Uncertainty Estimation in Deterministic Vision Transformer

Wenqian Ye *, **Yunsheng Ma** *, **Xu Cao**

New York University
{wy2029, ym2382, xc2057}@nyu.edu

Abstract

Though Transformers have achieved promising results in many computer vision tasks, they tend to be over-confident in predictions, as the standard Dot Product Self-Attention (DPSA) can barely preserve distance for unbounded input domain. Existing uncertainty quantification approaches, such as Deep Ensemble and MC Dropout, are inapplicable to sizable Vision Transformers, owing to their high computational and memory cost. In this paper, we fill this gap by proposing a novel CoBiLiR Self-Attention module. Specifically, we replace the dot product similarity with the distance within Banach Space and also normalize the term by a theoretical lower bound of the Lipschitz constant. Extensive experiments conducted on standard vision benchmarks demonstrate that our method outperforms the state-of-the-art single forward pass approaches in prediction, calibration, and uncertainty estimation.

1 Introduction

The remarkable performance of deep learning (DL) has made it widely employed in all kinds of inference and decision making systems. Despite that, it makes mistakes, making DL trust and safety an increasingly important topic (Amodei et al. 2016; Jiang, Kim, and Gupta 2018). Especially when a DL model's prediction affects critical decisions, such as self-driving cars (Huang and Chen 2020) and medical diagnosis (Esteva et al. 2017). A tempting way to tackle this problem is for a model to not only achieve high accuracy, but also be able to quantify its uncertainty over its predictions.

Recently, Vision Transformers (ViT) (Dosovitskiy et al. 2021), which rarely uses convolutional kernels, *i.e.* the heart of CNNs, have achieved state-of-the-art performance across numerous CV tasks and received a lot of attention in the CV community (Carion et al. 2020). Although ViTs have shown remarkable predictive performance, like other deep neural networks (DNNs), they lack proper uncertainty quantification and are inclined to make overconfident predictions. This limitation is significant as ViT is becoming the state-of-the-art basic model in CV. In this paper, we study the under-explored problem of uncertainty estimation in ViT, which can help downstream tasks to build reliable models.

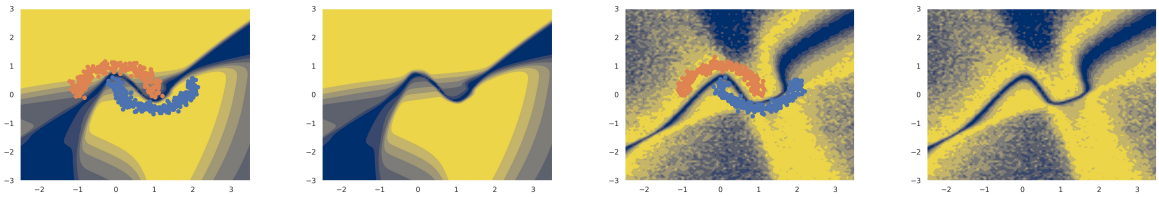
*These authors contributed equally.

Principled techniques to estimating a deep learning model's predictive uncertainty include (1) Bayesian deep learning (BDL) (Wilson and Izmailov 2020). For example, Blundell et al. (2015) proposed Bayes by Backprop to quantify the uncertainty of model weights. (2) Ensemble techniques, such as MC dropout (Gal and Ghahramani 2016), which uses dropout (Srivastava et al. 2014) as a regularization term to compute the prediction uncertainty and Deep Ensembles (Lakshminarayanan, Pritzel, and Blundell 2017). However, these multiple forward passes methods suffer heavy memory and computation cost, which limits their adoption in real-world applications.

