Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification

Jordy Van Landeghem [1 2] **Matthew Blaschko** [3] **Bertrand Anckaert** [2] **Marie-Francine Moens** [1]

Abstract

This paper experimentally reports on predictive uncertainty for real-world text classification tasks. We define a straightforward protocol to evaluate the quality of Deep Learning uncertainty estimation. We report on a Monte Carlo Dropout-based model and data uncertainties using 1-D convolutional neural networks on multi-class news topic and sentiment classification datasets. We find that our protocol effectively enables to test for novelty detection robustness showing that Bayesian quantities underestimate uncertainty and predictive entropy demonstrates superior performance.

1. Introduction

Reliable uncertainty quantification is indispensable for any machine learning system trusted in decision-making in many application domains such as medical diagnosis, self-driving cars and automated document processing. In any typical industrial application, predictive uncertainty is expected to communicate on a model's inability to learn from the training data, deal with noisy data and train-test skew. Supervised Deep Learning (DL) algorithms have been found to provide "catastrophically overconfident predictions" (Foong et al., 2019) under data distribution shift. Specifically, novel class distributions can emerge at inference time (Pimentel et al., 2014), which desirably should be detectable in a model's uncertainty. The early work of Bishop (1994) already proposed novelty detection as "the basis of a practical system for network validation".

The context of our study is a production-level text classification system for automatically handling incoming communications in information-intensive industries (e.g. legal, banking, insurance). Imagine an insurance client's surprise when a novel request for insuring a hobby drone prompts an automated email with a proposal to sign a car insurance contract. This shows that detection of novel class distributions is critical to keep errors in automation low.

We will focus on probabilistic novelty detection in the context of text classification, as it poses an important source of errors for a learned data-to-decisions automation system. More specifically, we investigate the reliability of Monte Carlo Dropout-based uncertainty estimates for unsupervised detection of novel class data in text classification and find that the studied methods underestimate uncertainty.

Our main contributions can be summarized as follows:

- We experimentally demonstrate on real-world text classification datasets that uncertainty modelling with Bayesian DL methods does not guarantee performance increase on classification and calibration metrics.
- We propose a methodology of *leave-one-class-out* to empirically compare the robustness of uncertainty quantities under novel class distribution shift.

2. Related work

In modern Deep Learning two common uncertainty (or inversely "confidence") estimates are the prediction probability over classes, known as *softmax-score*, and the *predictive entropy* over posterior class probabilities (Shannon, 1948; Zaragoza & d'Alché Buc, 1998). However, Guo et al. (2017)'s work on confidence calibration demonstrated these to be unreliable estimates of DL uncertainty.

Bayesian DL methods build on solid mathematical foundations and hold promise for more reliable learned uncertainty estimates (Wilson, 2020). The seminal work of Gal & Ghahramani (2016) on Monte Carlo (MC) Dropout and extensions (Kendall & Gal, 2017; Li & Gal, 2017; Gal et al., 2017) proposes efficient model uncertainty estimation by exploiting dropout regularization as a method for approximate variational inference.

Arguably, most research on uncertainty estimation focuses on regression and image classification tasks as they offer visual validation on uncertainty quality. Xiao & Wang (2019); Zhang et al. (2019) present some of few works focused on obtaining uncertainty in natural language processing (NLP) classification tasks. More specifically, the uncertainty estimation methods of Xiao & Wang (2019) form a major source of inspiration for our work. However, their study focuses on the performance increase of non-probabilistic measures (mean-squared error) and only reports sentiment regression

[1] Dept. of Computer Science (LIIR), KU Leuven, Belgium [2] Contract.fit, Brussels, Belgium [3] Dept. of Electrical Engineering (ESAT), KU Leuven, Belgium. Correspondence to: Jordy Van Landeghem <first.lastname@cs.kuleuven.be>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. Copyright 2020 by the author(s).

results. Moreover, we find no quantitative evaluation of the quality of the uncertainty scores and comparison to simpler measures of uncertainty, e.g., softmax score or predictive entropy. As such, we have formulated our experiments to verify how well the methods can estimate uncertainty.

In order to measure the quality of uncertainty estimates, robustness under data distribution shift is considered an appropriate probing task. Previous work on out-of-distribution (OOD) detection shows maximum softmax probability as a solid baseline (Hendrycks & Gimpel, 2016) often improved upon by Bayesian uncertainty estimation methods (Ovadia et al., 2019). Concretely, MC dropout-based uncertainties have been successfully assessed for novelty detection of unknown phenotypes (Dürr et al., 2018). However, Vernekar et al. (2019) argues against predictive uncertainties given by MC Dropout as they can only measure uncertainty in in-distribution settings. With the goal of a fair assessment of unsupervised novelty detection with MC Dropout uncertainties, we derive a new evaluation protocol where we consider cohesive class distributions close to the in-distribution data.

