Building Performance Evaluation Framework of Foundation Models for Nonproliferation Applications

Alexei N. Skurikhin*, Garrison S. Flynn, Michael A. Geyer, Giri R. Gopalan, Natalie E. Klein, Juston S. Moore, Mark G. Myshatyn, Nidhi K. Parikh, Rosalyn C. Rael, Selma L. Wanna, Emily M. Casleton
Los Alamos National Laboratory, Los Alamos, NM 87545, USA
*alexei@lanl.gov

Abstract
Recent progress in AI has culminated in foundation models (FMs) that can facilitate the development of innovative approaches for nuclear verification and geographic profiling of activities of interest. FMs are large-scale deep learning neural network models (e.g., transformer models) that are trained on very large general datasets and can then be tuned to a wide range of downstream tasks with relatively little additional task-specific training. FMs have already demonstrated a huge impact in natural language processing and are increasingly used in computer vision for tasks such as image-to-text mapping, image retrieval, and image tagging. However, while FMs are powerful models, their adaptation to the domain of nuclear nonproliferation comes with potential limitations due to inadequate quality and variety of data, as well as a possibility of bias in the data used to train the original FM. Test & evaluation (T&E) of FMs, including quantification of uncertainty, is crucial for nonproliferation applications, where there are unique challenges such as unavailability of all the modalities all the time, unequal distribution of information across modalities, and unequal distribution of annotated data across different modalities.

The paper will present an overview of T&E approaches for FMs, and issues such as computational complexity, scalability and deployability. The paper will also discuss T&E of FMs for computer vision to solve downstream tasks such as land-use, scene and image classification, object detection, localization, and segmentation, which are essential for the characterization of objects and activities of interest. Finally, we will consider an application of transformer models to scene classification using satellite imagery and compare transformers to convolutional neural networks using T&E metrics.

1 Introduction
More of the world is currently under surveillance than at any other time in history, due to open source and commercial high-resolution satellite and aerial surveillance, which can include visible wavelengths, multispectral or hyperspectral sensors. Textual data and time series data of various modalities, e.g., seismic and radio-frequency signals, is also being recorded at an enormous rate. This provides new opportunities for global transparency and innovative approaches to treaty verification. Recent progress in deep learning led to the development of foundation models (FMs), e.g., transformer models, and created a basis to potentially address some data deluge challenges [1].

FMs models are first trained on massive amounts of broad data to learn generic features that can be used for various downstream tasks. Once trained, they can be adapted for specific tasks using smaller task-specific datasets. FMs were initially developed within the context of natural language processing (NLP) tasks and achieved remarkable performance on a wide range of NLP tasks, including language translation, sentiment analysis, text classification, and question answering. Recently, FMs have been broadened to computer vision tasks, such as visual question answering and object detection. For

example, Vision Transformer (ViT) is a FM used for computer vision tasks [2]. Similar to FMs used for NLP tasks, FMs for computer vision tasks are pre-trained on large datasets, e.g., ground-based ImageNet imagery [3], and fine-tuned on task-specific data, e.g., satellite imagery [4].

The field of FMs is rapidly evolving, e.g., a new FM for NLP is released by industry nearly weekly and they have been extended to multi-modal data, such as images and text, video, audio and text [5]. However, FMs also face some challenges. A prodigous computational requirement is one of them. FMs, in particular language FMs, tend to have many more parameters than convolutional neural networks (CNNs). For example, the GPT-3 model developed by OpenAI has 175 billion parameters [6], and Switch Transformer developed by Google has 1.6 trillion parameters [7]. This limits their scalability and makes it difficult to deploy such models on resource-constrained devices.

Many accuracy-based metrics, such as classification accuracy and F1-score, are used to quantify performance of FMs in applications. However, while accuracy-based metrics are important, uncertainty quantification (UQ) in the model's predictions and model calibration are equally important. UQ and model calibration should be an integral part of a decision-making process, especially in high consequence real-world applications.