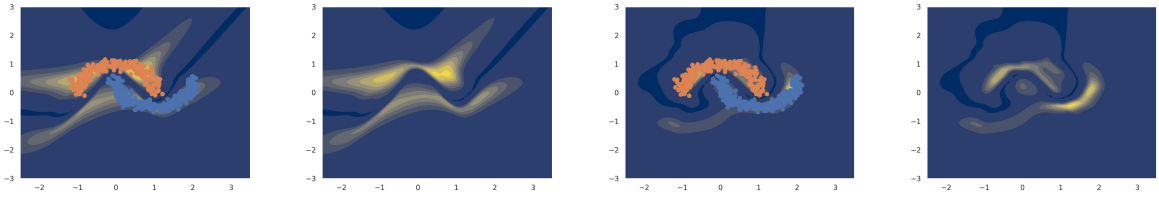
Alternatively, uncertainty quantification via single forward-pass neural networks, which has similar latency as a single deterministic network, has recently received lots of attention (Gillioz et al. 2020). SNGP (Liu et al. 2020) replaces the dense output layer with a Gaussian Process (GP) layer and applies Spectral Normalization (SN) (Miyato et al. 2018) to the hidden residual layers. DUE (van Amersfoort et al. 2021) builds upon GPDNN (Bradshaw, de G. Matthews, and Ghahramani 2017) and introduces additional constraints to the feature extractor in the form of residual connections in combination with SN (Miyato et al. 2018). These methods perform well on uncertainty estimation. However, they only focus on bounding the Lipschitz constants of certain CNN modules *i.e.*, convolution and batch normalization (Ioffe and Szegedy 2015) layers. Moreover, according to Lee et al. (2021), Transformer blocks are sensitive to the magnitude of Lipschitz constant, and training will progress slowly when SN is employed in self-attention modules. In this paper, we propose CoBiLiR self-attention based Transformer Gaussian Process, termed TGP, to address the above problems for uncertainty estimation without sacrificing its predictive ability.

In summary, our contributions are threefold:

- We propose a novel regularization method, termed *CoBiLiR*, to solve the distance-awareness in both aspects of Lipschitzness and Contraction problems by replacing the dot product similarity with the distance within Banach Space and normalizing the term by a theoretical lower bound of the Lipschitz constant.
- We develop a novel CoBiLiR self-attention based Transformer Gaussian Process, *i.e.* TGP that integrates distance-preserving hidden mappings in the transformer blocks via CoBiLiR, and GP as a distance-aware output layer for



(a) Deep Ensemble (b) DUE



(c) SNGP (d) **TGP (Ours)**

Figure 1: Uncertainty heat map of TGP and baseline approaches on the two moons 2D classification benchmark. Orange and blue points are positive and negative training samples respectively. Background color visualizes the predictive uncertainty of each model, where yellow stands for confidence and blue indicates uncertainty. The proposed TGP model (Figure 1(d)) achieves the closest to ideal uncertainty quantification on this benchmark.

high quality uncertainty estimation.

- We conduct extensive experiments on the commonly-used uncertainty benchmarks CIFAR-10/-100 vs SVHN and CIFAR-10/-100 vs CIFAR-100/-10, respectively. Compared to the state-of-the-art approaches, the results demonstrate the superiority of the proposed TGP model in prediction, calibration, and uncertainty estimation with no penalty on time complexity.

2 Method

We propose CoBiLiR self-attention based Transformer Gaussian Process (TGP), a novel solution to improve distance awareness of hidden space in ViT with a Contractive Bi-Lipschitz Regularization method (CoBiLiR) and a Gaussian Process Layer. The structure of TGP is built upon the foundation of GPDNN (Bradshaw, de G. Matthews, and Ghahramani 2017).

Contractive Bi-Lipschitz Regularization (CoBiLiR)

Scaled Dot-Product Self-Attention does not satisfy the *bi-Lipschitz condition* (See Appendix). To extend the generality of self-attention with high-quality uncertainty estimation, we propose a new regularization method (CoBiLiR) by replacing the self-attention function with a contractive Bi-Lipschitz expression without losing the original ability of representation. We will explicitly discuss in separate aspects to see how to achieve both Lipschitzness and Contraction in our method.