3. Methodology

Firstly, this section motivates a set of representative datasets and a baseline architecture with static hyperparameters for multi-class text classification. Subsequently, we introduce how to obtain uncertainty information during training and quantify predictive uncertainty in practice. Finally, we summarize the model setups and elaborate on the protocol to test unsupervised novelty detection.

3.1. Data and Architecture

Data We use three real-world text corpora characterized by a different number of classes and size of the documents *(Table 1*).

corpus	task	D	K	I	W	V
SemEval-2017 4A	message polarity	64,772	3	0.09	19	64,405
IMDB	movie review	348,415	10	0.03	325,6	115,073
Reuters ApteMod*	newswire topic	9,120	48	0.28	112,5	57,420

Table 1: D denotes the number of documents in the dataset, K the number of classes, I the class imbalance ratio (Tanwani & Farooq, 2009), W the average number of words per document, V the total vocabulary size respectively.

The first two datasets, SemEval 2017 task 4A (Rosenthal et al., 2017) comprising short social media text (Twitter), and IMDB movie reviews (Diao et al., 2014), share the task of sentiment classification which is characterized by ordinal labels ("negative-neutral-positive" and "1-10"). For both corpora we use pre-defined splits. We have modified the final corpus, Reuters-21578 ("ApteMod" version) (Apté et al., 1994), originally a multi-label text categorization dataset with 90 possible labels, to only contain documents with a single label attribution with a minimum label frequency of 3 in the corpus. This ensures an output space with mutually-exclusive and unambiguous classes (Liu et al., 2019). We generate randomized (seed 42) stratified splits of 65% for training, 15% validation and 20% for testing.

Base architecture We use 1-D Convolutional neural networks (TextCNN) for text classification, following the model structure of Kim (2014). We chose this architecture for its comparative simplicity and solid performance on a range of text classification tasks. Even as a light-weight model, it can deal with feeding in text sequences of varying sizes and learning n-gram-like structures over word embeddings, allowing a fair comparison across text datasets. An extensive hyperparameter study determined that regularization does not impact performance much (Zhang & Wallace, 2015).

Hyperparameters Our choice of hyperparameters is heavily based on Zhang & Wallace (2015) and Xiao & Wang (2019), where we propose one static setting. We constrain the input vocabulary to the 20,000 most frequent words, retain the original document lengths, upon which 300-D embeddings are uniformly initialized, and *UNK*/*PAD* tokens are masked throughout. Our TextCNN uses three different kernels (3,4,5) with 100 feature maps per kernel followed by a max pooling operation.

Additionally, for uncertainty estimation goals we apply dropout (Srivastava et al., 2014) with a rate of 0.5 after each non-linear layer, i.e., after each convolutional layer and before passing logits to the output layer, and adopt a global weight decay rate of 1e-4 (Krogh & Hertz, 1992; Loshchilov & Hutter, 2017). During training, Adam optimizes a categorical cross-entropy or heteroscedastic loss (see section 3.2) with a learning rate of 1e-3; batch size is set to 32 and training runs 45 epochs for 2000 iterations per epoch. At evaluation time, we estimate predictive mean and uncertainties by drawing T samples from the approximated posterior distribution. We have empirically set T to 10.

3.2. Quantification methods

In our experiments we replicate the model and data uncertainty quantification methods for text classification from Xiao & Wang (2019).

Quantifying "epistemic" *model uncertainty* using MC Dropout involves applying dropout both during training and evaluation. In the latter case, T stochastic weights are sampled from the variational Bernoulli distribution $\hat{\theta}_t \sim q(\theta)$ in order to calculate the lower-order moments of the approximate Gaussian posterior, respectively the predictive mean and variance.

