# Over-MAP: Structural Attention Mechanism and Automated Semantic Segmentation Ensembled for Uncertainty Prediction

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

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

ckantor@fas.harvard.edu leobix@mit.edu

## Abstract

Both theoretical and practical problems in deep learning classification require solutions for assessing uncertainty prediction but current state-of-the-art methods in this area are computationally expensive. In this paper, we propose a new confidence measure dubbed Over-MAP that utilizes a measure of overlap between structural attention mechanisms and segmentation methods, that is of particular interest in accurate fine-grained contexts. We show that this 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; Achille, Paolini, and Soatto 2019), meaning that it is difficult to track how and why such systems make decisions. Reasons for this unfortunate state of affairs include the large number of parameters in such networks; the considerable amount of data necessary for training these networks; the practical difficulties of curating the training data to ensure that all relevant cases are included; sensitivity to noise, poor annotations and adversarial attacks (Jin, Dundar, and Culurciello 2015); variation in input; and much more.

Besides, deep-learning and, more generally, AI 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.

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



## Contribution

For deep networks to become truly useful, they need to come 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). In short, these networks need to be aware of what they do not know (Nalisnick et al. 2018). 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. For example, diagnostic systems require knowing how reliable a reported classification is (De 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).

In this work, we advance a new confidence measure called Over-MAP that utilizes structural attention mechanisms, visual explanations from deep networks and segmentation methods, of particular interest to accurate fine-grained classification. We apply our methodology for real-world problems starting with large scale crowdsourced and collected data. Also of practical interest in the field of medical imaging, we specifically applied our uncertainty prediction tool in the context of wildlife analytics. Our deep learning methods offer opportunities for population monitoring. We developed accurate computer vision algorithms and proposed fine-grained classification innovations, now in deployment process on global platforms, to encourage the model to focus on areas of an image that are salient for identification.

## Related Work

Deep Neural Networks often provide good estimators for prediction tasks; however, producing a reliable, stable and trustworthy result is complicated and sometimes elusive. Establishing trust and explainability, the latter being a network's capacity to explain its results, is a way forward. In fine-grained classification, the problem is compounded

by the small differences between classes (Soni, Shah, and Moore 2020), making uncertainty more challenging to estimate and report.

**Additive Features Attributions** Lundberg and Lee (2017) point out that in the choice of a compromise between accuracy and explainability, the former often wins out because it is more easily quantifiable. Conversely, an accurate result that cannot be explained is often considered suspicious in critical domains like medical imaging. To address this problem, the authors present a unified framework for interpreting predictions by identifying a new class of additive feature importance measures and theoretical results showing a unique solution in this class with a set of desirable properties.

**Using Predictive Distribution Information** Koh and Liang (2017) use influence functions, a classical technique from robust statistics, to trace a mode prediction through the learning algorithm and back to its training data, thereby identifying training points most responsible for a given prediction. On linear models and convolutional neural networks, they demonstrate that influence functions are useful for multiple purposes: understanding model behavior, debugging models, detecting dataset errors, and creating visually indistinguishable training-set attacks.

To achieve the same goal of a reliable estimated result, estimating the uncertainty in prediction is a related but different approach. Indeed Feinman et al. (2017) seek to reduce uncertainty by actively using adversarial examples. Sensoy, Kaplan, and Kandemir (2018) design a predictive distribution for classification by placing a Dirichlet distribution on the class probabilities and assigning neural network outputs to its parameters. They fit this predictive distribution to data by minimizing the Bayes risk with respect to the L2-Norm loss which is regularized by an information-theoretic complexity term. The resultant predictor is a Dirichlet distribution on class probabilities, which provides a more detailed uncertainty model than the point estimate of the standard softmax-output deep nets. This approach is related to the use of Gaussian Processes in machine learning (Rasmussen 2003), which are at their heart an interpolation method providing a variance estimate throughout the domain of interpolation. Note that providing a measure of uncertainty is not the same as providing explainability.