Lipschitzness Following the proof of the statement that *Dot-Product Self-Attention is not Lipschitz* by Kim, Papamakarios, and Mnih (2021), suppose there exists such mapping $f(X)$, $X \in \mathbb{R}^N$:

$$f(X) = S \cdot X = \text{softmax}(aX \cdot X^\top) \cdot X = \begin{bmatrix} f_1(X) \\ \vdots \\ f_N(X) \end{bmatrix}$$

Its Jacobian Matrix is $J_f = [J_{ij}]_{N\times N}$, each entry can be written as:

$$J_{ij} = aX^\top S^{(i)} \left[E_{ji}X + \delta_{ij}X\right] + S_{ij}I \in \mathbb{R}^{N\times 1}$$

Thus for $i = j$:

$$J_{ii} = aX^\top S^{(i)} E_{ii}X + aX^\top S^{(i)}X + S_{ii} \qquad (1)$$

$X^\top S^{(i)}X$ is in the form of a variance of a discrete distribution. When $\mathbf{x}_i = \mathbf{0}$ for some i, some entries of the Jacobian of f grow proportionally to the sample variance of $\mathbf{x}_{\neq i}$.(The softmax probabilities $S_{i:}$ are constant with respect to $\mathbf{x}_{\neq i}$ when $\mathbf{x}_i = 0$.) This will lead to an unbounded Jacobian matrix. To avoid this pathology, we replace $Q \cdot K^\top$ by $b = -\|\mathbf{x}_i^\top Q - \mathbf{x}_j^\top K\|_2^2$ in $\text{Attention}(X)$. Here, the new similarity measurement lies on the Banach Space (complete vector space with norm $\|\cdot\|$), which is a more generalized space over Hilbert Space (complete inner product space) (Megginson 2012). This modification gives a strong theoretical guarantee on Lipschitzness with easy matrix multiplications during training.

Contraction Contraction of the Scaled Dot-Product Self-Attention is another crucial issue for achieving well-calibrated uncertainty. Deriving such contraction scalar requires a theoretical lower bound of Lipschitz constant on the Dot-Product Self-Attention function. A desirable contraction scalar could be non-strict but easy to compute during training.

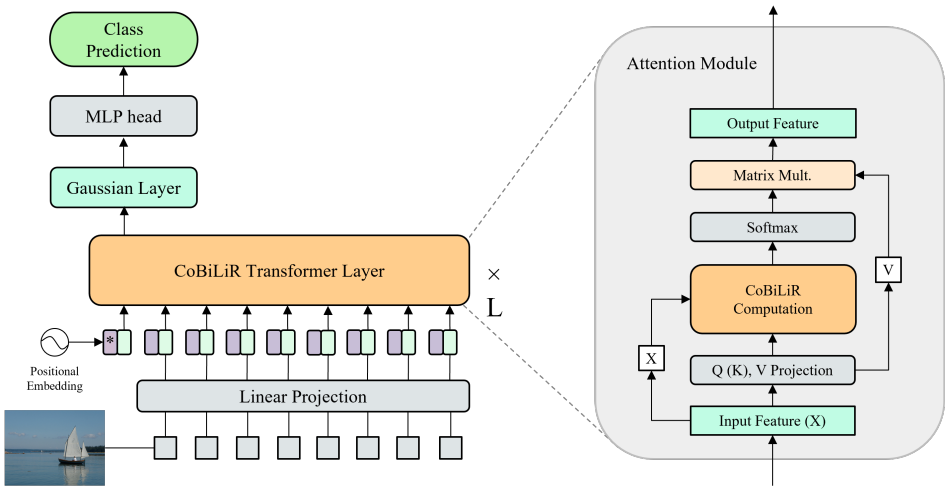


Figure 2: TGP Model Overview: The model prepends a [class] token and turns the input images into the embedding of patches with positional embedding. In the CoBiLiR Transformer Layer, we apply CoBiLiR function into the Attention Module with the rest parts same as standard ViT. The first learnable token is fed into the Gaussian Process Layer to get the class prediction. In the CoBiLiR Attention Module, we first project X into corresponding query, key and value matrices Q, K and V ($K = Q$). Then, we feed Q, K, X into the CoBiLiR function 4 and the output is activated through Softmax function. The module returns the matrix multiplication of the attention map and the value matrix V. With a Gaussian Process Layer with Random Fourier Features (RFF), the model will output the prediction of classes. In the following experiments, we refer to TGP with various ViT configurations using the same naming convention (*e.g.* TGP-Ti, TGP-S, and TGP-B).

Inspired by Theorem B.1 in Appendix, we introduce a proper regularization scalar function with a **Scalar Factor** α by replacing $\tilde{g}(X)$ with $Q \cdot K^\top$:

$$c(X) = \frac{\alpha}{\|Q\|_F \cdot \|X^\top\|_{(\infty,2)}} \tag{2}$$

α controls the scale of $\tilde{g}(X)$.