To estimate input-dependent, "heteroscedastic aleatoric", *data uncertainty* we slightly modify the model's architecture and objective function following Kendall & Gal (2017). Firstly, the output layer of model $f_{\hat{\theta}}$ is extended with a set of learnable variance variables $\boldsymbol{\sigma}$ per unique class output. The model's output logits, $\mathbf{u}$, are sampled from the stochastic output layer parametrized by $\mathcal{N}(f_{\hat{\theta}}(x), diag(\boldsymbol{\sigma}(x)^2))$. This model adaptation will be referred to as the *heteroscedastic model*. *Fig. 1* (Appendix) visualizes the difference in architecture with shared hyperparameters. Next, we incorporate a residual *heteroscedastic loss*:

$$\mathcal{L}_{\text{clf}}(\hat{\theta}) = \sum_{i=1}^{N} \log \frac{1}{T} \sum_{t=1}^{T} \exp \left(\mathbf{u}_{i,c}^{(t)} - \log \sum_{k} \exp \mathbf{u}_{i,k}^{(t)} \right) + \log T \quad (1)$$

with N the number of training examples passing through an instance t of the model $f_{\hat{\theta}_t}(x) + \boldsymbol{\sigma}^{(t)}$ to generate for example i a sampled logit vector $\mathbf{u}_i^t$, where predicted value for class k, $\mathbf{u}_{i,k}^{(t)}$, and c the index of the ground truth class. By learning to predict log variance with T dropout-masked samples, the model will be able to predict high variance (uncertainty) for inputs where the predictive mean is far removed from the true observation, which by design has a smaller effect on the total loss. This uncertainty modelling method is referred to as *Learned Loss Attenuation*.

Below follows a categorization of the uncertainty quantities within the scope of the experiments. To estimate for a new test sample x^* the prediction and uncertainty of model $f_{\hat{\theta}}(x^*)$ we typically seek to obtain the predictive posterior distribution $P(y^*|x^*, \hat{\theta})$ over class membership probabilities with $y_k^* \in \{1, \ldots, K\}$. Particularly when using MC Dropout at inference time we presume $P(y^*|x^*, \hat{\theta}) \approx \frac{1}{T} \sum_{t=1}^{T} P(y^*|x^*, \hat{\theta}_t)$, with prediction obtained after applying softmax function for sample t, $\hat{p}_t = P(y^*|x^*, \hat{\theta}_t)$, and predictive mean $\bar{p} = \frac{1}{T} \sum_{t=1}^{T} \hat{p}_t$.

Quantity	*Formula*
Softmax-score	$S = \underset{k}{\text{argmax}} \frac{\exp f_{\hat{\theta},k}(x^*)}{\sum_{i=1}^{K} \exp f_{\hat{\theta},i}(x^*)}$
Predictive Entropy	$H = -\sum_{k=1}^{K} P(y_k|x^*, \hat{\theta}) \log P(y_k|x^*, \hat{\theta})$
Model Uncertainty	$\hat{\sigma}_{model} = \frac{1}{T} \sum_{t=1}^{T} (\hat{p}_t - \bar{p})^2$
Data Uncertainty	$\hat{\sigma}_{data} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\sigma}_k^{(t)}(x^*)$

3.3. Benchmarking uncertainty quantities

Setups *Table 2* summarizes the model setups. During training models [1-3] are optimized by cross-entropy minimization, whereas models [4-5] also optimize the heteroscedastic loss. During testing, models [1,2,4] provide simple predictions, whereas models [3,5] estimate prediction and uncertainties from T stochastic forward samples. For all classification models we can compute the softmax-score and predictive entropy. When changing to the heteroscedastic architecture, we then quantify data uncertainty, and when stochastically sampling using MC dropout, model uncertainty is quantified. More specifically, data uncertainty in models [4,5] is quantified with as surrogate the average over variance logits $\boldsymbol{\sigma}$. Model uncertainty in models [3,5] is quantified by calculating the average softmax variance over the predictive mean from MC samples. Throughout the rest of the work we will respectively refer to the model setups as **N**o **D**ropout (ND), **B**aseline (B), **M**odel **U**ncertainty (MU), **D**ata **U**ncertainty (DU), **D**ata & **M**odel **U**ncertainty (DMU).

	Monte Carlo dropout		
Architecture	deterministic	stochastic	***no dropout**
softmax	2	3	1
heteroscedastic	4	5	

Table 2: A summary of the 5 model setups, varying across the base architecture and if MC Dropout sampling is activated. For a fair comparison, we include a softmax model without any dropout.