UQ and model calibration are related, but not identical. UQ refers to the process of estimating the level of uncertainty in a model's predictions. This can include both aleatoric or data uncertainty, and epistemic uncertainty, which arises from the model's lack of knowledge or understanding of the true data generating mechanism. Model calibration, on the other hand, refers to the process of ensuring that a model's predicted probabilities align with the true probabilities of the predicted events. A well-calibrated model will assign probabilities that reflect the true likelihood of the events occurring.

UQ and model calibration are related in the sense that a well-calibrated model will typically have more accurate uncertainty estimates. In other words, a model that is poorly calibrated may overestimate or underestimate the uncertainty in its predictions, which can lead to incorrect decisions in nonproliferation applications where uncertainty plays a critical role. Therefore, it is important to both calibrate a model and quantify its uncertainty in order to ensure accurate and reliable predictions.

In this paper, we review methods for T&E of FMs and perform a case study of T&E of ten deep learning models, including CNNs and transformer models, applied to the task of image scene classification using satellite and aerial imagery. We target post-hoc evaluation of FMs, i.e., the evaluation of trained "black-box" models, when an end user does not have access to the internals of the model yet needs to know if the model is well-calibrated. The paper continues by reviewing, in Section 2, UQ approaches and metrics used for T&E of FMs, their constraints, and open problems. Section 3 summarizes computer vision tasks that are of interest for nonproliferation applications. Section 4 presents results of our case study. Finally, conclusions and directions for future work are presented in Section 5. While our experiments are done in the imaging domain, our results are generalizable and can be applicable to other data modalities.

2 Uncertainty Quantification and Calibration of Foundation Models

Our goal is to develop a T&E framework of FMs for nonproliferation applications. As we work towards our goal, in this section, we examine and summarize approaches that can be used for a variety of T&E tasks, as well as open problems in T&E of FMs. Broadly, UQ methods can be categorized

into two groups: (1) intrinsic, when there is an access to the model's internal structure (e.g., model's nodes and internode weights), and (2) extrinsic, when there is no or limited access to the model's internals, e.g., a pre-trained "black-box" model.

(1) For intrinsic UQ, popular approaches in the machine learning community include:

- *Ensemble methods*: Ensemble methods [e.g., 8] involve training multiple models with different initializations and/or architectures and combining their predictions, e.g., by taking the average or maximum prediction across all models. Variance in predictions across the ensemble can be used as a measure of model uncertainty. Compared to combining models of the same type, combining structurally different models (e.g., CNNs and transformers) can provide a more robust prediction and an estimate of the uncertainty in the prediction. The choice of models to ensemble, the number of models in the ensemble, and the method of combining predictions impacts the performance and confidence of the ensemble model and required computational resources.

- *Bayesian Deep Learning (BDL)*: BDL, including Bayesian neural networks, is a probabilistic approach that incorporates uncertainty into deep learning models [9, 10]. This approach involves placing a prior distribution over the parameters of the neural network, such as internode connection weights and node biases, which can be updated to a posterior distribution via approximate Bayesian inference. The resulting posterior distribution can be used to quantify the uncertainty in the model's predictions. BDL encompasses a variety of methods for inferring posterior distribution, including variational inference [11, 12], drop-out variational inference [13, 14], sampling approaches [15-17], approximate inference based on Stochastic Weight Averaging Gaussian [18, 19], and Laplace approximations [20, 21]. Among drop-out techniques, the Monte Carlo dropout [13] is often used. Instead of only dropping out neural network units during training, dropout is also applied at test time, and multiple predictions are made for a given input, which can be used to calculate the variance and uncertainty. Bayesian methods can also be combined with ensembles, e.g., Bayesian nonparametric ensemble [22]. The computational cost of BDL depends on the specific inference technique used, complexity of the model, availability of hardware accelerators, and the amount of data being used.

(2) For extrinsic UQ, the commonly used methods include:

- *Data augmentation (DA)*: DA involves applying different transformations to the input data during inference and averaging the predictions across the different transformations [23]. We can use DA both at training and test times. The last is known as test-time data augmentation [24, 25], and can be used for post-hoc model's evaluation and improvement. DA helps to estimate the uncertainty in the model's predictions by measuring the variability across the different predictions.

