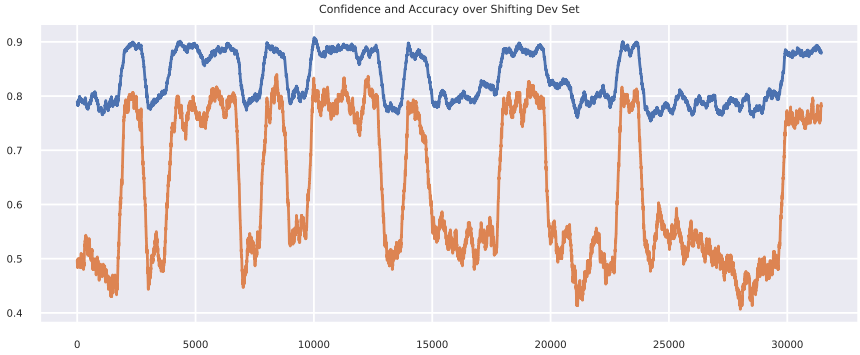
Figure 1: LSTM on Newsgroups-20



(a) Density Plot

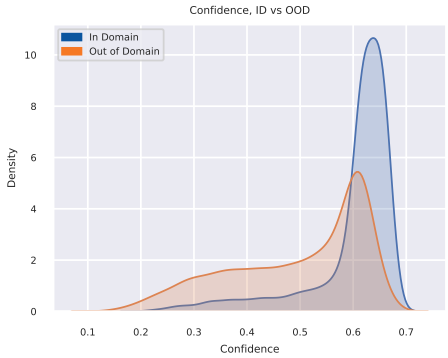


(b) Decay over Noise

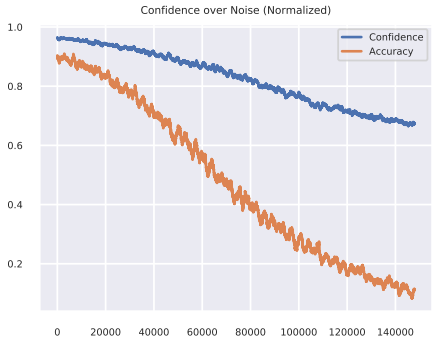


(c) Direct Comparison

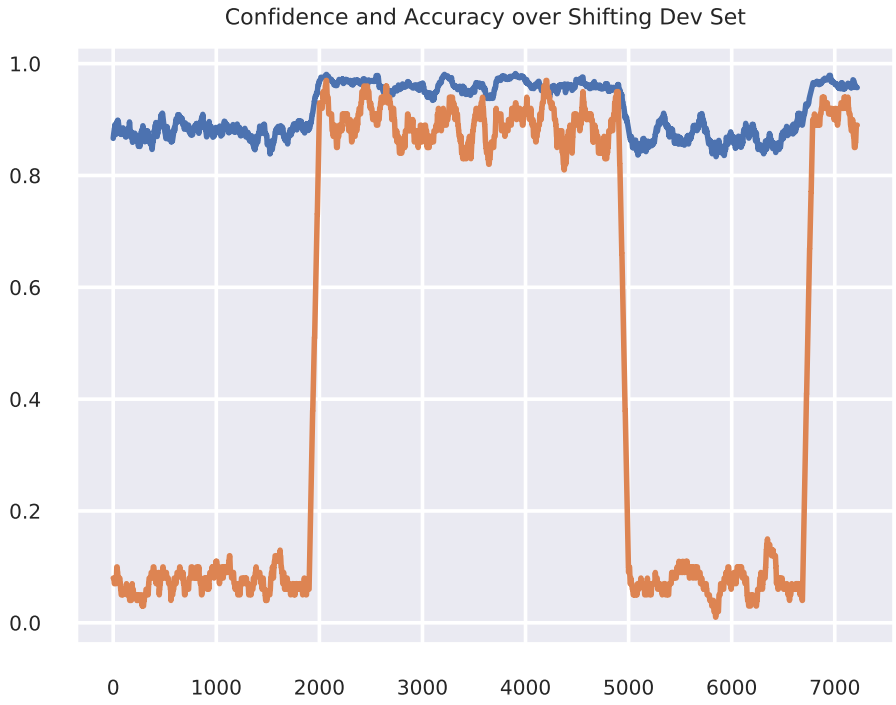
Figure 2: LSTM on SNLI



(a) Density Plot



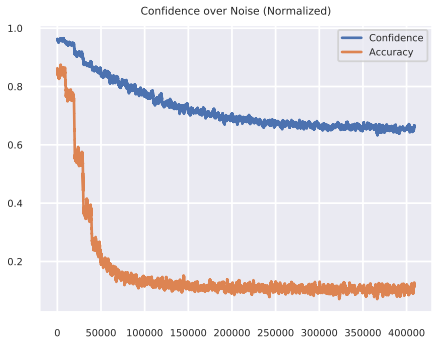
(b) Decay over Noise



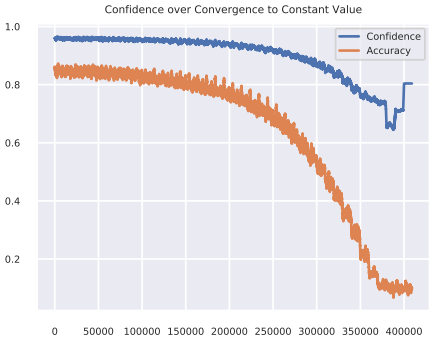
