Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

German
Research Center
for Artificial
Intelligence

Bonn-Aachen
International Center for
Information Technology

Benchmarking Out-of-Distribution Detection in 2D Object Detection

Thesis Defense

March 23, 2022

Jaswanth Bandlamudi

Supervisors

Prof. Dr. Paul G Plöger
Prof. Dr. Nico Hochgeschwender
Prof. Dr. Matias Valdenegro Toro
M.Sc. Octavio Arriaga

Introduction

- Deep Neural Networks, current State-Of-The-Art (SOTA) performers in
  - Classification
  - Object Detection
  - Segmentation
- Trained with *closed world assumption*, test data $\sim$ train data
- Deployed in open world $\implies$ Out-of-Distribution(OOD) data
- Applications
  - Product recommendations, recoverable
  - Time series prediction, partially reversible
  - Autonomous driving / Medical diagnosis, irreversiable and catastrophic

Out-of-Distribution (OOD) detection (1/3)

- What is OOD data ?
  - Data that is outside the semantic space formed by the images used for training
  - Input with objects which are not used in training but have features closer to the object of interest.

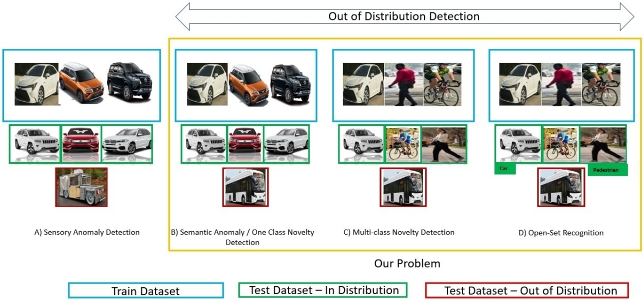


Figure: Class differentiation in generalized OOD detection framework

Out-of-Distribution (OOD) detection(2/3)

Different types of OOD data

- Data from a different domain
- Data with poor quality of features
- Data with inputs that are neither used nor prominent in the training data

Out-of-Distribution (OOD) detection(3/3)

Current Object Detection model performance on OOD data

(a)

(b)

Figure: Examples of failures in object detection

OOD detector - Expectations

- Produce a ***Novelty Score (NS)***.
- NS can be a distance metric, a class-dependent probabilistic value, an entropy value, or a descriptive statistic value
- OOD detection as a binary classification problem and AUROC score for comparision.

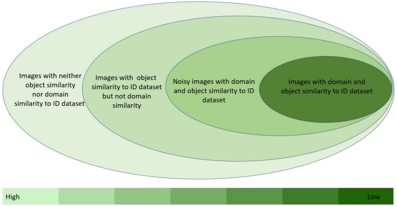


Figure: Expected behavior of OOD detector.

$$X = \begin{cases} \text{ID}, & \text{if } NS \geq \delta \\ \text{OOD}, & \text{otherwise} \end{cases} \quad (1)$$

Previous works

Table: Previous works on OOD detection

Method	Works Proposed
Metric based methods	Devries and Taylor [2018], Oberdiek et al. [2018], Hendrycks et al. [2018] , Lee et al. [2018]
Inconsistency based methods	Liang et al. [2017]
Generative methods	Hendrycks and Gimpel [2017a], Ren et al. [2019], Van Den Oord et al. [2016]
Uncertainty based methods	Malinin and Gales [2018], Lakshminarayanan et al. [2017], Van Amersfoort et al. [2020]

- Works only for classification problem
- Not directly adaptable to object detection problem

Uncertainty Quantification SoTA

- Bayesian Neural Networks
  1. Bayes by Backprop [Blundell et al., 2015]
  2. Variatonal Dropout and Reparameterization [Kingma et al., 2015]
  3. Flipout sampling based Bayesian Neural Networks [Wen et al., 2018]
- Monte carlo Dropout [Gal, 2016]
- Ensemble methods
  1. Deep Ensembles [Lakshminarayanan et al., 2017]
  2. Deep Sub-Ensembles [Valdenegro-Toro, 2019]

Proposed OD^2 benchmark

Table: Table showing various type of images to address the OOD cases

Purpose	Dataset Source	Classes	Novelty Score Behavior	Task
In-Distribution	BDD100K [Yu et al., 2020]	Pedestrian, Rider, Car, Truck, Bus, Motorcycle, Bicycle, Traffic sign	Low Novelty score	Object detector performance
Low light and bad image quality	BDD100K (non-clear weather) and Climate-GAN [Schmidt et al., 2021] generated Smog images	Pedestrian, Rider, Car, Truck, Bus, Motorcycle, Bicycle, Traffic sign	Medium Novelty Score	Detector Robustness
Classes with semantic-variance	IDD [Varma et al., 2019]	Trucks, Motorcycles, Traffic Sign	High Novelty Score	OOD detection
Novel Classes	IDD	Auto-Rickshaws	High Novelty Score	Multi class novelty detection
Out-of-Domain images	Climate-GAN generated Flood and Fire images	Pedestrian, Rider, Car, Truck, Bus, Motorcycle, Bicycle, Traffic sign	High Novelty Score	Out-Of-Domain detection