- *Sensitivity analysis (SA)*: SA is related to DA methods and involves adding noise or perturbations to the input data and measuring the impact on the model's predictions and calibration [26, 27]. SA can be used to identify which features have the most impact on the model's prediction [28]. By permuting the values of each feature and measuring the change in the prediction, the importance of each feature can be estimated.

– *Prediction Intervals*: Prediction intervals can provide a range of values that is likely to contain the true value of the prediction. Common approaches are drop-out and bootstrapping. Bootstrapping is a resampling technique that can be used to estimate model uncertainty or assess the stability of a statistical estimate by creating multiple datasets through random sampling with replacement from the original dataset. Bootstrapping can be used for both the train and test datasets, depending on the specific application and the question being investigated.

  Estimation of prediction intervals can be combined with data augmentation [29, 30]. By making predictions on each of augmented versions of the input data and then computing the variance of the predictions, it is possible to estimate prediction intervals associated with the model's predictions. It is important to note that the choice of data augmentation techniques can have an impact on the estimated prediction intervals.

– *Evaluation of the model's calibration*: For a well-calibrated model, the predicted probabilities reflect the true probabilities of the events being predicted. Expected calibration error (ECE) [31, 32] and Brier score [35, 36] are metrics for evaluating the calibration of machine learning models. They are not techniques for calibration themselves. Instead, they are used to assess the accuracy of the predicted probabilities and to identify potential issues with the calibration of the model.

  The ECE is a scalar metric that measures the difference between the predicted probabilities and the actual frequencies of the events being predicted:

$$ECE(B) = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)|, \quad (1)$$

where n_b is the number of probabilities in bin b of the histogram B and N is the size of the dataset,

$$acc(b) = \frac{1}{n_b} \sum_{i \in b} 1(\hat{y}_i = y_i), \quad conf(b) = \frac{1}{n_b} \sum_{i \in b} \hat{p}_i, \quad (2)$$

$\hat{y}_i$ is obtained from the highest probability and $\hat{p}_i$ is the highest probability. Model predictions are partitioned into separate bins b_i (Fig. 1) based on their associated confidence scores.

ECE is closely related to reliability diagrams that represent model calibration by plotting accuracy as a function of confidence (Fig.1) [32, 33]. Reliability diagrams can be helpful for interpreting the ECE and identifying patterns in the model's predictions. ECE is a summary statistic that quantifies the calibration error across all the bins of the reliability diagram.

Both ECE and reliability diagrams are sensitive to the choice of binning. While there are several approaches that try to address the binning choice, such as adaptive calibration error [34] and kernel density estimator, the choice of binning remains an open problem. The other challenge is that all the metrics can be affected by class imbalance.

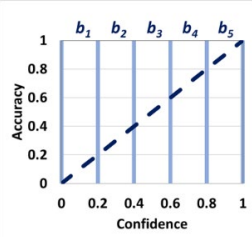

Figure 1. Reliability diagram. A well calibrated model is represented by the diagonal line. Deviation from the diagonal indicates model's miscalibration, such as under-confidence (points above the diagonal) or over-confidence (points below the diagonal) in model's predictions.

The Brier Score (BS) measures the mean squared error between the predicted probabilities and the true labels. For multi-class predictions:

$$BS = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(Z_{ik} - p_{ik})^2, \quad (3)$$

where K is the number of classes, $Z_{ik} = \{0,1\}$ is the indicator variable of class k for observation i, p_{ik} is the predicted probability of observation i to belong to class k. The lower the Brier score the better.

– *Confidence calibration*: It is essential to evaluate the model's calibration, as poorly calibrated models can lead to incorrect decisions. Confidence calibration techniques include temperature scaling, isotonic regression, ensemble calibration, and Bayesian calibration [e.g., 32, 37].

  Confidence calibration is challenging. The open problems include: (1) calibration of the model under data shift, when the statistical properties of the training data and the test data differ, (2) time-varying calibration, when the underlying distribution of the data changes over time, (3) dealing with high-dimensional data, and (4) multi-class calibration.