Novelty detection - *how well can the model identify and communicate uncertainty on samples of novel class distributions?* In the worst case, classifiers "fail silently" and wrongly attribute high confidence to an in-distribution class. (Goodfellow et al., 2014; Amodei et al., 2016). In the best case, the model either lowers its confidence or signals uncertainty. Prior work hypothesizes model uncertainty to be mostly impacted (Kendall & Gal, 2017; Leibig et al., 2017). With this experiment we simulate the conditions of novel class data by removing a single class during training. For the Reuters dataset, this class is very distinct from the remaining classes, i.e., (i) by not appearing often in the originally multi-label annotated dataset jointly with the remaining classes, and (ii) occurring frequently enough to guarantee representative results. Since the dataset has been explicitly annotated for multi-label, we can draw statistics on the label co-occurrence rates, and find that the second-most frequent topic "Acquisitions" (id:0) occurs in 94% of documents as a single topic. This makes it an ideal candidate for testing novelty detection in a multi-class text classification setting. For both sentiment classification datasets, we isolate the middle class (respectively, "neutral" and rating "5" out of the 10 ratings) from training and expect the models to allocate prediction mass to a label close to the holdout class (ratings "4" or "6").

4. Experimental results and discussion

Text classification results The results in *Table3* show that for Reuters newswire classification, learned loss attenuation and applying the MC Dropout procedure improves all metrics. Surprisingly, uncertainty quantification does not guarantee classification improvement over a model without dropout regularization in the case of both sentiment classification tasks. Relative to the other regularized models, modelling uncertainty does marginally improve performance. Xiao & Wang (2019) conjectures a limited classification output space to be the reason for only marginal classification increase when modelling uncertainty. While this also holds

Measure / Method	Acc	MSE (↓)	F1(m)	F1(M)	NLL(↓)	ECE(↓)	Brier(↓)	Softmax (μ)	Entropy (μ)	MU (μ)	DU (μ)
SemEval No Dropout	**0.5831**	**0.5389**	**0.5766**	0.5618	0.9979	0.1494	0.5728	0.7325	0.901	/	/
SemEval Baseline	0.568	0.5923	0.5652	0.5525	**0.9158**	**0.0195**	**0.5491**	0.5838	1.2829	/	/
SemEval Model Uncertainty	0.5712	0.5785	0.5666	0.5526	0.9601	0.0979	0.5653	0.6692	1.0765	0.115	/
SemEval Data Uncertainty	0.567	0.5928	0.5657	0.5554	0.9172	0.0245	0.55	0.5895	1.2718	/	0.0181
SemEval DMU	0.5808	0.5591	0.5761	**0.563**	0.9466	0.0915	0.558	0.6714	1.0741	0.0055	0.0143
IMDB No Dropout	**0.4164**	**3.0908**	**0.3958**	**0.3563**	**1.4786**	0.0139	**0.6807**	0.4208	2.142	/	/
IMDB Baseline	0.405	3.5007	0.3724	0.3287	1.5641	0.0671	0.7034	0.3379	2.5217	/	/
IMDB Model Uncertainty	0.4069	3.4787	0.3787	0.3349	1.5247	**0.0124**	0.6932	0.3954	2.2661	0.1426	/
IMDB Data Uncertainty	0.4067	3.3854	0.377	0.3358	1.558	0.0536	0.7022	0.3531	2.4685	/	0.0033
IMDB DMU	0.4071	3.3377	0.3774	0.3371	1.5263	0.0148	0.6945	0.4131	2.2109	0.0026	0.0026
Reuters No Dropout	0.923	30.2168	0.9145	0.6464	0.3329	0.0265	0.1147	0.9403	0.3308	/	/
Reuters Baseline	0.9293	28.1707	0.9228	**0.7193**	0.3364	0.0337	0.1123	0.8978	0.6704	/	/
Reuters Model Uncertainty	0.9277	27.8746	0.9209	0.7131	0.3311	**0.0147**	0.1054	0.9351	0.3667	0.052	/
Reuters Data Uncertainty	0.9301	**25.0199**	0.9243	0.7184	0.3286	0.0314	0.1112	0.8993	0.6555	/	0.0246
Reuters DMU	**0.932**	26.0086	**0.9255**	0.6957	**0.319**	0.016	**0.1023**	0.9369	0.3539	0.0003	0.0087

Table 3: This table reports on the effectiveness of the text classification using the 3 datasets.We report all metrics on the test data, respectively *classification* scores: Accuracy, Mean-Squared Error, weighted and macro F1; *calibration* metrics: Negative Log Likelihood, Expected Calibration Error (Guo et al., 2017) and Brier score (Brier, 1950); *uncertainty* measures when available and averaged over all samples, Softmax-score, Predictive Entropy, Model Uncertainty and Data Uncertainty.

for our results, we additionally suspect that the ambiguous and complex class decision boundaries counterbalance the benefits of quantifying uncertainty. Another clear observation is that applying MC Dropout increments average softmax-score, providing calibration for IMDB and Reuters, yet raising Expected Calibration Error for SemEval.