Sample Images

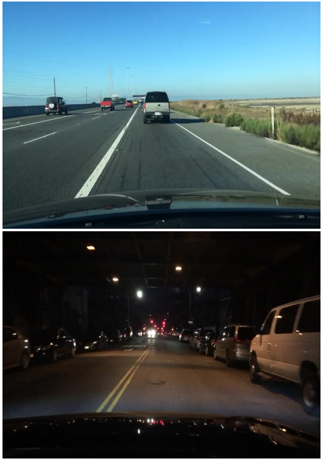


Figure: BDD100K (In-Distribution)

Figure: Indian Driving Dataset (OOD)

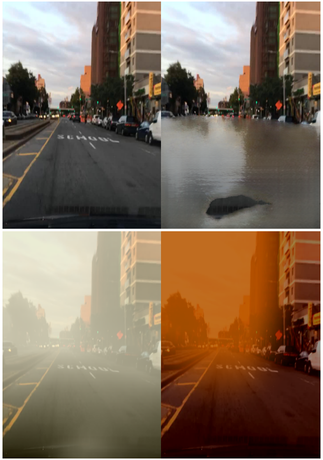


Figure: Climate-GAN images (OOD)

Experiment methodology

- Object Detector: Single Shot multi-box Detector [Liu et al., 2016]
- Out of Distribution detection methods:
  1. Max-Softmax [Hendrycks and Gimpel, 2017b]
  2. ODIN [Liang et al., 2017]
  3. Mahalanobis distance based OOD detection [Lee et al., 2018]
  4. Uncertainty-based OOD detection
     - Bayesian Neural Networks with Flipout sampling[Wen et al., 2018]
     - SubEnsembles [Valdenegro-Toro, 2019]

SSD Performance

Table: AP values of SSD300 model with different prior boxes

Prior boxes	Agents								
	Pedestrian	**Rider**	**Car**	**Truck**	**Bus**	**Motorcycle**	**Bicycle**	**Traffic Sign**	**Mean**
vanilla	0.006	0.004	0.095	0.083	0.15	0.045	0.092	0.001	0.059
tuned	0.165	0.135	0.479	0.389	0.389	0.163	0.213	0.186	0.267

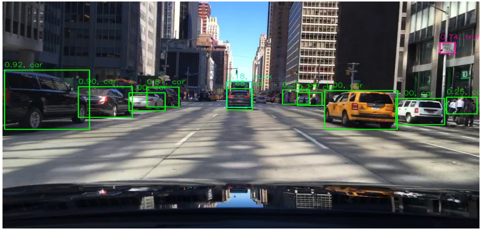


Figure: sample detection from BDD100K dataset

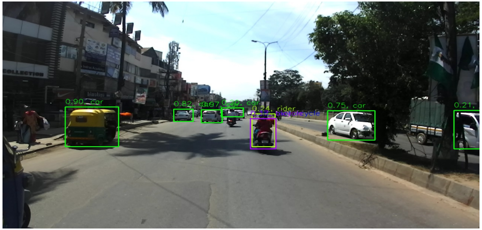


Figure: sample detection from IDD dataset

OOD Detection - MaxSoftmax

- AUROC score of 48 %
- poorer than un-biased random classifier
- complex scenarios, class overlap

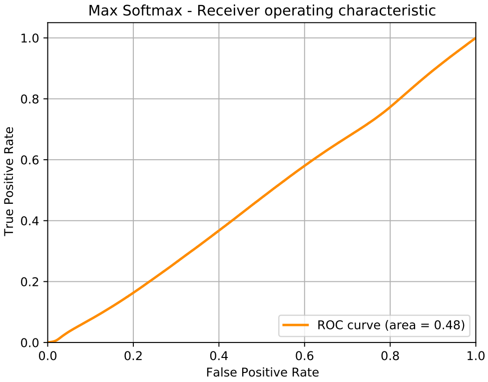


Figure: AUROC curve for OOD detection using softmax scores.

OOD Detection - ODIN (1/2)

- gradient of loss w.r.t input image is calculated and scaled by a perturbation magnitude.
- input image is modified by subtracting the perturbation.

