Reliability benchmarks for image segmentation

E. Kelly Buchanan, Michael W. Dusenberry, Jie Ren, Kevin Patrick Murphy
Balaji Lakshminarayanan, Dustin Tran
Google Research, Brain Team
`ekb2154@columbia.edu` `trandustin@google.com`

Abstract

Recent work has shown the importance of *reliability*, where models are assessed under stress conditions pervasive in real-world deployment. We examine reliability tasks in the setting of semantic segmentation, a dense output problem that has typically only been evaluated using in-distribution predictive performance—for example, the mean intersection over union score on the Cityscapes validation set. To reduce the gap toward reliable deployment in the real world, we compile a benchmark involving existing (and newly constructed) distribution shifts and metrics. We evaluate natural baselines to determine how well segmentation models can simultaneously make robust predictions across multiple types of distribution shift, detect out-of-distribution inputs, and make calibrated predictions. We find that Gaussian process and BatchEnsemble last-layers work well out-of-the-box, improving existing state-of-the-art across tasks. There also remain open challenges in measuring out-of-distribution detection for segmentation.[1]

1 Introduction

We consider a model to be "reliable" if it can perform well over a large collection of decision making scenarios [1, 13, 16]. To test the reliability of deep learning models, our field has and continues to establish several benchmarks along multiple directions: (i) new datasets that can capture different types of distribution shifts, changes in the data which were not included as examples in the training set; (ii) new methods, to reduce the noise introduced by the different distribution shift; and (iii) new evaluation scores or metrics, to evaluate the confidence in the prediction and robustness of a model. However, pushing along multiple directions has led to a fragmentation of the literature, where it is unclear how a method developed to improve model performance under a specific type of distribution shift, fares when another distribution shift or task is present.

Recently Tran et al. [16] developed a stress-test suite to evaluate the uncertainty, robustness, and adaptation abilities of models for a wide variety of datasets and tasks. Tran et al. [16] showed that last layer replacements for dense layers were able to provide consistent performance improvements over deterministic linear layers. While Tran et al. [16] focused on classification tasks for vision and language models, we find that such a fragmentation also exists for semantic segmentation tasks.

Contributions We provide a benchmark for reliable semantic segmentation. The benchmark includes two in-distribution datasets (Cityscapes and ADE20k), over three types of distribution shift (covariate shift, open set shift, and natural shift), and three tasks (predictive performance, out-of-distribution detection, calibration). Our results show that a minimal substitution of the last layer in a deterministic model for a GP or BatchEnsemble layer, produces results more reliable across distributions shifts and the tasks. Our framework is open source, implemented in JAX, and can be easily extended to include additional models and tasks.

[1] The code is available at `https://github.com/google/uncertainty-baselines`.

Workshop on Distribution Shifts, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

