Gradient-Based Localization and Spatial Attention for Confidence Measure in Fine-Grained Recognition using Deep Neural Networks (Abstract)

Charles A. Kantor,[1,2,3] Léonard Boussioux,[1,3,6] Brice Rauby,[3,5] and Hugues Talbot[3,4]

[1]KLASS, AI Research (AIR)*,
[2]MILA, Quebec Artificial Intelligence Institute, Montreal, QC, Canada,
[3]Paris-Saclay University, (ECP) Ecole CentraleSupélec Paris, Greater Paris, France,
[4]INRIA, Greater Paris, France, [5]Polytechnique Montréal, QC, Canada
[6]Operations Research Center, MIT, Cambridge, MA, USA,

ckantor@fas.harvard.edu

Abstract

Both theoretical and practical problems in deep learning classification benefit from assessing uncertainty prediction. In addition, current state-of-the-art methods in this area are computationally expensive: for example, (Loquercio, Segu, and Scaramuzza 2020) is a general method for uncertainty estimation in deep learning that relies on Monte-Carlo sampling. We propose a new, efficient confidence measure later dubbed Over-MAP that utilizes a measure of overlap between structural attention mechanisms and segmentation methods. It does not rely on sampling or retraining. We show that the classification confidence increases with the degree of overlap. The associated confidence and identification tools are conceptually simple, efficient and of high practical interest as they allow for weeding out misleading examples in training data. Our measure is currently deployed in the real-world on widely used platforms to annotate large-scale data efficiently.

Introduction

The vast majority of deep learning systems today operate mostly as black-boxes (Castelvecchi 2016). Reasons for this include a large number of parameters; the considerable amount of data necessary for training; the practical difficulties of curating the training data to ensure that all relevant cases are included; sensitivity to noise, poor annotations, adversarial attacks; variation in input; and more.

Besides, deep learning systems operate in an uncertain world that is very different from the policed variability introduced in most benchmarks. As a trivial example, systems trained on ImageNet can only recognize elements within the thousand of its training classes. While performance may be excellent within that set, it falls to zero for any class element missing in the training set. Deep networks would be more useful with some metric that tells the user how confident in their predictions they are (Osband 2016). In this way, if a network is given as input something that they have never trained on, they might reply that the input is unknown (Blundell et al. 2015). If the input is corrupted, ambiguous, or noisy, the network should also reply that it is hesitant to conclude. This ought to be done without training the network on every possible contingency, which is impossible by definition (Kendall and Gal 2017).

Reporting confidence and uncertainty is critical in many systems. Diagnostic systems, for example, require knowing how reliable a reported classification is (Fauw et al. 2018). This is even more so the case of fine-grained classification since reliable classification must often rely on tiny details (Soni, Shah, and Moore 2020).

Contribution

Small details are usually overwhelmed by a rich surrounding context. Grad-CAM (Selvaraju et al. 2017) uses reverse gradient propagation as an auditing tool to check whether networks were likely to pay attention to the background instead of focusing on the object or region of interest. Factorizing the object of interest and feeding the resulting picture as a prior to the classification network could improve performance. We built an automated particularizing algorithm to cut down the background. However, merely focusing on foreground objects may induce some limitations, such as neglecting the spatial conjunction between the region of interest and inner parts. Jointly employing attention models is required to exploit subtleties and local differences.

We used an attention module for feed-forward convolutional neural networks (CNN), end-to-end trainable. We chose this architecture for its good classification performance and its ability to separate the spatial attention mask from the channel attention. In the same vein as class activation mapping, we built attention before data classification to avoid getting large background noise areas of lower relevance and further reducing overlap.