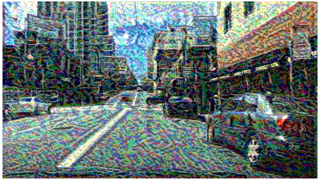


Figure: magnitude=0.2, Temperature=10

Figure: magnitude=0.005, Temperature=100

Figure: magnitude=0.005, Temperature=10

- hyperparameters are tuned using a fraction of test images sampled from IDD dataset
- Perturbation Magnitude 0.2 and Temperature of 1000

OOD Detection - ODIN (2/2)

- AUROC score of 54 %
- improvement over max-softmax method
- effect of perturbation is not observed in smaller objects

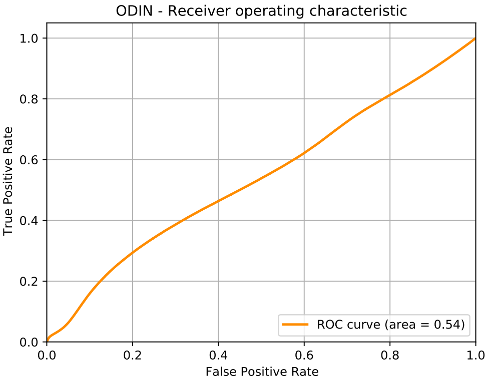


Figure: AUROC curve for OOD detection using softmax scores after applying ODIN

OOD Detection - Mahalanobis distance (1/1)

- class-wise mean vectors of each class and a tied covariance matrix using the features from the penultimate layer
- images with other classes masked out by the mean of the image

  - flattened penultimate layer of SSD is of shape (78588×1)
  - the covariance matrix is of shape (78588×78588)
  - 25GB of RAM / GPU per image.

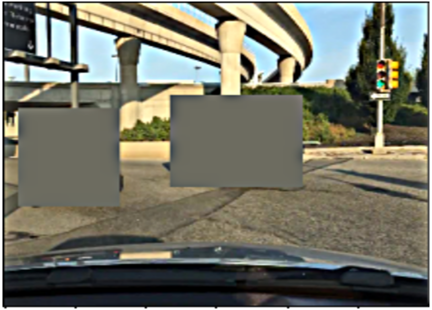


Figure: Class masked images to extract class specific mean and variance

UQ models - Performance (1/6)

Layers highlights in green are chosen as

- Flipout layers to model Bayesian-SSD
- Task network to model SubEnsemble SSD

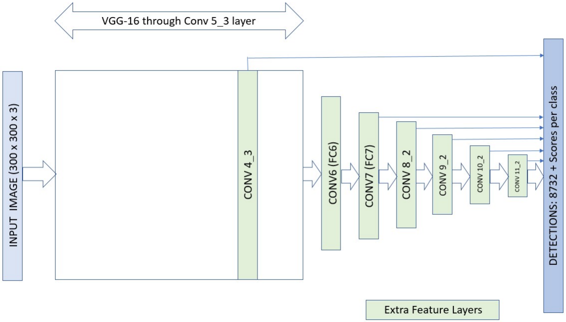


Figure: modified SSD for uncertainty quantification

UQ models - Performance (2/6)

Table: AP values of various SSD300 models

Models	Agents								
	Pedestrian	Rider	Car	Truck	Bus	Motorcycle	Bicycle	Traffic Sign	Mean
SSD300	0.165	0.135	0.479	0.389	0.389	0.163	0.213	0.186	0.267
Bayesian SSD300	0.172	0.149	0.476	0.4	0.401	0.196	0.232	0.184	0.276
SubEnsemble SSD300	0.167	0.144	0.47	0.393	0.396	0.181	0.211	0.171	0.267

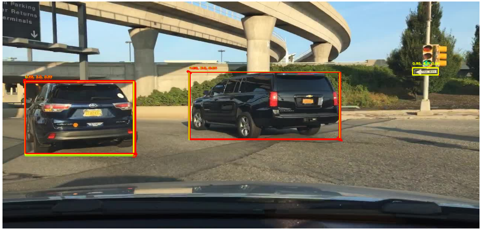


Figure: UQ on image from BDD100K dataset

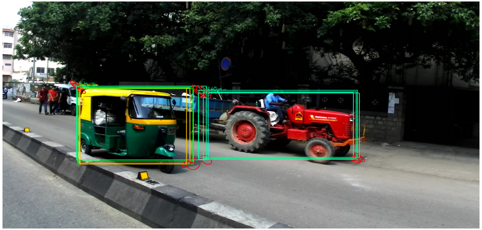
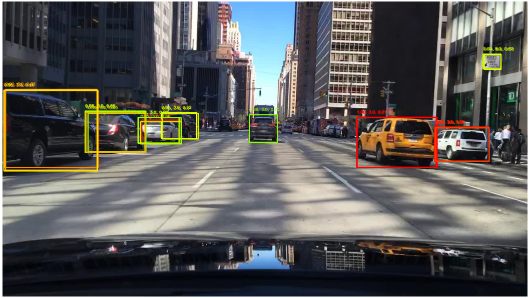
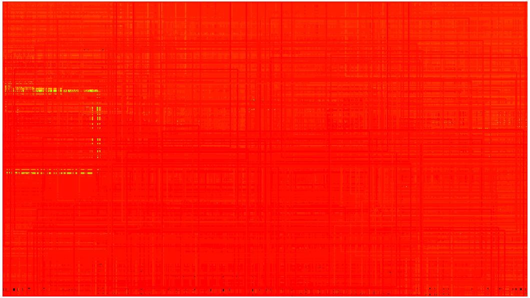