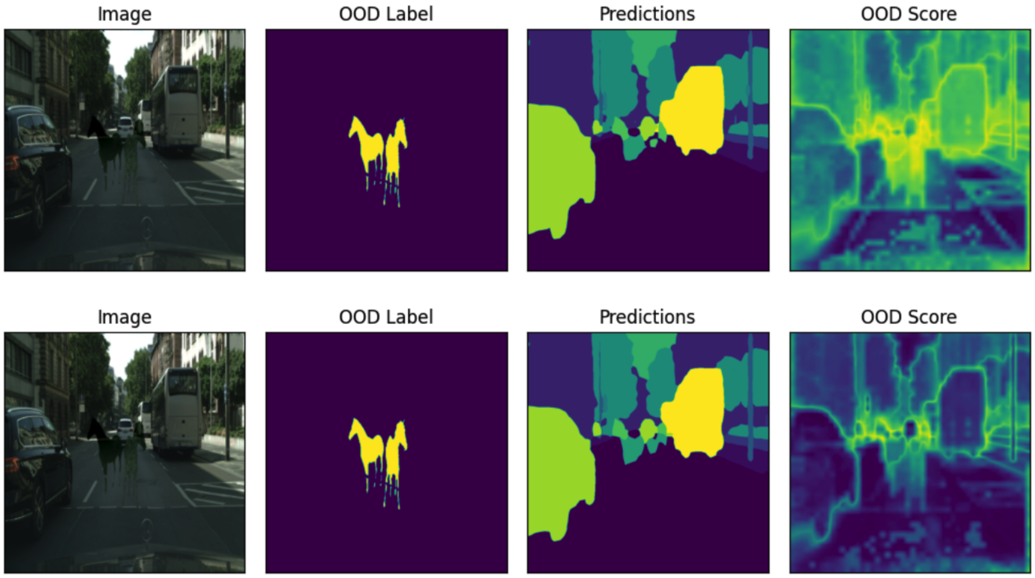

Figure 1: Predictions from a segmenter-deterministic (**top**) and a segmenter-GP model (**bottom**). From left to right: input image with OOD object, OOD mask, model predictions, OOD feature map. The fourth panel shows us that all the logits are not all proportionally scaled equally as the OOD features are more clearly defined to the GP model compared to the deterministic model.

2 Methods

Following the setup of Tran et al. [16, Fig. 4], we start with a Vision Transformer (ViT) model pretrained on Imagenet21k and finetuned on Imagenet2012 [14]. Given an image, $\boldsymbol{x} \in \mathbb{R}^{H \times W \times C}$, a ViT encoder maps the input $\boldsymbol{x}$ to the patch encodings, the positional enconding and class tokens. When a ViT is trained for image classification, the class tokens are mapped to different classes via a dense layer. Strudel et al. [15] proposed a segmentation model based on a ViT backbone, where the patch and positional encodings are mapped to the pixel classes via linear decoding consisting of a dense layer and upsampling. A Segmenter L/16 model is shown to achieve state of the art performance for semantic segmentation tasks such as Cityscapes and ADE20k.

Reliability methods We evaluate reliability methods from Plex [16], which has been shown to work well for image classification. Namely, we replace the linear decoder in Segmenter for one of the following decoders (see Fig. 1 for an illustrative example of how these improve model reliability):

- BatchEnsemble [17] decoder: a BatchEnsemble layer substitutes a single decoder or head for “multiple heads", where these heads have a shared structure and can be factorized into vectors. In turn, these vectors enable learning efficient ensembles, while still providing multiple diverse predictions that are averaged for an ensemble prediction.
- Gaussian Process (GP) [11] decoder: a Gaussian Process prior is placed on top of the hidden representations of the ViT backbone. In turn, this prior induces a posterior predictive distribution over the data likelihood. As computing the GP posterior distribution is intractable and expensive, [11] approximates the posterior using a Laplace approximation to the random feature expansion of the GP, resulting in an approximate posterior that can be learned in closed form.
- Heteroscedastic [6] decoder: a Gaussian distribution is assigned to the model logits, where the noise term is modeled as independent but not identically distributed for each logit. Adding this noise term supports data with variable noise. The intractable posterior is approximated using Monte Carlo sampling. We note that [12] is a special case of this work.

We train the Segmenter models following [15]. For Cityscapes, we train a Segmenter L/16 model for 100 epochs with a learning rate of 1e-4 with adam following a polynomial learning rate schedule. For

ADE20k, we train a Segmenter L/16 model for 100 epochs with a learning rate of 3e-5 with Adam following a polynomial learning rate schedule.

Datasets We train Segmenter models with deterministic or Plex layers on two standard scene parsing datasets as in-distribution. The Cityscapes dataset [7] contains street scenes from 50 different cities, and includes 2900 train images for 19 classes, and 500 images for the validation set. The image dimensions are 1024x2048. The ADE20K dataset [18] contains 20,210 images for 150 classes for the training set, and 500 images for the validation set. The image dimensions are 640x640. These datasets allow us to evaluate the changes in performance, across different conditons: in a low data regime, as Cityscapes has much fewer classes than ADE20k, for different granularity level, i.e. image resolution, and for different number of classes.