3 Computer Vision Tasks

As our case study is in the image domain, in this section, we outline computer vision tasks of interest for nonproliferation applications and order them based on the complexity of the analysis they require.

Image classification is an assignment of an image to a certain class or category, such as identifying whether an image contains a nuclear or coal-fired power plant (Figs. 2&3). *Localization* goes a step

Figure 2. Example of image classification (whether the image contains a nuclear power plant or not) and object localization.
This is an image of Nogent-sur-Seine nuclear power plant, France, acquired December 2005.
Image credit © Google Earth.

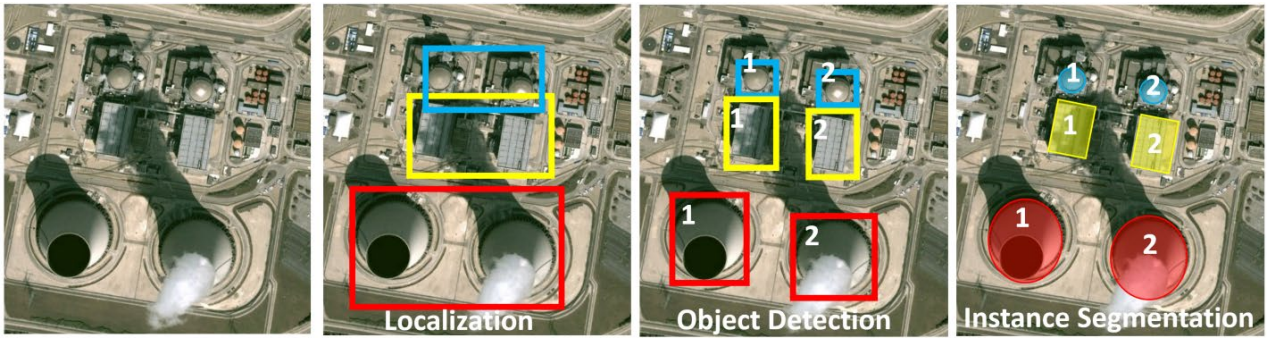

Figure 3. Illustration of localization, object detection and instance segmentation of different components of the nuclear power plant shown in Fig. 2. Red highlights cooling towers, yellow outlines turbines, and blue outlines reactors. Numbers represent IDs of individual instances of cooling towers, turbines, and reactor buildings. Image credit © Google Earth.

further by not only identifying the object of interest but also drawing a bounding box around it to indicate its location in the image. *Object detection* is more complex as it requires localizing multiple instances of different objects in an image, often with overlapping bounding boxes. By detecting each individual object, we can count the number of instances of each object type, such as cooling towers and reactor buildings. *Instance segmentation* goes even further by not only identifying and localizing objects but also segmenting each instance of an object from its surroundings, by assigning a unique class ID and an instance ID to *each pixel in the object of interest*. By segmenting each individual instance of an object in an image, it allows for more accurate measurements of the object's characteristics, such as its size, shape, and location relative to other objects in the scene. In contrast to the previous tasks, *semantic and panoptic segmentations* classify *all the pixels* in an image. Semantic segmentation labels each pixel of an image with a class ID, without differentiating between different instances of the same object class. Panoptic segmentation is the most complex by differentiating different instances and assigning instance ID to each pixel.

These tasks build upon each other, with image classification providing a basic understanding of the content of an image, and instance and panoptic segmentations providing a detailed understanding of the spatial relationships between different objects in the image.

4 Evaluation of the Model's Calibration for Remote Imagery Scene Classification

As an initial case study of T&E of FMs, we present results of the evaluation of the calibration of deep learning models. We consider an application of CNNs and transformer models to remote sensing image classification. We compare the models using conventional metrics (such as classification accuracy) and compare calibration estimates using expected calibration error and Brier score.