Figure: UQ on image from IDD dataset

UQ models - Performance (3/6)



(a) entropy in box color and variances represented as ellipses



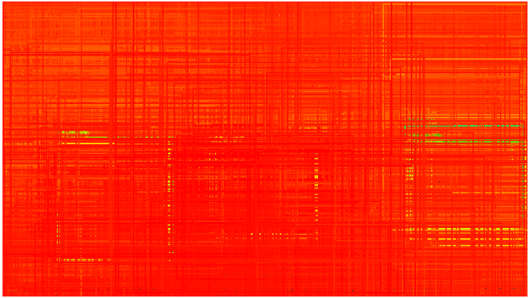
(b) Visualizing all 8732 boxes regressed by SSD300

(c) Entropy Colormap

Figure: Inference of Bayesian-SSD300 model on a sample image from BDD

UQ models - Performance (4/6)

(a) entropy in box color and variances represented as ellipses



(b) Visualizing all 8732 boxes regressed by SSD300

Figure: Inference of Bayesian-SSD300 model on a sample image from IDD

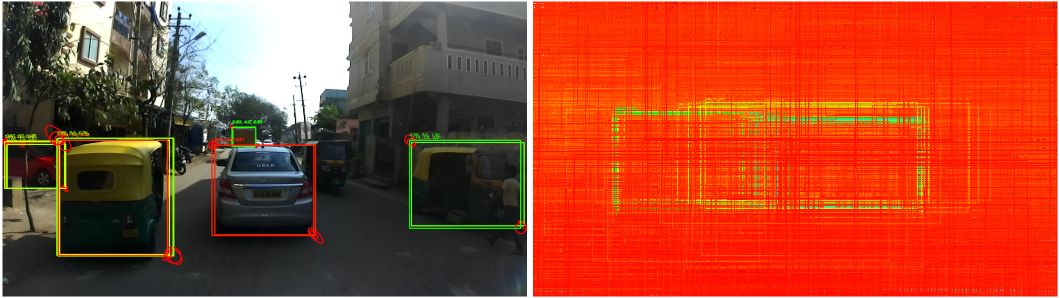
UQ models - Performance (5/6)

(a) entropy in box color and variances represented as ellipses

(b) Visualizing all 8732 boxes regressed by SSD300

Figure: Inference of SubEnsemble-SSD300 model on a sample image from BDD

UQ models - Performance (6/6)

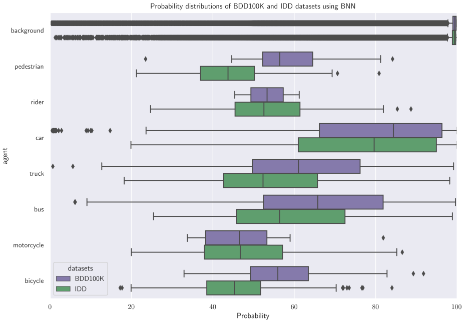


(a) entropy in box color and variances represented as ellipses

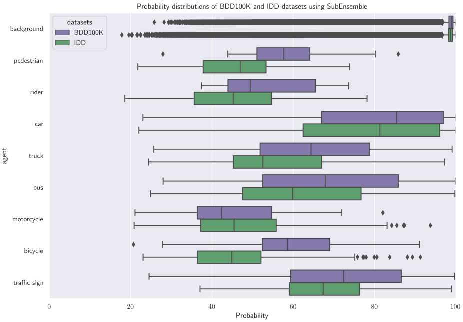
(b) Visualizing all 8732 boxes regressed by SSD300

Figure: Inference of SubEnsemble-SSD300 model on a sample image from IDD

UQ models - OOD Detection (1/7)



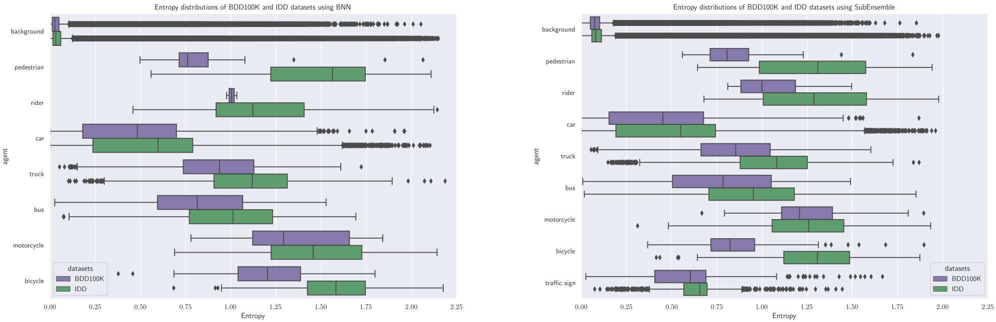
(a) Probabilities between BDD and IDD using Bayesian SSD300