Distribution shifts We examine distribution shifts pervasive in the real world: changes in the input environment (covariate shift) and changes in the label classes (open set recognition). Table 1 provides a summary. To simulate environment changes, we apply the synthetic corruptions in [8] to the Cityscapes and ADE20K datasets such as Gaussian noise and under varying intensities. To simulate changes in the label classes, we employ the Fishyscapes dataset [3], constructed by adding Pascal VOC objects to different scenes in Cityscapes. For ADE20k, we construct an ADE20k-InD dataset by dropping a subset of the classes (three classes including chair, sofa and couch) in the original ADE20k dataset from the training set. We drop this classes by masking out the pixels corresponding to these classes. The ADE20k-Open dataset corresponds to the validation set in ADE20k where there are only two classes, class 0 being all the pixels which correspond to a class included in ADE20k-InD and class 1 being all the pixels which correspond to three objects of chair, sofa, and couch.

Evaluation Metrics For in-distribution predictive performance, we evaluate Segmenter on the validation set of Cityscapes and ADE20k using the mean intersection over union (MeanIOU) score, which captures the overlap between the true segmentation map and the predicted segmentation map [4]. We also evaluate the MeanIOU under covariate shift. While covariate shift changes the appearance of the input, the labels do not change, and we can directly compare the InD MeanIOU and the OOD MeanIOU. For calibration performance, we employ the calibration AUC-ROC which measures the ranking performance of the uncertainty score; and unlike the Expected Calibration Error (ECE), it is not sensitive to class imbalance.

For open set recognition, we use $1 - \mathrm{msp}$, where the msp is the maximum softmax probability, as the OOD score [9, 4, 3]. In semantic segmentation, the goal of open set recognition is to identify pixels in the image which do not correspond to a class in the training set. The model produces a score per pixel between 0 to 1. Given the score, an OOD binary mask is created by setting a threshold λ at which a pixel is considered InD vs OOD. Because the binary masks depends on the threshold λ, we evaluate across thresholds using the area under the receiver operator characteristic curve (AUC-ROC), and the precision recall curve (AUC-PR).

Table 1: Segmentation benchmark for robustness evaluation.

InD dataset	Covariate shift	Open set recognition
Cityscapes	Cityscapes-C	Fishyscapes
ADE20k	ADE20k-C	ADE20k-Open

3 Results

We report the performance of Segmenter with the deterministic and Plex layers for Cityscapes in Table 2 and for ADE20K in Table 3. The Plex models perform comparably well or outperform the deterministic models both in-distribution and out of distribution. In particular, the GP layer has benefits over other models when the environment changes (covariate shift) and when there is an anomalous object (open set) in the low data regime; while the BatchEnsemble and Heteroskedastic layers can give us gains even as the number of classes is large (>150).

The GP layer has slightly better performance than the deterministic layer on both datasets in-distribution. Moreover, the GP layer has significant improvements on both covariate shift and open

Table 2: Comparison of reliability models on Cityscapes dataset.

Model	Cityscapes Val MeanIOU (↑)	Cityscapes Calibration AUC (↑)	Cityscapes-C Val MeanIOU (↑)	Fishyscapes AUC-ROC (↑)	Fishyscapes AUC-PR (↑)
Deterministic	0.760	0.905	0.620	0.777	0.067
Batch Ensemble	0.764	0.911	0.620	0.808	0.076
GP	**0.765**	**0.917**	**0.633**	**0.960**	**0.309**
Heteroskedastic	0.756	0.909	0.625	0.891	0.144

Table 3: Comparison of reliability models on ADE20k dataset.