(c) Direct Comparison

Figure 3: BERT on Newsgroups-20

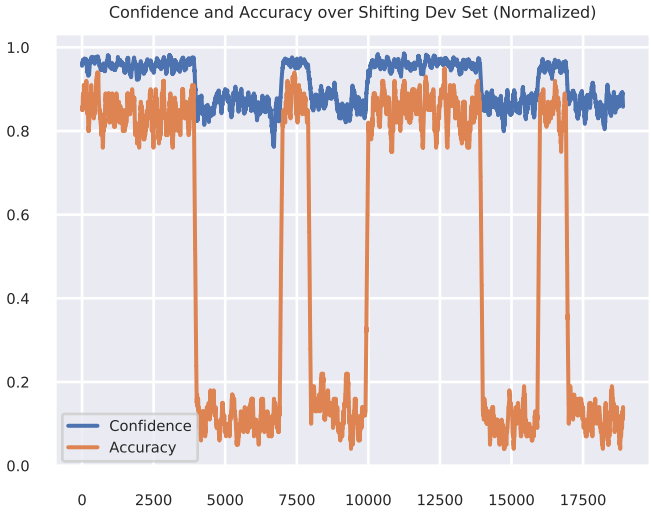
(a) Density Plot



(b) Decay over Noise

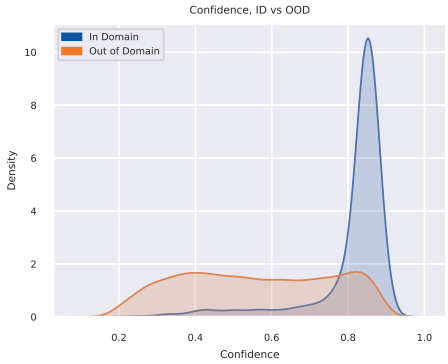


(c) Decay over Convergence to Constant

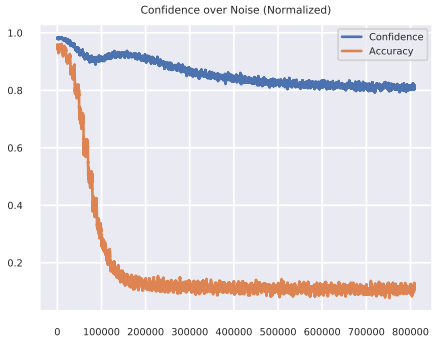


(d) Direct Comparison

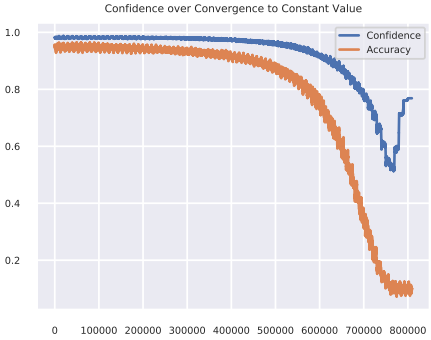
Figure 4: ResNet-18 on CIFAR-10



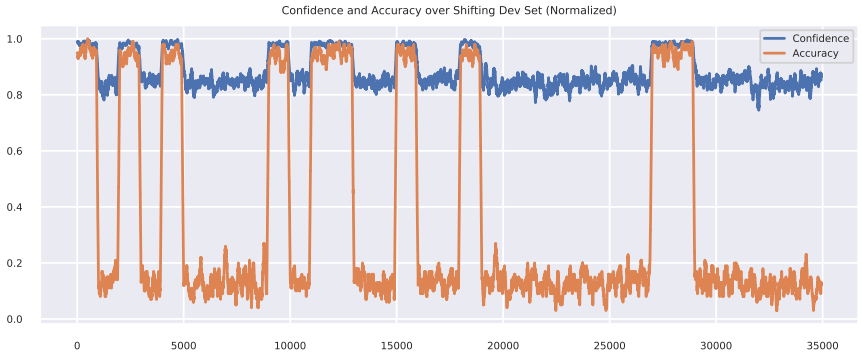
(a) Density Plot