(b) Probabilities between BDD and IDD using SubEnsemble SSD300

Figure: Probabilities from Bayesian and SubEnsemble SSD300 model

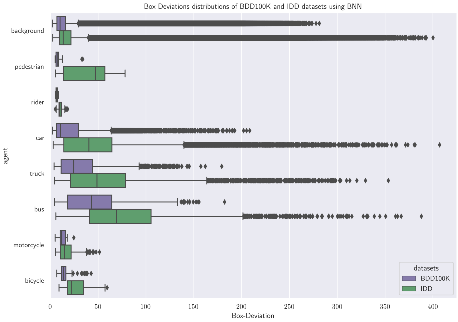
UQ models - OOD Detection (2/7)



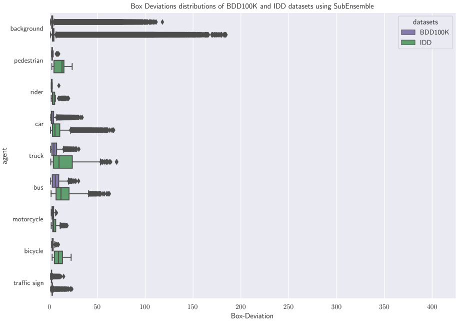
(a) Entropy between BDD and IDD using Bayesian SSD300 (b) Entropy between BDD and IDD using SubEnsemble SSD300

Figure: Entropy from Bayesian and SubEnsemble SSD300 model

UQ models - OOD Detection (3/7)



(a) Box-Deviation between BDD and IDD using Bayesian SSD300



(b) Box-Deviation between BDD and IDD using SubEnsemble SSD300

Figure: Box-Deviation from Bayesian and SubEnsemble SSD300 model

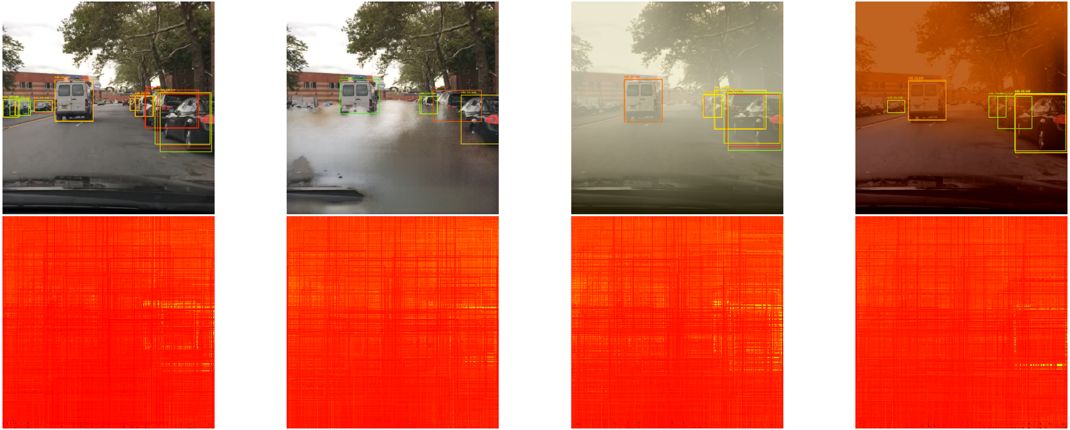
UQ models - OOD Detection (4/7)

Table: AUROC scores for classifying BDD100K and IDD datasets using UQ metrics

Models	**Metrics**	**Agents**									
		Background	Pedestrian	Rider	Car	Truck	Bus	Motorcycle	Bicycle	Traffic Sign	Mean
Bayesian-SSD300	Probability	0.45	0.14	0.75	0.45	0.37	0.38	0.49	0.23	-	0.45
	Entropy	0.56	0.88	0.59	0.59	0.67	0.64	0.6	0.84	-	0.56
	Box_deviation	0.64	0.93	0.82	0.76	0.69	0.7	0.62	0.85	-	0.64
SubEnsemble SSD300	Probability	0.44	0.21	0.34	0.46	0.34	0.4	0.53	0.19	0.41	0.44
	Entropy	0.56	0.86	0.72	0.58	0.71	0.62	0.51	0.88	0.62	0.56
	Box_deviation	0.76	0.95	0.84	0.79	0.74	0.75	0.71	0.94	0.88	0.75