**Uncertainty Characterization** Kaplan et al. (2018) point out that machine-learning systems make mistakes that occur when the trained systems operate in situations beyond their training domain. Instead of forcing a decision, it is crucial for ML systems to characterize a degree of uncertainty relative to the similarity of the current observations to the training data. It is also important and to explain the uncertainty to a human decision-maker. This way, the ML system can alert its users when it realizes that it can no longer provide quality inferences. Their paper uses subjective Bayesian networks and evidential neural networks to achieve this uncertainty awareness iteratively.

**Visualization** Finally, as a powerful approach to both interpretability and reliability, visualization of trained network activity has proved useful. Zeiler and Fergus (2014) propose to visualize the activity within a trained model. Their approach reveals the trained features to possess intuitively desirable properties such as compositionality, increasing invariance and class discrimination as they explore increasingly abstract layers. They also demonstrate through a series of occlusion experiments that a standard, trained deep-learning model is highly sensitive to local structure in the image and does not just use a broad scene context. In particular, even small occlusions can significantly change classification results.

Related to this approach, Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al. 2017) has proved particularly useful and seminal since it allows visualization of the relative importance of weights in network decision-making by back-propagation of the last convolutional layer through the entire network. In Over-MAP, a step towards uncertainty assessment propose to combine Grad-CAM back-propagation and a measure of overlap with a segmented region to obtain a novel notion of confidence, depending on the amount of weighted overlap between the two.

## Organizational Approach

*Our motivation is to provide simple and efficient uncertainty measures for deep classification networks, in particular in the context of fine-grained classification.*

### Efficient Vision Model

In the context of fine-grained classification, small details are usually overwhelmed by a rich surrounding context.

**Gradient-Boosting Visualization** We initially implemented Grad-CAM using reverse gradient propagation as an auditing tool to check whether our networks were likely to pay attention to the background instead of focusing on the object or region of interest. Leveraging visual modalities by focusing on discriminative segments is key, to benefit the most from sparse visual cues. Thus, factorizing the object of interest and feeding the resulting picture as a prior to the classification network could improve performance. We propose to build 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 its parts. Jointly employing attention models is required to exploit subtleties and local differences.

**Detection Phase** During object detection (as opposed to segmentation), we use Mask R-CNN, which provides several classifications at various locations. Detected bounding boxes typically spread much wider than the true objects of interest. As such, object proposals around these regions are hypothesized to contain one or several objects of interest for later classification. We trained a deep classification neural

network building class activation map, using Mask R-CNN for pixel labels refinement.

**Instance Segmentation** To improve our fine-grained classification, we thus generate segmentation masks. We also used a Mask R-CNN (He et al. 2017) network pre-trained on the COCO instance segmentation dataset (Lin et al. 2014) and fine-tuned it on a small subset of our dataset. This approach is possible because for our objects of interest (wildlife analytics), the segmentation task is very similar to segmenting everyday objects present in natural images and therefore, pre-training was very efficient. We annotated a small subset (10% of the total data) used for pre-training ourselves and we qualitatively assessed the segmentation performance. The segmentation results obtained were good enough to be used in the *guided attention* context and for *uncertainty prediction*.

**Generalization: Prior Neural Knowledge Using Visual Saliency** Visual saliency is a concept in computer vision (Malik and Perona 1990) that relates to areas of images that are perceived to be important by the human visual system. Visual saliency detection algorithms associate a high value to such areas. Recently, deep-learning-based methods have provided the state-of-the-art for saliency detection (Li and Yu 2015). In our work, we can use saliency detection instead of segmentation to highlight areas of interest in our images.

**Attention Module for Feed-Forward Deep CNNs** As the gist of our approach is to leverage attention mechanisms, we used an attention module for feed-forward convolutional neural networks (CNN), end-to-end trainable. We chose this architecture for its good classification performances on several benchmarks 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, synonym of data-redundancy.

## Overlap-Based Precision Estimation

**Motivation for Over-MAP** One limitation of most deep learning models is their incapacity to assess their uncertainty. Indeed, classical deep learning models used in classification are generally very confident in their predictions, even when they are incorrect (Goodfellow, Shlens, and Szegedy 2015). This is unfortunate, so it can be interesting to assess uncertainty and use it to reject predictions below a given confidence threshold. A benefit of such confidence measurement is that high-confident predictions might require less expert confirmations to be considered reliably annotated data.