Since modelling uncertainty does not guarantee strictly increasing accuracy or calibration, we have formulated the novelty experiment to tease out what exactly the uncertainty quantities can measure.

Novelty detection The main results are collected quantitatively in *Table 4* and visually in *Fig. 2* (Appendix).

While in some cases model uncertainty does significantly increase when presented with novel class data, our experiments do not support its hypothesized top ranking. Overall, the same trend holds over the datasets with predictive entropy resulting in the most robust novelty uncertainty estimate. While the complexity of OOD-tasks differs greatly with respect to classification boundaries, the quantities measured are impacted similarly given only small relative differences in quantity rank across datasets. The visual results detail how the different quantities are impacted.

Generally, adding dropout regularization (B) and modelling data uncertainty (DU) have the most positive effect on entropy, closely followed by softmax. As already indicated in the standard results, MC Dropout increments the average softmax-score of the most probable class, which deteriorates the measure's ability to discriminate novel samples. Overall, the experiment demonstrates that quantifying model and data uncertainty does not yield a good estimator for novel class presence in input data in contrast to predictive entropy.

Dataset	SemEval		IMDB		Reuters		
measure	PCC	Rank	PCC	Rank	PCC	Rank	**Avg Rank**
nodropout softmax-score	0.0922*	12	0.1035*	10	0.2894*	12	12
nodropout entropy	-0.1115*	11	-0.1339*	6	-0.3381*	11	10
baseline softmax-score	0.1419*	6	0.1332*	8	0.6066*	5	6
baseline entropy	-0.1590*	**1**	-0.1636*	**1**	-0.6367*	**3**	**1**
MU softmax-score	0.1339*	8	0.1304*	9	0.5270*	9	9
MU entropy	-0.1571*	4	-0.1471*	**3**	-0.5732*	6	4
MU model uncertainty	-0.0734*	13	0.0052	14	0.0027	14	14
DU softmax-score	0.1396*	7	0.1414*	5	0.6370*	**2**	5
DU entropy	-0.1590*	**2**	-0.1595*	**2**	-0.6558*	**1**	**1**
DU data uncertainty	-0.1465*	5	0.0106	12	-0.5539*	8	8
DMU softmax-score	0.1298*	9	0.1336*	7	0.5677*	7	7
DMU entropy	-0.1585*	**3**	-0.1440*	4	-0.6118*	4	**3**
DMU data uncertainty	-0.0170	14	-0.0546*	11	-0.0849*	13	13
DMU model uncertainty	-0.1253*	10	0.0098	13	-0.4635*	10	11

Table 4: We report the Pearson Correlation Coefficient and p-value<0.05 with * between uncertainty values and binary variable IID-OOD. Higher absolute correlation score points to stronger association of uncertainty and novelty detection. The final rank over datasets confirms the superior robustness of predictive entropy.

5. Conclusion

We have evaluated predictive uncertainty and their value in novel class detection in a text classification setting. This study has led to the following conclusions:

- Necessary regularization for uncertainty estimation proves to not always guarantee increase in model performance. This is an important insight to be considered when adopting uncertainty quantification.
- MC Dropout-based uncertainty quantities do not perform well under extrapolation. In the novelty detection experiment predictive entropy and softmax-score outperform current data and model uncertainty metrics.

Admittedly, it is hard to compare uncertainty estimates without a controlled setting. In our experiments we adopted the logic of evaluating uncertainty using a well-tuned reliable base classification model on a representative set of real-world text classification datasets. As such we have empirically verified the drawbacks and the applicability scope of uncertainty methods, most notably their underestimation of uncertainty under novel class distribution shift.

Going forward, we seek to extend our methodology with (i) a larger scope of uncertainty estimation methods and (ii) more probing experiments covering situations where we expect predictive uncertainty to be crucial.

Acknowledgements

The authors acknowledge the financial support of VLAIO (Flemish Innovation & Entrepreneurship) through the Baekeland PhD mandate (HBC.2019.2604).

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P. F., Schulman, J., and Mané, D. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

Apté, C., Damerau, F., and Weiss, S. M. Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, 1994.

Bishop, C. M. Novelty Detection and Neural Network Validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4):217–222, 1994.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Diao, Q., Qiu, M., Wu, C.-Y., Smola, A. J., Jiang, J., and Wang, C. Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 193–202, 2014.