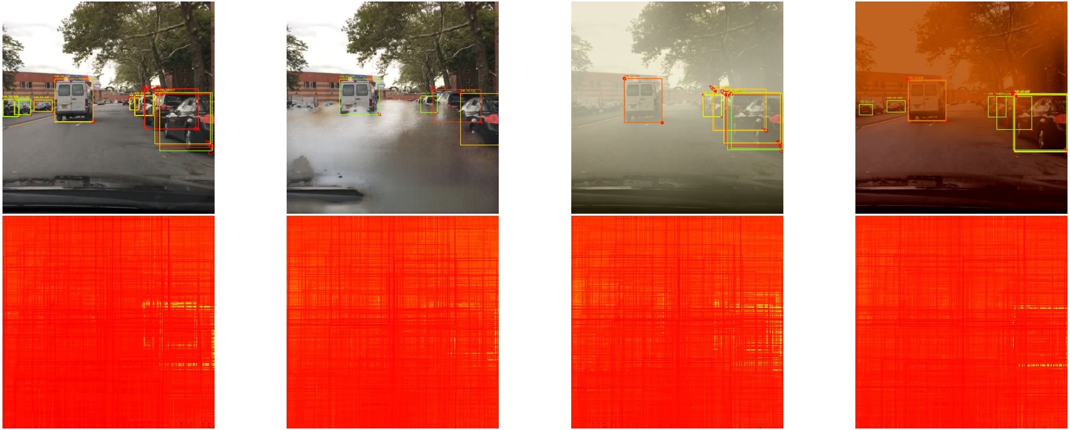
- Box deviation performed well in detecting OOD.
- performance is similar in Car, Truck, Bus, and Rider

UQ models - OOD Detection (5/7)



Performance of BNN on BDD100K weather dataset

UQ models - OOD Detection (6/7)



Performance of SubEnsembles on BDD100K weather dataset

UQ models - OOD Detection (7/7)

Table: Uncertainty quantification metrics calculated using Bayesian and SubEnsemble versions of SSD300 model

		Bayesian SSD300	SubEnsemble SSD300
Flood	Probability	0.43	0.5
	Entropy	**0.57**	0.5
	Box Deviation	0.46	0.5
Smog	Probability	0.43	0.5
	Entropy	**0.57**	0.5
	Box Deviation	0.44	0.5
Wild Fire	Probability	**0.51**	**0.53**
	Entropy	0.49	0.46
	Box Deviation	0.49	0.46

- performance has deteriorated especially in the case of flood images.
- uncertainty quantification methods performed almost similar to an unbiased random classifier.
- Results complies with the results reported by [Ovadia et al., 2019]

Contributions (1/1)

- Proposed seminal benchmark dataset for Out-of-Distribution detection in Object Detection
- Max Softmax, ODIN, Mahalanobis distance, and uncertainty based OOD detector are modelled.
- SSD300 is trained on BDD100K dataset and tuned the prior boxes
- Entropy and box-deviation are used to quantify uncertainty
- Extensive experimentation on all the three available OOD detectors for object detection purposes.
- Class-specific behavior of the uncertainty quantification metrics is studied.

Observations (1/1)

- Object Detector performance is dependent on the prior knowledge of the dataset.
- Deep learning based object detectors struggle when deployed in open environments.
- SubEnsemble based uncertainty quantification with box deviation as a metric out-performed all other methods.
- Entropy struggled in detecting samples that are ambiguous due to their semantic appearance.
- The OOD detection methods proposed in this work did not work as expected on the BDD100K-Weather data in the OD^2 benchmarking dataset.

Future-work (1/1)

- An object detector without the usage of background class would be useful for improved OOD performance.
- Exploring the effects of uncertainty calibration methods Guo et al. [2017] on OOD detection is still a open-ended question.
- combining Entropy and Box-Deviation to obtain a single novelty score might result in a better representation of OOD detection ability.
- Disentangling of Predictive uncertainty.
- we believe OD^2 can be further modified and extended to include more class-agnostic tasks.

References (1/5)

Terrance Devries and Graham W Taylor. Learning Confidence for Out-of-Distribution Detection in Neural Networks. 2018.

Philipp Oberdiek, Matthias Rottmann, and Hanno Gottschalk. Classification uncertainty of deep neural networks based on gradient information. In **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, volume 11081 LNAI, pages 113–125, 2018. ISBN 9783319999777.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In **arXiv**, 2018.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Technical report, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. **arXiv preprint arXiv:1706.02690**, 2017.

References (2/5)

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In **5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings**, 2017a. ISBN 1610.02136v3.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In **arXiv**, 2019.

Äron Van Den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional image generation with PixelCNN decoders. In **Advances in Neural Information Processing Systems**, pages 4797–4805, 2016.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In **Advances in Neural Information Processing Systems**, volume 2018-December, pages 7047–7058, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In **Advances in Neural Information Processing Systems**, volume 2017-December, pages 6403–6414, 2017.