Here, we assign it as a hyperparameter in control of the corresponding Lipschitz constants for proper contraction of each model. A small alpha result in loss of information while a large alpha cause the model tending to be non-Lipschitz.

Summary Here is our formal definition of Contractive Bi-Lipschitz Regularization (CoBiLiR) on the Self-Attention function:

$$S_{ij} := b \cdot c(X) = -\frac{\alpha \|\mathbf{x}_i^\top W_Q - \mathbf{x}_j^\top W_K\|_2^2}{\|Q\|_F \cdot \|X^\top\|_{(\infty,2)}} \tag{3}$$

This pair-wise operation can alternatively be implemented as a matrix version for improved computational efficiency:

$$S = -\alpha \cdot \frac{\|Q\|_{\text{row}}^2 - 2QK^\top + \|K\|_{\text{row}}^{2^\top}}{\|Q\|_F \cdot \|X^\top\|_{(\infty,2)}} \tag{4}$$

Gaussian Process Layer

In TGP, to perserve the distance awareness between input test sample and previously seen training data, we simply replace the typical dense output layer with a Gaussian Process (GP) with an RBF kernel following SNGP (Liu et al. 2020). This approach makes sure the model returns a uniform distribution over output labels when the input sample is OoD.

To make it end-to-end trainable, the Gaussian Process layer can be implemented a two-layer network:

$$\text{logits}(x) = \Phi(x)\beta, \quad \Phi(x) = \sqrt{\frac{2}{M}} * \cos(Wx + b) \tag{5}$$

Here x is the input, and W and b are frozen weights initialized randomly from Gaussian and uniform distributions, respectively. $\Phi(x)$ is Random Fourier Features (RFF) (Williams and Rasmussen 2006). β is the learnable kernel weight similar to that of a Dense layer. The layer outputs the class prediction $\text{logits}(x) \in \mathbb{R}_{\text{NumClasses}}$.

3 Experiments

In this section, we conduct ablation studies and compare TGPs with several SOTA methods. We design ablation experiments including module comparison under ViT-Ti, searching for a proper scalar factor α and validating the reliability of pretrained models.

Setup

Benchmarks We evaluate the performance of the proposed TGP model on the OoD benchmark (Miyato et al. 2018) using SVHN (Netzer et al. 2011) as the OoD dataset for the model trained on CIFAR-10/-100 (Krizhevsky 2009). OoD data is never seen during training, whereas ID samples are semantically similar to training samples.

Baselines We chose the following methods as baselines to compare TGP to state-of-the-art approaches for uncertainty prediction: (1) Deep Ensemble with 5 models (5-Ensemble) (Lakshminarayanan, Pritzel, and Blundell 2017), (2) SNGP

Method	Accuracy (↑)	ECE (↓)	NLL (↓)	OoD AUROC (↑)		OoD AUPR (↑)	
				SVHN	CIFAR-100	SVHN	CIFAR-100
5-Ensemble*	96.6 ± 0.01	0.010 ± 0.001	0.114 ± 0.01	0.967 ± 0.005	-	0.964 ± 0.01	0.888 ± 0.01
DUQ*	94.7 ± 0.02	0.034 ± 0.002	0.239 ± 0.02	0.940 ± 0.003	-	0.973 ± 0.01	0.854 ± 0.01
DUE*	95.6 ± 0.04	0.018 ± 0.002	0.187 ± 0.01	0.958 ± 0.005	-	-	-
SNGP*	95.9 ± 0.01	0.018 ± 0.001	0.138 ± 0.01	0.940 ± 0.006	-	0.990 ± 0.01	0.905 ± 0.01
TGP-S (ours)	$\mathbf{97.2 \pm 0.01}$	$\mathbf{0.012 \pm 0.001}$	$\mathbf{0.100 \pm 0.01}$	$\mathbf{0.983 \pm 0.005}$	$\mathbf{0.914 \pm 0.01}$	$\mathbf{0.993 \pm 0.01}$	$\mathbf{0.911 \pm 0.01}$

Table 1: Comparison between proposed TGP-S and SOTA methods on CIFAR-10 vs SVHN/CIFAR-100 benchmarks. The best method among single-network approaches is highlighted in **bold**. *Results from the original papers.