For the evaluation, we use the publicly available and well-characterized Remote Sensing Image Scene Classification dataset (RESISC45) [4]. We evaluate ten deep learning models that were either trained from scratch on the RESISC45, or were pre-trained on the ImageNet1K [3] image dataset followed by fine-tuning on the RESISC45 [38]. RESISC45 dataset contains 31,500 images, covering 45 scene categories with 700 images in each category, and with spatial resolution that varies from 0.2 to 30 m per pixel. This dataset was collected over different locations and under different conditions and possesses rich variations in viewpoint, object appearance, spatial resolution, and background. The test dataset is a subset of RESISC45 and consists of 6,300 images with 140 images per category. Included in the categories are thermal power stations, storage tanks, parking lots, harbors, industrial areas, commercial areas, bridges, and airplanes. ImageNet1K is a benchmark in object detection and classification that spans 1,000 object classes and contains ~1.2 million images. The evaluated deep learning models include CNNs and transformers. Among the CNN models are earlier and broadly used models, such as AlexNet [39], VGG16 [40], ResNet50 and ResNet152 [41], DenseNet161 [42], and more recent models, such as EfficientNet [43], MLPMixer [44], and ConvNeXt [45]. FM-like transformer models are represented by Vision Transformer (ViT) [2] and Swin transformer [46].

Our results are summarized in Figs. 4 – 7. First, the results indicate that pre-training does improve classification performance and calibration of all the evaluated models (Figs. 4-6). Second, the transformers perform better than the CNN competitors when they are pre-trained and fine-tuned (Fig. 4). The transformers also exhibit smaller Brier scores and expected calibration errors, suggesting that they are better calibrated than other models (Figs. 5 and 6). However, some of the CNN models, e.g., ResNet50, are only slightly behind in terms of accuracy, Brier score, and ECE. It is interesting to note

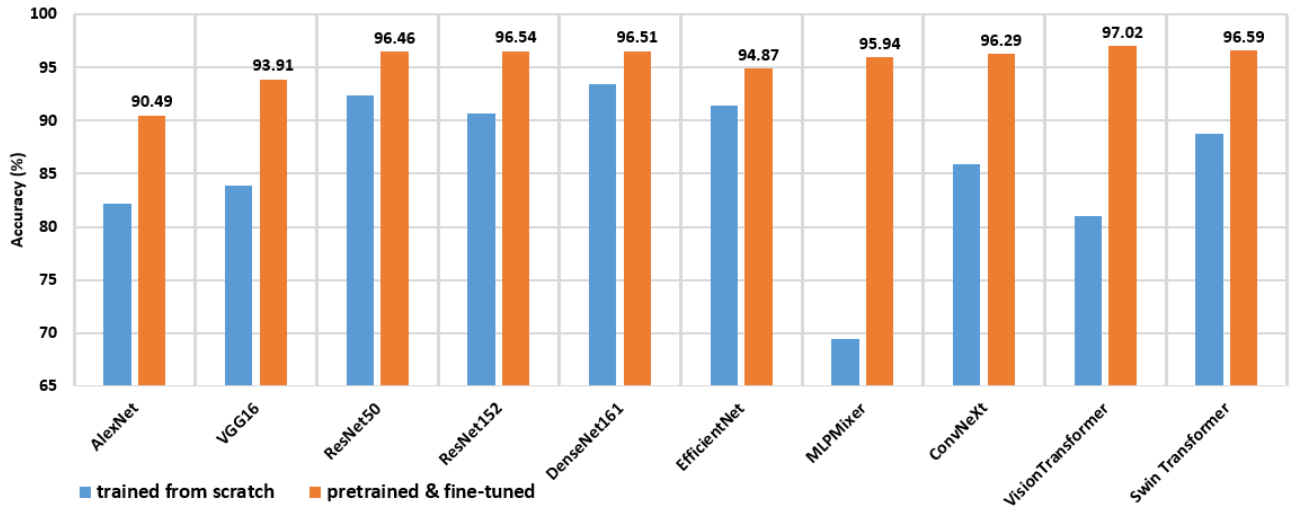

Figure 4. Classification accuracy (%) of the models on the test dataset.