**Semantic Process** To that end, for each image, we propose a new confidence measure based on the overlap of several possible combinations of masks. We propose a stage-by-stage comparison starting with the intersection analysis between the CAM-extracted saliency masks and binarized R-CNN-extracted segmentation masks. We also measure the overlap between attention mechanism masks and binarized R-CNN-extractions.

**Threshold-Based Rejection** For this measure, predictions are rejected if the overlap is below a parametric threshold - we provide three specific values to illustrate their impact and we suggest guidance for managerial choice. 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. However, the underlying assumption of this overlap measure is the accuracy of the Mask R-CNN segmentation. When it failed to segment our images, we rendered *estimation not available*, meaning that no mask is available, and therefore, the overlap with the attention masks will be 0.

# Experiments

In this section, we describe our overall pipeline including the classification and segmentation process. Specifically, a foremost automated segmentation is designed to leverage the shape-based prior knowledge of pictures. Besides, we generate independently saliency maps from both feed-forward attention CNNs and gradient-based localization. Finally, we introduce the threshold-based rejection measure to weed out specific sightings.

Through our collaborations, we gathered sightings with date and time information. In this paper, we worked with 1000 annotated labels (level of hierarchy). To evaluate our algorithms, we experiment with a fine-grained subset of 100,000 labeled pictures. We split the data into a train (80%) and test (20%) set.

## Optimization and Training

For the classification task, we used a Residual Neural Network augmented with an attention module for feed-forward convolutional neural networks (CNN) pre-trained on Imagenet (Deng et al. 2009). We ran our models with varying batch sizes. We decayed the learning rate factor every 20 epochs. We optimized this steady number through experiments. To prevent over-fitting, we set a weight decay parameter and we add a dropout layer (Srivastava et al. 2014) before the last fully-connected layer. Data-augmentation comprises random rotation, flipping, rescaling and cropping during training. The best weights on the validation set were saved and training was interrupted when no improvement occurred for more than 50 epochs. To obtain preliminary results without changing the class-balancing parameters, we worked with a balanced, reduced by 90% dataset version that we created, containing the 100 most common labels (focusing on hierarchical level of interest), with a same number of samples. We noted that including our module for feed-forward CNNs boosts level-oriented classification performance by 2% on the test set. Models are all trained within 12 hours.

| Performance Accuracy | Customized Attention Model (Top-1 acc.) | Customized Attention Model (Top-3 acc.) | Hierarchical Recovered Family | Hierarchical Recovered Genus |
|---|---|---|---|---|
| Micro acc. | 54.31 (0.70) | 72.93 (0.45) | 84.68 (0.40) | 72.14 (0.43) |
| Macro acc. | **81.29** (0.61) | **92.81** (0.31) | **97.64** (0.52) | **91.99** (0.23) |

Table 1: Foreword generated baseline CNN (before Over-MAP) model performance with and without CNN Attention Module and Guided-Attention. We provide the average accuracy obtained over 3 different seeds and the standard deviation between parenthesis. Best accuracies are in bold. We report macro accuracy, which is the total number of correct observations over the total number of observations, and micro accuracy, which is the average performance of each class. Top-1 and Top-3 refer to the fraction of test images for which the correct label is among the one or three labels considered most probable by the model.