References (3/5)

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. Technical report, 2020.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In **Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37**, ICML'15, page 1613–1622. JMLR.org, 2015.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, **Advances in Neural Information Processing Systems 28**, pages 2575–2583. Curran Associates, Inc., 2015.

Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. Technical report, 2018.

Yarin Gal. **Uncertainty in Deep Learning**. PhD thesis, University of Cambridge, 2016.

Matias Valdenegro-Toro. Deep Sub-Ensembles for Fast Uncertainty Estimation in Image Classification. Technical report, 2019. URL `http://arxiv.org/abs/1910.08168`.

References (4/5)

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In **IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, June 2020.

Victor Schmidt, Alexandra Sasha Luccioni, Mélisande Teng, Tianyu Zhang, Alexia Reynaud, Sunand Raghupathi, Gautier Cosne, Adrien Juraver, Vahe Vardanyan, Alex Hernandez-Garcia, and Yoshua Bengio. Climategan: Raising climate change awareness by generating images of floods, 2021.

G. Varma, A. Subramanian, A. Namboodiri, Manmohan Chandraker, and C. V. Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. **2019 IEEE Winter Conference on Applications of Computer Vision (WACV)**, pages 1743–1751, 2019.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, **Computer Vision – ECCV 2016**, pages 21–37, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46448-0.

References (5/5)

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. **Proceedings of International Conference on Learning Representations**, 2017b.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. **Advances in Neural Information Processing Systems**, 32, jun 2019. ISSN 10495258. URL `https://arxiv.org/abs/1906.02530`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In **Proceedings of the 34th International Conference on Machine Learning - Volume 70**, ICML'17, page 1321–1330. JMLR.org, 2017.

Single Shot multi-box Detector (SSD) model (1/2)

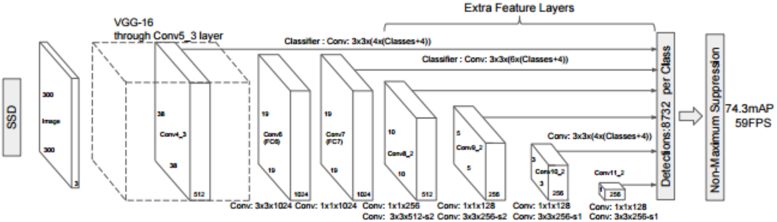


Figure: SSD framework proposed by [Liu et al., 2016, p. 24].

- Single network for detection and classification
- No Fully-Connected layers
- Low input resolution

Single Shot multi-box Detector (SSD) model (2/2)

- Default boxes
- Matching strategy is used,
  - $IoU^{groundtruth}_{defaultbox}$ > 0.5
  - overlapped objects and simple learning
- Processing of features from multiple layers
  - Deep feature maps
  - Shallow feature maps
- Loss

$$L(x,c,l,g) = \frac{1}{N}\left(L_{\text{conf}}(x,c) + \alpha L_{\text{loc}}(x,l,g)\right)$$

  - L_{conf} is Softmax Loss
  - L_{loc} is Smooth L_1 Loss
- Filter boxes with low confidence and NMS with 0.45 IOU
- Top 200 detections are considered

OOD methods (1/6)

- Max-Softmax

  Maximum value of softmax scores are used as novelty score

$$s\left(\mathbf{x}^*\right)=\max_c P\left(y_c \mid \mathbf{x}^* ; \mathcal{D}\right) \tag{2}$$

- ODIN

$$\tilde{\boldsymbol{x}}=\boldsymbol{x}-\operatorname{esign}\left(-\nabla_{\boldsymbol{x}} \log S_{\hat{y}}(\boldsymbol{x} ; T)\right) \tag{3}$$

$$S_i(\boldsymbol{x} ; T)=\frac{\exp \left(f_i(\boldsymbol{x}) / T\right)}{\sum_{j=1}^N \exp \left(f_j(\boldsymbol{x}) / T\right)} \tag{4}$$

  - ϵ is the perturbation magnitude
  - T is the Temperature

OOD methods (2/6)

- Mahalanobis distance based OOD detection
  assuming intermediate layer features follow class-conditional Gaussian distributions with tied covariances

$$M(x) = \max_{c} -\left(f(x) - \hat{\mu}_c\right)^T \hat{\Sigma}^{-1} \left(f(x) - \hat{\mu}_c\right) \tag{5}$$

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_c=c} f\left(x_i\right)$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{c} \sum_{i:y_c=c} \left(f\left(x_i\right) - \hat{\mu}_c\right)\left(f\left(x_i\right) - \hat{\mu}_c\right)^T$$

OOD methods (3/6)

- Uncertainty based OOD detection
  - quantifies trustworthiness in the model output
    - » epistemic uncertainty, higher in areas of low data density.
    - » aleatoric uncertainty, labelling and measurement noise
  - OOD data is implicitly modeled by epistemic uncertainty
  - To quantify epistemic uncertainty, we used
    - » Bayesian Neural Netowrk.
    - » Deep SubEnsembles