During object detection, we use Mask R-CNN (He et al. 2017), which provide several classifications at various locations. The detected bounding boxes typically spread much wider than the true objects of interest. As such, object proposals around these regions are assumed to contain one or

*KLASS AI Research (AIR), www.klass.global

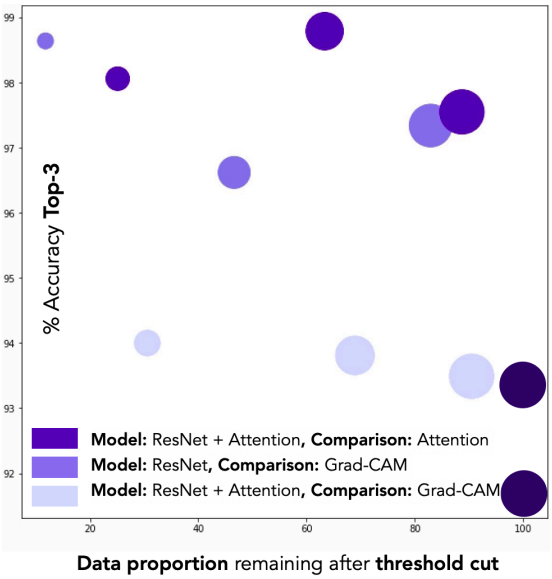

Figure 1: Influence of rejecting uncertain classification on the Top-3 macro test accuracy score by different methods. Disk diameters are proportional to data remaining. Darkest disks represent baseline accuracy without driven rejection.

several objects of interest for later classification. We trained a deep classification neural network building class activation maps for pixel label refinement.

We thus generate customed segmentation masks. We annotated it for pre-training ourselves and qualitatively assessed the segmentation performance, sufficient to be used in our customed *guided attention* context and for *uncertainty prediction* assessment.

Visual saliency (Malik and Perona 1990) detects areas of images that are perceived to be important by humans. Recently, deep-learning-based methods have provided the state of the art (Li and Yu 2015). In our work, as a generalization, we can use saliency detection as a prior to segmentation to highlight areas of interest in our images.

Classical deep learning models used in classification are generally confident in their predictions, even when they are incorrect (Goodfellow, Shlens, and Szegedy 2015). To ameliorate this, we assess uncertainty and use it to reject predictions below a given confidence threshold. For each image, we propose to use the overlap of several possible combinations of our generated masks. A stage-by-stage comparison starts with the intersection analysis between the CAM-extracted maps and binarized R-CNN-maps. We also measure the overlap between attention mechanism extractions.

For this measure, predictions are rejected if the overlap is below a parametric threshold. A low overlap value can be interpreted as the network likely basing too much its prediction on regions outside of the region of interest, i.e., the background and foreground. However, the underlying assumption of this overlap measure is the accuracy of the *guided segmentation*. When it failed to segment our images, we rendered *estimation not available*, meaning that no mask is available, and thus, attention-based overlap is set to 0.

Conclusion

From the results, effecting only a small amount of selection (12-18%) results in improved accuracy scores by 6-10 percentage points, which is very significant. Accuracy keeps rising with an increased level of selection beyond 18%, but to a much lesser degree.

We conclude that creating Grad-CAM overlap with *guided segmentation* as a confidence measure allowed us to weed out ambiguous or noisy samples from the *training* and *test* dataset and that once these samples were removed, performance remained at a high level.

References

Blundell, C.; Cornebise, J.; Koray, K.; and Wierstra, D. 2015. Weight uncertainty in neural networks. *International Conference on Machine Learning* 1613–1622. ArXiv preprint arXiv:1505.05424.

Castelvecchi, D. 2016. Can we open the black box of AI? *Nature News* 538(7623): 20.

Fauw, D.; et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* 24(9): 1342–1350.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. ISSN 2380-7504. doi:10.1109/ICCV.2017.322.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5455–5463.

Loquercio, A.; Segu, M.; and Scaramuzza, D. 2020. A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters* 5(2): 3153–3160.

Malik, J.; and Perona, P. 1990. Preattentive texture discrimination with early vision mechanisms. *JOSA A* 7(5): 923–932.

Osband, I. 2016. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *NIPS Workshop on Bayesian Deep Learning*, volume 192.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Soni, R.; Shah, N.; and Moore, J. D. 2020. Fine-grained Uncertainty Modeling in Neural Networks. *arXiv preprint arXiv:2002.04205* .