Method	Accuracy (↑)	ECE (↓)	NLL (↓)	OoD AUROC (↑)		OoD AUPR (↑)	
				SVHN	CIFAR-10	SVHN	CIFAR-10
5-Ensemble*	80.2 ± 0.01	0.021 ± 0.004	0.666 ± 0.02	-	-	0.888 ± 0.01	0.780 ± 0.01
DUQ*	78.5 ± 0.03	0.119 ± 0.001	0.980 ± 0.02	-	-	0.878 ± 0.01	0.732 ± 0.01
SNGP*	79.9 ± 0.03	0.025 ± 0.012	0.847 ± 0.01	-	-	0.923 ± 0.01	$\mathbf{0.801 \pm 0.01}$
TGP-S (ours)	$\mathbf{85.2 \pm 0.03}$	$\mathbf{0.018 \pm 0.005}$	$\mathbf{0.538 \pm 0.01}$	$\mathbf{0.896 \pm 0.01}$	$\mathbf{0.799 \pm 0.01}$	$\mathbf{0.955 \pm 0.01}$	0.777 ± 0.01

Table 2: Comparison between proposed TGP-S and the SOTA methods on CIFAR-100 vs SVHN and CIFAR-100 vs CIFAR-10 benchmark. The best method among single-network approaches is highlighted in **bold**. *Results from the original papers.

(Liu et al. 2020), (3) DUQ (Van Amersfoort et al. 2020) and (4) DUE (van Amersfoort et al. 2021).

Implementation Details In the following experiments, we resize input image to 224×224 pixels and set the patch size to 16. The ViT model is initialized with weights pre-trained on ImageNet-1K dataset (Russakovsky et al. 2015) except for the attention layers. All models are trained for 100 epochs with 5 different random seeds on 2 NVIDIA A100 GPUs.

Ablation Study

Module Comparison In this section, we compare TGP-Ti with existing uncertainty estimation approaches **applied to ViT-Ti**. For both Standard ViT-Ti and Deep Ensemble we take the predictive entropy as uncertainty. For SNGP, the entropy of the average of the Monte Carlo softmax samples is used as uncertainty. Note that for DUE and SNGP, only the GP output layers are applied to the ViT-Ti model. We do not compare with DUE for the CIFAR-100 dataset, as its training does not converge.

The accuracy, NLL, AUROC, AUPR results are shown in Appendix. The AUROC metric indicates the quality of uncertainty, since it measures the probability that in-distribution (ID) and OoD samples can be separated (Mukhoti et al. 2021). From the results, we have the following observations:

(1) For OoD detection, *The proposed TGP model outperforms other methods applied to ViT including Deep Ensemble as well as all single forward pass methods on CIFAR-10 vs SVHN and CIFAR-100 vs SVHN benchmarks*.

(2) Notably, the superior performance in OoD is achieved without sacrificing TGP's predictive performance. On the contrary, TGP even outperforms standard ViT in terms of classifications accuracy on the CIFAR-100 dataset, making TGP achieve the best performance in terms of all the metrics compared with all other single-network methods.

(3) Furthermore, the proposed CoBiLiR self-attention can be computed efficiently using matrix operations, with minimal overhead compared to the original dot-product self-attention. This ensures TGP's performance gains come without compromising computation cost.

Comparison with SOTA

Based on the ablation study above, we apply our method in larger TGP variants for achieving SOTA performance in prediction, calibration, and uncertainty estimation. Following Touvron et al. (2022), we adopt an existing training setup, namely the A3 procedure of Wightman, Touvron, and Jégou (2021). We adjust the learning rate of the A3 procedure when training TGPs, since it was originally designed for training ResNet-50 models. In our experiments, we set the learning rate to 0.006 for TGP-S when pretraining on ImageNet-1K and 0.004 while finetuning on CIFAR-10/-100. For TGP-B, we reduce the learning rate to 0.003 when pretraining on ImageNet-1K and 0.002 when finetuning on CIFAR-10/-100.