(b) Decay over Noise

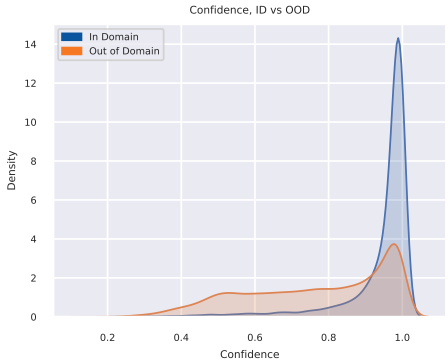


(c) Decay over Convergence to Constant

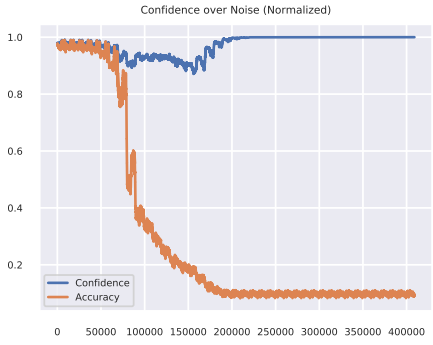


(d) Direct Comparison

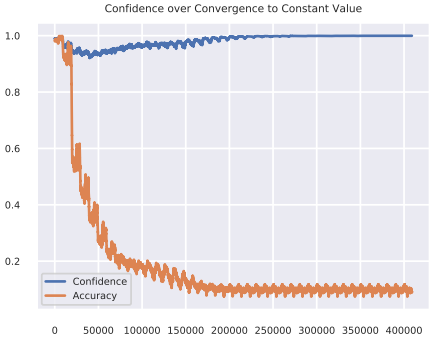
Figure 5: DLA on CIFAR-10



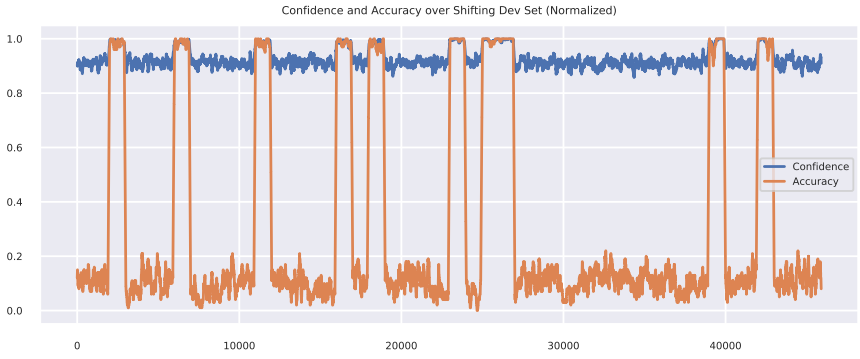
(a) Density Plot



(b) Decay over Noise



(c) Decay over Convergence to Constant



(d) Direct Comparison

Figure 6: ResNet-18 on MNIST