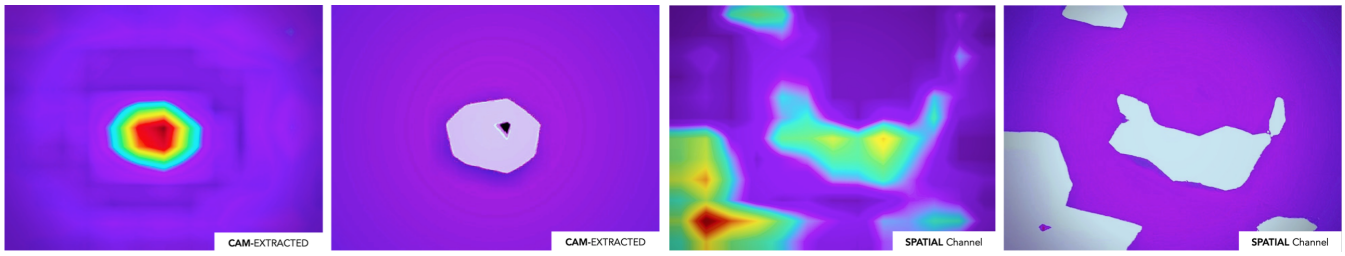


Figure 1: 1. Saliency heatmap *CAM-generated* before metamorphosis *(left)*. Metarmorphosed map, fashioned using delineation for selected relevant area *(CAM)*. 2. Saliency heatmap *channel-generated* before metamorphosis *(left)*. Metarmorphosed map, fashioned using delineation for selected relevant area *(Spatial Channel)*.

## Extraction in Probability Map

Activations are taken as the pixel-wise probabilities of the corresponding class. In this stage, CAM is adopted to collect our picture's saliency map to localize the region of interest. The saliency map indicates the representative regions used by the CNN to identify the object class. In our process, we decide to extract salient object regions of images obtained by performing a binarization and connectivity area extraction on the saliency maps. We performed the same selection to generate attention mechanism maps, keeping only the most *intense* areas (see Figure 1). In our process, we preserve the areas encircled by our dotted lines, visually corresponding to the RGB values of interest.

For each image, an overlap percentage is computed using the *Sørensen–Dice* (DSC) and *Intersection over Union* (IoU) coefficients. We set $X_1$ as the ensemble of pixels kept by convolutional-block translated maps and $X_2$ as the ensemble of pixels kept by CAM-extraction.

$$DSC := \frac{2|X \cap Y|}{|X| + |Y|} \quad and \quad IoU := \frac{|X \cap Y|}{|X \cup Y|}$$

where $Y$ is the map generated from the segmenting network. The Sørensen index equals twice the number of elements common to both sets divided by the sum of the number of elements in each set. As compared to Euclidean distance, the Sørensen-Dice distance retains sensitivity in more heterogeneous data sets and gives less weight to outliers. Both coefficients equal zero of there is no overlap between the masks, and 1 if the masks are identical. From a Machine-Learning perspective, they both ignore true negatives, i.e., they are insensitive to the size masks' size compared to that of the background.

## Distribution Analysis

We now consider the overlap distribution for the whole *test set*. We select several thresholds on the distribution curve to determine a relationship between a rejection ratio and a level of precision. We introduce and compute the probability of getting a correct class prediction given that images taken into account yield an overlap greater than the varying threshold (see left column of tables 2 and 3). In Table 2, the overlap is computed by considering attention module for feed-forward CNNs together with automated segmenatation. Therefore, it can only be applied to module-augmented DNNs. We compare the performance with our standard score giving full accuracy. In Table 3, the overlap is computed between the CAM-generated saliency maps and those from the automated segmentation, which can be applied to a broader range of networks since an additional attention module is not used, and thus not necessary for this second context. We also provide in Figures 2 and 3 a scatter plot representation of *macro test accuracies* from *top-1* and *top-3* columns of the tables. *Top-1* and *top-3* refer to the fraction of test images for which the correct label is among the one or three labels considered most probable by the model. We report macro accuracy, which is the total number of correct observations over the total number of observations. Upper right points are those of interest, showing an improvement in accuracy with a high percentage of the data remaining.

| Overlap threshold | Model | Top-1 acc. | Top-3 acc. | Data proportion remaining after threshold cut |
|---|---|---|---|---|
| 0% | Attention-augmented | 80.95 (0.45) | 93.35 (0.20) | 100% |
| 20% | Attention-augmented | 80.92 (0.55) | 93.49 (0.22) | 90.47 (3.09)% |
| 40% | Attention-augmented | 81.80 (0.67) | 93.81 (0.29) | 68.93 (8.64)% |
| 60% | Attention-augmented | 83.05 (1.25) | 94.00 (0.38) | 30.6 (9.97)% |