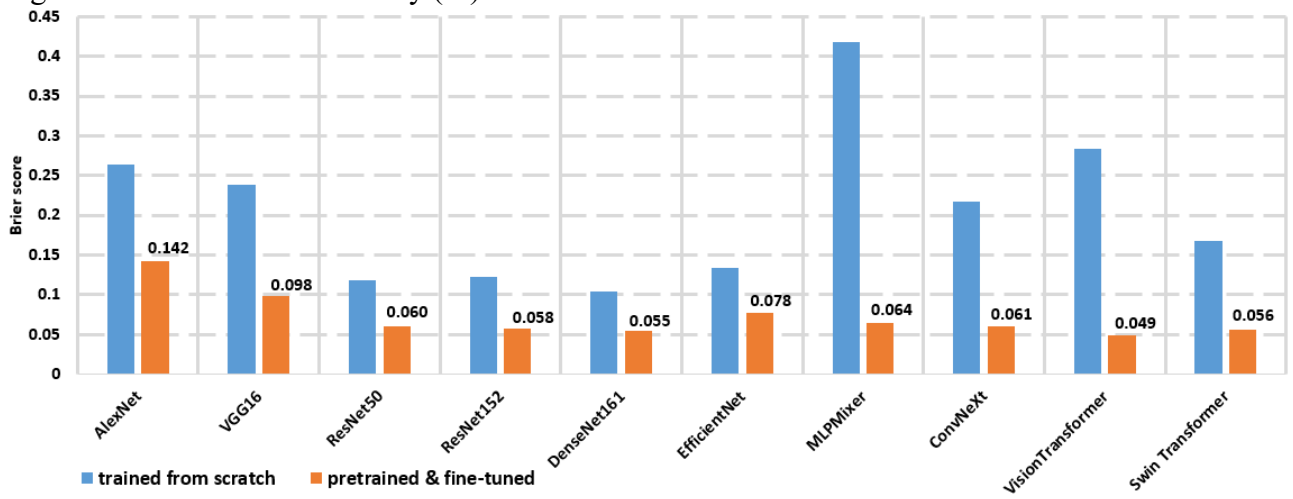

Figure 5. Brier scores for the models trained from scratch and models pre-trained and fine-tuned.

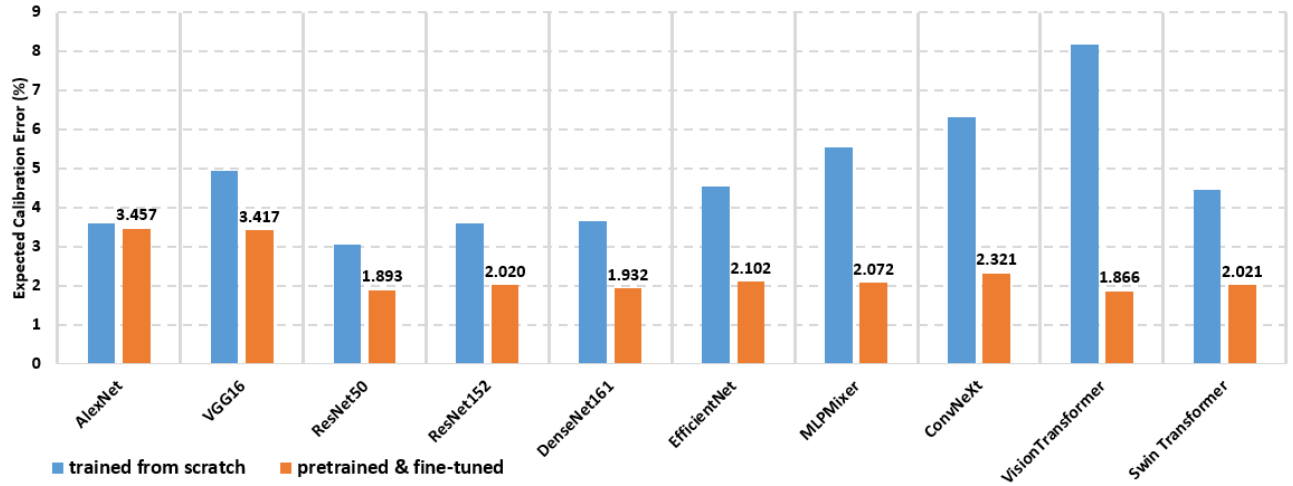

Figure 6. Expected calibration errors (%) for the evaluated models.

