Building Safe and Reliable AI systems for Safety Critical Tasks with Vision-Language Processing

Building Safe and Reliable AI systems for Safety Critical Tasks with Vision-Language Processing

Shuang Ao[1][0000−0003−2648−3082]

The Open University, Walton Hall, Kents Hill, Milton Keynes MK7 6AA
shuang.ao@open.ac.uk

Abstract. Although AI systems have been applied in various fields and achieved impressive performance, their safety and reliability are still a big concern. This is especially important for safety-critical tasks. One shared characteristic of these critical tasks is their risk sensitivity, where small mistakes can cause big consequences and even endanger life. There are several factors that could be guidelines for the successful deployment of AI systems in sensitive tasks: (i) failure detection and out-of-distribution (OOD) detection; (ii) overfitting identification; (iii) uncertainty quantification for predictions; (iv) robustness to data perturbations. These factors are also challenges of current AI systems, which are major blocks for building safe and reliable AI. Specifically, the current AI algorithms are unable to identify common causes for failure detection. Furthermore, additional techniques are required to quantify the quality of predictions. All these contribute to inaccurate uncertainty quantification, which lowers trust in predictions. Hence obtaining accurate model uncertainty quantification and its further improvement are challenging. To address these issues, many techniques have been proposed, such as regularization methods and learning strategies. As vision and language are the most typical data type and have many open source benchmark datasets, this thesis will focus on vision-language data processing for tasks like classification, image captioning, and vision question answering. In this thesis, we aim to build a safeguard by further developing current techniques to ensure the accurate model uncertainty for safety-critical tasks.

Keywords: Deep learning · Model calibration · Uncertainty.

1 Introduction

Despite the impressive performance of AI algorithm in various fields, their safety and reliability is still a concern. Recent studies have achieved successful performance in areas like image [5] and text classification [18], object detection [10], segmentation [9], image captioning [20], visual question answer [8] and graph scene generation [19], and some tasks obtain near-perfect results. However, AI has not been fully deployed in sensitive fields like autonomous driving, medical diagnosing, or assistance for socially vulnerable groups. The major limitation lies in the lack of safeguard in these safety-critical tasks. One shared characteristic of these tasks is their risk sensitivity: it raises serious concern as the mistake

by the AI algorithm can be expensive and even endanger human life. Guidelines from academic papers [2] and industry whitepapers [1] for deploying AI systems for safety-critical tasks include: identifying common causes of failure detection and out-of-distribution (OOD) detection, identifying overfitting in training data, quantifying uncertainty in prediction, and making the model robust to data perturbations. However, these recommendations are not fully satisfied in specific tasks, leading to the limitation of deployment of AI systems in these fields.

One serious limitation of current AI systems is that they tend to give the wrong prediction confidently [16, 12]. Humans feel uncertain when their decision is potentially wrong or ambiguous in the decision-making process, and AI systems are supposed to have similar behaviors. In the recent decade, the quality of network architectures has significantly improved by utilizing deeper and wider networks such as VGG [15] and ResNet [5]. These state-of-the-art networks significantly improve feature learning for text and image but also raise the question of models with poor uncertainty quantification and being over-confident. It refers to when the overall confidence score is higher than the overall accuracy for testing data. Specifically, the model is supposed to show low confidence when the prediction is ambiguous or likely to be wrong and vice versa. The over-confident issue leads to the concern of inaccurate uncertainty quantification and trustworthiness of predictions.

The confidence and accuracy level of the system should match so that human experts can tell when the system tends to make mistakes. Hence addressing the over-confidence issue is essential to building a safe and reliable AI system. Even though a deep learning model can output the prediction for trained tasks, it cannot provide feedback about the quality of its prediction. For example, which class is poorly performed or if the overall output is reliable or not. In other words, such quality refers to how doubtful or uncertain the model for its prediction is, known as the model uncertainty. Ideally, when the model uncertainty is high, the model should suggest a second opinion and defer the task to human experts to re-examine it. With human intervention, the unexpected behaviour or wrong predictions from the AI system can be prevented. This process is crucial for safety-critical tasks as it can enhance failure detection, meaning a model to detect its own wrong predictions during the deployment or real-time applications without checking with the ground truth. Hence precise quantification and sufficient improvement of model uncertainty lie in the heart of building a safe and reliable AI.

In practice, many safety-critical tasks require multi-modality processing, especially with vision and language data. For example, autonomous driving systems process images, audio from user's input, and also signal data from sensors. Models for medical diagnosing deal with the image data such as MRI and text data of patients records. As real-world applications are more complicated than single modality data processing such as solely image or text classification, I will research reliable multi-modal data processing of image and text, and possibly other data sources.

2 Research Question

The main goal of my research is to build a safeguard for vision-language processing, by developing techniques and learning strategies to improve estimates of model uncertainty. RQ1.2 and RQ 2.1 will be discussed at the Doctoral Consortium.

RQ1: Can model uncertainty quantification be improved without adding additional computational complexity to build safe AI systems?