Table 2: Model performance on selective dataset subset considering the overlap between spatial attention masks and masks issued from Mask R-CNN segmentation. Results have been run with 3 different seeds. We provide the mean and the standard deviation in parenthesis.

| Overlap threshold | Model | Top-1 acc. | Top-3 acc. | Data proportion remaining after threshold cut |
|---|---|---|---|---|
| 0% | Standard | 79.54 (0.70) | 91.72 (0.49) | 100% |
| 0% | Attention-augmented | 80.95 (0.45) | 93.35 (0.20) | 100% |
| 20% | Standard | 87.09 (0.58) | 97.34 (0.33) | 82.9 (0.14)% |
| 20% | Attention-augmented | 90.52 (0.77) | 96.62 (0.20) | 88.69 (0.91)% |
| 40% | Standard | 90.55 (0.51) | 98.64 (0.20) | 46.64 (0.52)% |
| 40% | Attention-augmented | 88.32 (0.82) | 97.55 (0.39) | 63.34 (1.07)% |
| 60% | Standard | 91.87 (0.96) | 98.79 (0.48) | 11.8 (0.24)% |
| 60% | Attention-augmented | 89.44 (1.26) | 98.06 (0.67) | 25.13 (2.45)% |

Table 3: Model performance on selective dataset subset considering the overlap between masks issued from CAM-generated maps and masks issued from Mask R-CNN segmenting network. Results have been run with 3 different seeds. We provide the mean and the standard deviation between parenthesis.

## Discussion

For a given overlap threshold, accuracy improves more using Grad-CAM masks than attention masks (Tables 1-2). However, attention masks can advantageously be computed directly during the inference without reference, whereas Grad-CAM requires a back-propagation pass after each inference.

### Uncertainty Assessment

From the results on Table 2, we see that accuracy scores are improved using a thresholded map based on attention module version. Moreover, the more we select samples with high overlap, the higher the score. However, measurable improvement only occurs for high levels of selection: +2% in Top-1 accuracy if we discard about 70% of the samples.

Given the limited amount of improvement, in that setting we cannot recommend using this technique as a confidence measure.

The results from Table 3 using our proposed Grad-CAM overlap measure tell a different story. In this case, effecting only a small amount of selection (12-18%) results in improved accuracy scores by 6-10 percentage points, which is more significant. Accuracy keeps rising with an increased level of selection beyond 18%, but to a lesser degree.

We conclude that using Grad-CAM overlap with a segmentation mask 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.

### Recommendation

Given the continuity of our approach, we recommend using our approach whenever a binary segmentation map of the input data can be sufficiently easily derived. We can wonder what happens if the automated segmentation is not correct. This is in fact not a problem. Even a mask which is incomplete or incorrect to some degree helps with the classification by focusing attention to the salient part. This approach might be possible when the segmentation task is sufficiently similar to the task of segmenting other objects present in a common dataset, and therefore, when pre-training might be very effective. As a solution, our high performance in high level labels classification also narrows down the manual annotation by experts.

Our approach's critical element seems to be the quality of the segmentation on the one hand, and the method used to highlight the network's attention on the other. The use of Mask R-CNN or U-Nets is currently a reasonable approach for segmentation but may not be optimal. We