Model	ADE20k-InD Val MeanIOU (↑)	ADE20k-InD Calibration AUC (↑)	ADE20k-InD-C Val MeanIOU (↑)	ADE20k-OOD AUC-ROC (↑)	ADE20k-OOD AUC-PR (↑)
Deterministic	0.482	0.793	0.391	0.726	0.046
Batch Ensemble	0.489	**0.822**	0.395	**0.765**	**0.052**
GP	0.487	0.753	0.394	0.694	0.036
Heteroskedastic	**0.494**	0.782	**0.397**	0.719	0.045

set recognition, with gains up to 5%. Liu et al. [11] posits that the variance component introduced by the GP layer induces lower confidence predictions to samples away from the training data. We hypothesize that the gains from using a GP layer over a deterministic layer are particularly pronounced for open set recognition for this reason.

BatchEnsemble provides gains both in-distribution and out-of-distribution. In particular, it achieves the best performance on ADE20k. In future work we plan to examine if the gains provided by the batch ensemble and GP layer can be compounded as suggested by [16].

The heteroscedastic layer improves the in-distribution performance of Segmenter on both Cityscapes and ADE20k. This is congruent with Fig. 4 in [5], which showed that replacing a dense layer by a heteroscedastic layer improved both quantitatively and qualitatively the segmentation maps provided by a Deeplabv3+ network trained on MSCOCO. Because ADE20k has a much larger number of classes than Cityscapes (150 and 19 respectively), we may expect a higher likelihood of label noise, and thus more benefit using a heteroscedastic layer over a deterministic layer for ADE20k versus Cityscapes. However, the benefits appear to be comparable for Cityscapes and ADE20k in-distribution and out-of-distribution.

AUC-PR scores are generally low across the board. This is due to the large class imbalance between true positives and true negatives in the open set images, i.e. only a few pixels in the image are OOD compared to the large number of InD pixels, as noted in previous work [3, 10]. The AUC-ROC values are also particularly pronounced for Fishyscapes, as shown in Table 2. Several works have noted that the unseen objects in Fishyscapes [3] have different hue compared to the objects in Cityscapes [2, 8], and thus post-processing methods can capture these changes easily. Our results show that the GP layer can easily capture this changes without additional post-processing steps. In the full version of the manuscript we also plan to include the Street Hazards dataset [10].

4 Conclusion

We provided a benchmark to evaluate the reliability of semantic segmentation models. We evaluated the layers in Plex [16] and found that both the GP and BatchEnsemble layers have benefits both in-distribution and out-of-distribution across multiple datasets and distribution shifts over deterministic layers for semantic segmentation tasks. Through this framework, we show how Plex Layers can be easily applied out of the box, even for semantic segmentation tasks. Across the Plex layers, we find that an additional benefit of using the GP layer is that it requires less parameter search than the Batch Ensemble and Heteroscedastic layers. Our results show that replacing the last dense layer of a semantic segmentation model provides reliability gains, consistently and with minor hyperparameter search.

References

[1] Richard E Barlow and Frank Proschan. Statistical theory of reliability and life testing: probability models. Technical report, Florida State Univ Tallahassee, 1975.

[2] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

[3] Hermann Blum, Paul-Edouard Sarlin, Juan Nieto, Roland Siegwart, and Cesar Cadena. Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[4] Robin Chan, Svenja Uhlemeyer, Matthias Rottmann, and Hanno Gottschalk. Detecting and learning the unknown in semantic segmentation, 2022. URL https://arxiv.org/abs/2202.08700.

[5] Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. A simple probabilistic method for deep classification under input-dependent label noise. *arXiv preprint arXiv:2003.06778*, 2020.

[6] Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. Correlated input-dependent label noise in large-scale image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1551–1560, June 2021.

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. In *International Conference on Learning Representations*, 2019.

[9] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018.

[10] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8759–8773. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hendrycks22a.html.

[11] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

[12] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems*, 33:12756–12767, 2020.

[13] Patrick O'Connor and Andre Kleyner. *Practical reliability engineering*. John Wiley & Sons, 2012.

[14] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021.

[15] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[16] Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards reliability using pretrained large model extensions, 2022. URL https://arxiv.org/abs/2207.07411.

[17] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020.

[18] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.