To evaluate the model's OoD detection performance, we adopt two OoD tasks suggested by SNGP: (1) using SVHN as the OoD dataset for a model trained on CIFAR-10/-100; (2) using CIFAR-100/-10 as the OoD dataset for a model trained on CIFAR-10/-100, respectively. We select TGP-S to fairly compare with other SOTA approaches for the following two reasons. (1) TGP-S has 19.9M parameters while backbones (WRN-28-10 (Zagoruyko and Komodakis 2016)) used in DUQ and SNGP has 36.5M parameters. (2) The performance gap between TGP-S and TGP-B is insignificant, as shown in Appendix. Table 1 and 2 show the main comparison results. TGP-S outperforms the other single forward pass approaches in all the metrics of CIFAR-10 and most of the metrics of CIFAR-100. Moreover, TGP-S also achieves better results than 5-Ensemble of WRN-28-10, which requires around $5\times$ as much time to execute as TGP-S and other single forward

pass approaches.

4 Conclusion

We propose CoBiLiR, an effective regularization method for ensuring the Bi-Lipschitz constraints and the contraction of self-attention mappings with theoretical guarantees. The developed TGP model consists of CoBiLiR self-attention layers in the ViT and a Gaussian Process output layer, which enables distance awareness for high quality uncertainty quantification. The extensive experiments conducted on the various OoD benchmarks demonstrate the efficiency and effectiveness of our method. Importantly, the superior performance in OoD is achieved without sacrificing TGP's predictive performance.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *ArXiv*, abs/1606.06565.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *International Conference on Machine Learning*, 1613–1622. PMLR.

Bradshaw, J. F.; de G. Matthews, A. G.; and Ghahramani, Z. 2017. Adversarial Examples, Uncertainty, and Transfer Testing Robustness in Gaussian Process Hybrid Deep Networks. *arXiv: Machine Learning*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. *ArXiv*, abs/2005.12872.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*, abs/2010.11929.

Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J. M.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542: 115–118.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Gillioz, A.; Casas, J.; Mugellini, E.; and Khaled, O. A. 2020. Overview of the Transformer-based Models for NLP Tasks. *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 179–183.

Huang, Y.; and Chen, Y. 2020. Autonomous Driving with Deep Learning: A Survey of State-of-Art Technologies. *ArXiv*, abs/2006.06091.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv*, abs/1502.03167.

Jiang, H.; Kim, B.; and Gupta, M. R. 2018. To Trust Or Not To Trust A Classifier. In *NeurIPS*.

Kim, H.; Papamakarios, G.; and Mnih, A. 2021. The Lipschitz Constant of Self-Attention. In *ICML*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *NIPS*.

Lee, K.; Chang, H.; Jiang, L.; Zhang, H.; Tu, Z.; and Liu, C. 2021. ViTGAN: Training GANs with Vision Transformers. In *International Conference on Learning Representations*.

Liu, J. Z.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax-Weiss, T.; and Lakshminarayanan, B. 2020. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. *ArXiv*, abs/2006.10108.

Megginson, R. E. 2012. *An introduction to Banach space theory*, volume 183. Springer Science & Business Media.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. *ArXiv*, abs/1802.05957.

Mukhoti, J.; Kirsch, A.; van Amersfoort, J. R.; Torr, P. H. S.; and Gal, Y. 2021. Deterministic Neural Networks with Inductive Biases Capture Epistemic and Aleatoric Uncertainty.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3): 211–252.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15: 1929–1958.

Touvron, H.; Cord, M.; El-Nouby, A.; Verbeek, J.; and J'egou, H. 2022. Three things everyone should know about Vision Transformers. *ArXiv*, abs/2203.09795.

Van Amersfoort, J.; Smith, L.; Teh, Y. W.; and Gal, Y. 2020. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, 9690–9700. PMLR.

van Amersfoort, J. R.; Smith, L.; Jesson, A.; Key, O.; and Gal, Y. 2021. On Feature Collapse and Deep Kernel Learning for Single Forward Pass Uncertainty.

Wightman, R.; Touvron, H.; and Jégou, H. 2021. ResNet strikes back: An improved training procedure in timm. *arXiv e-prints*, arXiv–2110.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wilson, A. G.; and Izmailov, P. 2020. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. *ArXiv*, abs/2002.08791.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.