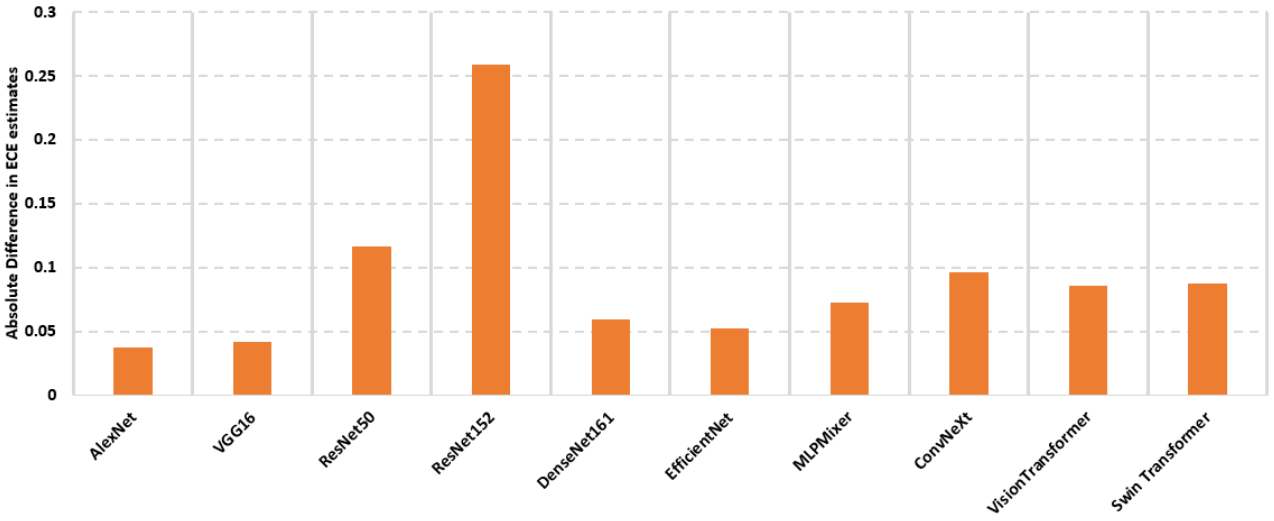

Figure 7. Difference in the ECE estimates using 10 & 30 bins for the pre-trained & fine-tuned models.

that ConvNeXt, a CNN-like model recently designed to challenge transformer models, has slightly lower performance than a more conventional CNN model (ResNet50). Finally, Fig. 7 demonstrates the ECE dependence on the choice of binning, though in our study this is not significant.

5 Conclusions

We present a review of approaches for UQ and calibration of deep learning models, as well as open problems facing these approaches. In the context of our case study, we outline computer vision tasks that are relevant to nonproliferation applications. Our results show that pre-training followed by fine-tuning results in both improved model performance and calibration.

There are a number of ways to expand our study. Next steps can include estimation of prediction intervals using bootstrapping, data augmentation, and sensitivity analysis or correction of model miscalibration using confidence calibration approaches. Going beyond post-hoc T&E will include examination of drop-out approaches, Bayesian deep learning, and ensemble methods.

UQ and deep model calibration are complex fields, with many open problems. The difficulties stem from FMs' complexity, data quality, data distribution shift from training and test data, a possible lack of access to the model's internals, scalability of UQ methods, and data fusion (e.g., combining electro-optical, synthetic aperture and hyperspectral image data), among others. Continued research in UQ of FMs is crucial for improving the reliability of FMs in a wide range of applications.

References

[1] Bommasani, R. et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[2] Dosovitskiy, A. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[3] Russakovsky, O. et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211-252.

[4] Cheng, G. et al. (2017). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, *105*(10), 1865-1883.