OOD methods (4/6)

– Bayesian Neural Network

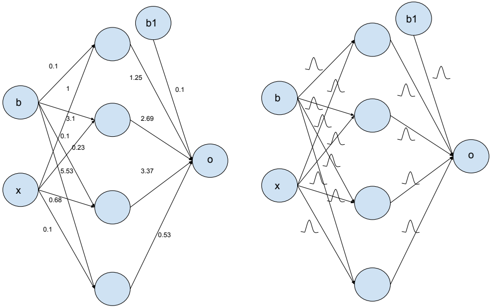


Figure: Bayesian Neural Network

- Bayesian Flipout layers [Wen et al., 2018]
- Reparameterization trick for training [Kingma et al., 2015]
- Prior $P(w) \sim N(0,1)$
- multiple forward passes for uncertainty quantification

OOD methods (5/6)

– SubEnsemble Network

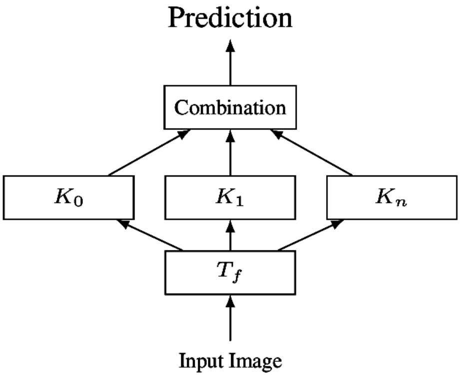


Figure: SubEnsemble Network

- Model is divided into Trunk and Task layers
- Trunk layers has best performing weights restored and cannot be trained
- Task layers are randomly initialized and re-trained
- Random initialization of layers creates ensemble model.

OOD methods (6/6)

– Novelty Score

Entropy

$$\text{Entropy} = -\sum_{i=1}^{C} P\left(c_i \mid \mathbf{x}^*; \mathcal{D}\right) \ln P\left(c_i \mid \mathbf{x}^*; \mathcal{D}\right) \tag{6}$$

Box deviation is the square root of the trace of the covariance matrix $C(x^*)$.

$$C\left(\mathbf{x}^*\right) = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{v}}_{\mathbf{x}^*}^{i} \hat{\mathbf{v}}_{\mathbf{x}^*}^{i\,T} - \mathbf{I}_{\mathbf{x}^*} \mathbf{I}_{\mathbf{x}^*}^{T} \tag{7}$$

SSD Object Detection Results (1/2)

Table: AP values for various classes using vanilla-SSD Prior boxes

Class	**score**
Pedestrian	0.006
Rider	0.004
Car	0.095
Truck	0.083
Bus	0.15
Motorcycle	0.045
Bicycle	0.092
Traffic Sign	0.001
Mean	0.059

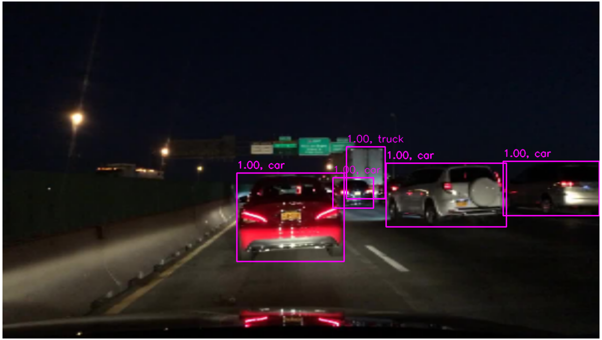


Figure: Positively matched vanilla prior boxes with ground truth boxes.

- Poor performance, can be improved by tuning

SSD Object Detection Results (2/2)

Table: AP values for various classes using tuned Prior boxes

Class	**score**
Pedestrian	0.165
Rider	0.135
Car	0.479
Truck	0.389
Bus	0.389
Motorcycle	0.163
Bicycle	0.213
Traffic Sign	0.186
Mean	0.265

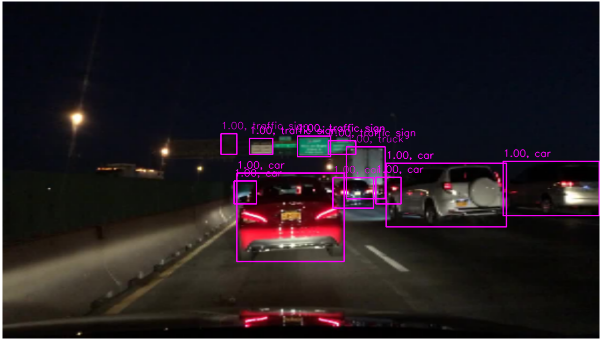


Figure: Positively matched tuned prior boxes with ground truth boxes..

- improved performance