- **RQ1.1: Can we improve upon the uniform distribution in Label Smoothing by generalizing the soft label in a more reasonable way for different applications?**
  The traditional label smoothing clips the hard label into uniformly distributed soft labels, but not the case in practice. Hence we need to tackle this issue for a more accurate soft label generalization.
- **RQ1.2: How to efficiently conduct automatic failure detection (FD) for identify model's own wrong prediction during inference time?**
  FD is a significant criteria for the trustworthiness of a model.
- **RQ1.3: For curriculum learning, can we rank difficulty of samples more accurately to build the framework of curriculum learning?**
  We will use the model confidence as a proxy for ranking the difficulty of training samples.

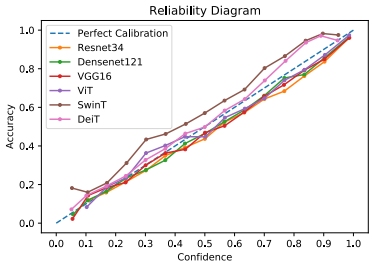
RQ2: By integrating the techniques of safe AI systems developed in RQ1, can we improve the reliability and robustness of the model in the application of vision-language processing?

- **RQ2.1: For vision-language processing, how to build an end-to-end pipeline to reduce the dependency on the prior procedure, as well as the computational cost?**
  Tasks like image captioning and VQA include several processing steps such as object detection and feature extraction. Hence it is necessary to reduce the dependency of the prior step, to reduce its influence to latter processing.

3 Preliminary Result

We investigate 3 CNNs (ResNet34 [5], DenseNet121 [6] and VGG16 [15]) and 3 transformers models (ViT [4], SwinT [11] and Deit [17]) for ImageNet [14] and CIFAR100 [7] dataset. The evaluation metrics are accuracy and Expected Calibration Error (ECE) [13], and ECE is the metric to measure calibration. In figure 1, the table shows that models that perform better in accuracy are not always better in ECE, such as VGG16 and SwinT, which is illustrated further in the reliability diagram in 1 (A). Hence it is essential to build a model with good calibration and high accuracy to improve the trustworthiness and reliability.

The Risk Coverage Curve (RCC) demonstrates the efficiency of automatic failure detection (AFD) as shown in Figure 2. We utilize the predictive uncertainty to distinguish correct from incorrect samples by following [3]. The data



(A)

Models	ECE(%)	Accuracy(%)
ResNet34	4.0	71.2
DenseNet121	3.0	71.8
VGG16	3.3	69.9
ViT	2.2	83.3
SwinT	8.7	84.4
DeiT	6.4	81.3

(B)

Fig. 1. Left (A): Reliability diagram for ResNet34, DenseNet121, VGG16, ViT, SwinT, and DeiT models with ImageNet dataset. Right (B): Results of accuracy and ECE of various model architectures on ImageNet dataset. All results are shown in percentages. The lower the ECE the better.

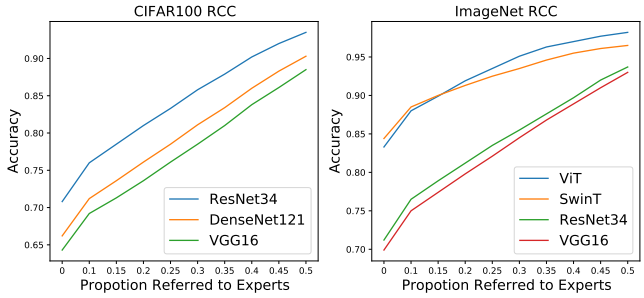


Fig. 2. Risk Coverage Curve for CIFAR100 and ImageNet dataset of CNNs and transformers. The x-axis indicates the percentage of data removed from the entire test set, and the accuracy is calculated on the remaining test set. The higher accuracy means better performance of automatic failure detection.

belonging to "Referred to experts" are wrongly predicted. Comparing ViT and SwinTrans models in the ImageNet RCC, SwinTrans has higher accuracy with the entire test set, but ViT outperforms it with more wrong predictions removed. It suggests that ViT is more reliable than SwinT as the model can detect more wrong predictions.

4 Conclusion

In this thesis, I will explore the topic of safe and reliable AI, focusing on the applications in vision-language processing, such as image captioning and visual question answering. The main goal of my research is to build a safeguard for safety-critical tasks with multi-modal data processing.

References

1. Aptiv, A., Apollo, B., Continenta, D., FCA, H., Infineon, I.V.: Safety first for automated driving. In: Continental, Daimler, FCA, HERE, Infineon, Intel, and Volkswagen, pp. 1–157. White Paper (2019)
2. Ashmore, R., Calinescu, R., Paterson, C.: Assuring the machine learning lifecycle: Desiderata, methods, and challenges. ACM Computing Surveys (CSUR) **54**(5), 1–39 (2021)
3. Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. Advances in Neural Information Processing Systems **32** (2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
7. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
8. Li, W., Sun, J., Liu, G., Zhao, L., Fang, X.: Visual question answering with attention transfer and a cross-modal gating mechanism. Pattern Recognition Letters **133**, 334–340 (2020)
9. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A.: H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging **37**(12), 2663–2674 (2018)
10. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
11. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
12. Müller, R., Kornblith, S., Hinton, G.E.: When does label smoothing help? Advances in neural information processing systems **32** (2019)
13. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)

17. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
19. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–685 (2018)
20. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13041–13049 (2020)