[5] Akbari, H. et al. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, *34*, 24206-24221.
[6] Brown, T. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901.
[7] Fedus, W. et al. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, *23*(1), 5232-5270.
[8] Lakshminarayanan, B. et al. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, *30*.
[9] Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision?. *Advances in Neural Information Processing Systems*, *30*.
[10] Clark, A. et al. (2015). Weight uncertainty in neural networks. In: *Int. Conf. on Machine Learning*.
[11] Blei, D. M. et al. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859-877.
[12] Zhang, C. et al. (2018). Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(8), 2008-2026.
[13] Gal, Y, and Ghahramani, Z. (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: *Int. Conf. on Machine Learning*, pp. 1050-1059, PMLR.
[14] Gal, Y., & Ghahramani, Z. (2015). Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*.
[15] Chen, T., et al. (2014). Stochastic gradient Hamiltonian Monte Carlo. In *Int. Conf. on Machine Learning* (pp. 1683-1691). PMLR.
[16] Betancourt, M. (2019). The convergence of Markov chain Monte Carlo methods: from the Metropolis method to Hamiltonian Monte Carlo. *Annalen der Physik*, *531*(3), 1700214.
[17] Hernández, S. et al. (2020). Improving predictive uncertainty estimation using dropout-Hamiltonian Monte Carlo. *Soft Computing*, *24*, 4307-4322.
[18] Maddox, W. J. at el. (2019). A simple baseline for Bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, *32*.
[19] Seckler, H., & Metzler, R. (2022). Bayesian deep learning for error estimation in the analysis of anomalous diffusion. *Nature Communications*, *13*(1), 6717.
[20] Ritter, H. et al. (2018). A scalable Laplace approximation for neural networks. In *6th Int. Conf. on Learning Representations, ICLR 2018-Conference Track Proceedings* (Vol. 6).
[21] Lee, J. et al. (2020). Estimating model uncertainty of neural networks in sparse information form. In *Int. Conf. on Machine Learning* (pp. 5702-5713). PMLR.
[22] Liu, J. et al. (2019). Accurate uncertainty estimation and decomposition in ensemble learning. *Advances in Neural Information Processing Systems*, *32*.
[23] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 1-48.
[24] Kim, I., Kim, Y., & Kim, S. (2020). Learning loss for test-time augmentation. *Advances in Neural Information Processing Systems*, *33*, 4163-4174.
[25] Lu, H. et al. (2022). Improved Text Classification via Test-Time Augmentation. *arXiv preprint arXiv:2206.13607*.
[26] Zhang, J., & Li, C. (2019). Adversarial examples: Opportunities and challenges. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(7), 2578-2593.
[27] Ahmed, U. et al. (2023). Robust adversarial uncertainty quantification for deep learning fine-tuning. *The Journal of Supercomputing*, 1-32.

[28] Covert, I. et al. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33, 17212-17223.
[29] Bahat, Y., & Shakhnarovich, G. (2020). Classification confidence estimation with test-time data-augmentation. *arXiv e-prints*, arXiv-2006.
[30] Li, J. et al. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Int. Conf. on Machine Learning* (pp. 12888-12900). PMLR.
[31] Naeini, M. P. et al. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 29, No. 1).
[32] Guo, C. et al. (2017). On calibration of modern neural networks. In *Int. Conf. on Machine Learning* (pp. 1321-1330). PMLR.
[33] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of Int. Conf. on Machine learning* (pp. 625-632).
[34] Nixon, J. et al. (2019). Measuring Calibration in Deep Learning. In *CVPR workshops* (Vol. 2, No. 7).
[35] Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1-3.
[36] Gneiting, T., & Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359-378.
[37] Ding, Z. et al. (2021). Local temperature scaling for probability calibration. In *Proceedings of the IEEE/CVF Int. Conf. on Computer Vision* (pp. 6889-6899).
[38] Dimitrovski, I. et al. (2023). Current trends in deep learning for Earth Observation: An open-source benchmark arena for image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *197*, 18-35.
[39] Krizhevsky, A., et al. (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
[40] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
[41] He, K. et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 770-778).
[42] Huang, G. et al. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 4700-4708).
[43] Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. on Machine Learning* (pp. 6105-6114). PMLR.
[44] Tolstikhin, I. O. et al. (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems ms*, *34*, 24261-24272.
[45] Liu, Z. et al. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (pp. 11976-11986).
[46] Liu, Z. et al. (2022). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (pp. 12009-12019).