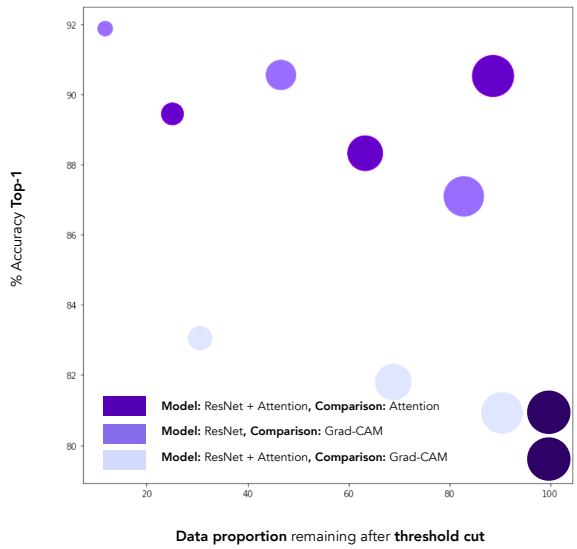


Figure 2: Influence of rejecting uncertain classification on the Top-1 macro test accuracy score by different methods, with results from Tables 2 and 3. Disk diameters are proportional to data remaining. Darkest disks represent baseline accuracy without overlap-driven rejection.

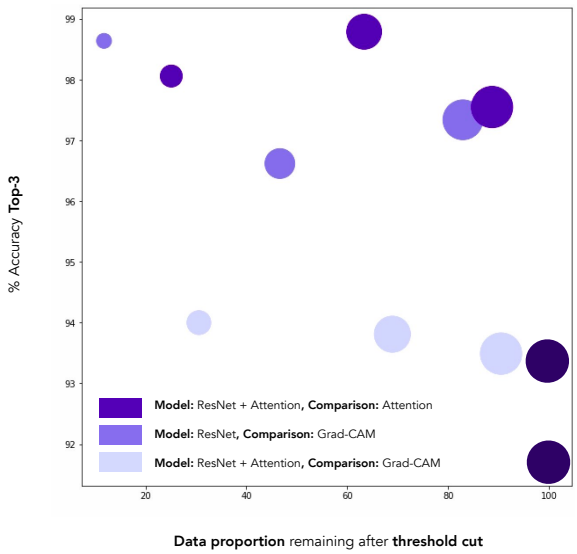


Figure 3: Influence of rejecting uncertain classification on the Top-3 macro test accuracy score by different methods, with results from Tables 2 and 3. Disk diameters are proportional to data remaining. Darkest disks represent baseline accuracy without overlap-driven rejection.

plan to test and evaluate other customed methods based on variants.

The reason why overlap-based selection works well needs to be analyzed, and in particular, our basis hypothesis, i.e., that we reject noisy or ambiguous samples. This can be done by simple inspection of the rejected samples.
Our approach is general enough to be applied to any fine-grained classification tasks, especially in biomedical imaging. We plan to extend it to other classification tasks in future work.

## Conclusion

We have sought to improve the fine-grained classification by deep learning, using in particular attention-based CNNs. To this end, we have proposed a novel confidence measure for classification based on the overlap of various generated maps obtained by segmentation, attention module for feed-forward CNNs, and saliency maps. This approach is useful to assist fine-grained identifications and to help annotate large-scale wildlife data. We also showed that our models are highly accurate when choosing the right overlap threshold and more robust to detail-rich environments. Our approach is general enough to be applied to other fine-grained recognition tasks. Our tools are now in a deployment phase in recognized institutions and provide solutions to a real need in the research community, showing potential impact and connection to citizen scientists.

## References

Achille, A.; Paolini, G.; and Soatto, S. 2019. Where is the information in a deep neural network? *arXiv preprint arXiv:1905.12213* .

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

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

De Fauw, J.; Ledsam, J. R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D.; et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* 24(9): 1342–1350.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Feinman, R.; Curtin, R. R.; Shintre, S.; and Gardner, A. B. 2017. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410* .

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.

Jin, J.; Dundar, A.; and Culurciello, E. 2015. Robust convolutional neural networks under adversarial noise. *arXiv preprint arXiv:1511.06306* .

Kaplan, L.; Cerutti, F.; Sensoy, M.; Preece, A.; and Sullivan, P. 2018. Uncertainty-aware AI ML: why and how. *arXiv preprint arXiv:1809.07882* .

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.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730* .

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.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

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

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; Gorur, D.; and Lakshminarayanan, B. 2018. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136* .

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.

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer School on Machine Learning*, 63–71. Springer.

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.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, 3179–3189.

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

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* 15(56): 1929–1958. URL http://jmlr.org/papers/v15/srivastava14a.html.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.