Dürr, O., Murina, E., Siegismund, D., Tolkachev, V., Steigele, S., and Sick, B. Know When You Don't Know: A Robust Deep Learning Approach in the Presence of Unknown Phenotypes. *Assay and drug development technologies*, 16(6):343–349, 2018.

Foong, A. Y. K., Burt, D. R., Li, Y., and Turner, R. E. On the Expressiveness of Approximate Inference in Bayesian Neural Networks, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Gal, Y., Hron, J., and Kendall, A. Concrete Dropout. In *Advances in neural information processing systems*, pp. 3581–3590, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

Hendrycks, D. and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv preprint arXiv:1610.02136*, 2016.

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

Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv preprint arXiv:1408.5882*, 2014.

Krogh, A. and Hertz, J. A. A Simple Weight Decay Can Improve Generalization. In Moody, J. E., Hanson, S. J., and Lippmann, R. P. (eds.), *Advances in Neural Information Processing Systems 4*, pp. 950–957. Morgan-Kaufmann, 1992.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging Uncertainty Information from Deep Neural Networks for Disease Detection. *Scientific reports*, 7(1): 1–14, 2017.

Li, Y. and Gal, Y. Dropout Inference in Bayesian Neural Networks with Alpha-divergences. *arXiv preprint arXiv:1703.02914*, 2017.

Liu, H., Burnap, P., Alorainy, W., and Williams, M. A Fuzzy Approach to Text Classification With Two-Stage Training for Ambiguous Instances. 6:227–240, 04 2019. doi: 10.1109/TCSS.2019.2892037.

Loshchilov, I. and Hutter, F. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you Trust your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Advances in Neural Information Processing Systems*, pp. 13991–14002, 2019.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. A Review of Novelty Detection. *Signal Processing*, 99:215–249, 2014.

Rosenthal, S., Farra, N., and Nakov, P. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp. 502–518, 2017.

Shannon, C. E. A Mathematical Theory of Communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Tanwani, A. K. and Farooq, M. Classification Potential vs. Classification Accuracy: a Comprehensive Study of Evolutionary Algorithms with Biomedical Datasets. In *Learning Classifier Systems*, pp. 127–144. Springer, 2009.

Vernekar, S., Gaurav, A., Abdelzad, V., Denouden, T., Salay, R., and Czarnecki, K. Out-of-distribution Detection in Classifiers via Generation. *arXiv preprint arXiv:1910.04241*, 2019.

Wilson, A. G. The Case for Bayesian Deep Learning. *arXiv preprint arXiv:2001.10995*, 2020.

Xiao, Y. and Wang, W. Y. Quantifying Uncertainties in Natural Language Processing Tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7322–7329, 2019.

Zaragoza, H. and d'Alché Buc, F. Confidence Measures for Neural Network Classifiers. In *Proceedings of the Seventh Int. Conf. Information Processing and Management of Uncertainty in Knowlegde Based Systems*, 1998.

Zhang, X., Chen, F., Lu, C.-T., and Ramakrishnan, N. Mitigating Uncertainty in Document Classification. *arXiv preprint arXiv:1907.07590*, 2019.

Zhang, Y. and Wallace, B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *arXiv preprint arXiv:1510.03820*, 2015.

Appendix

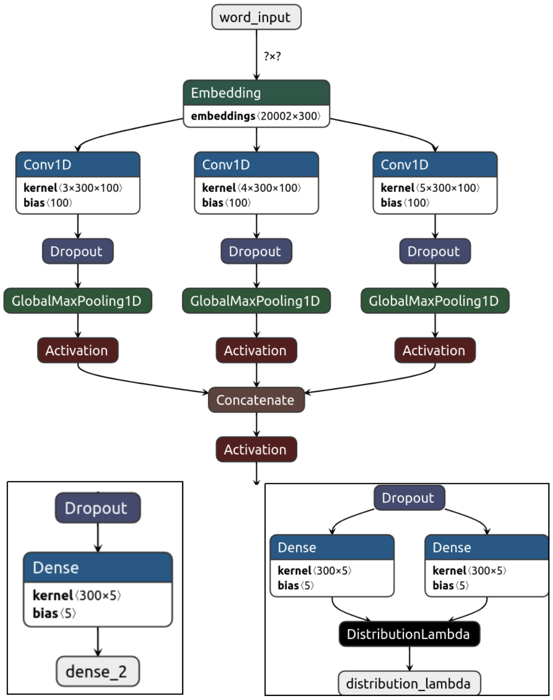
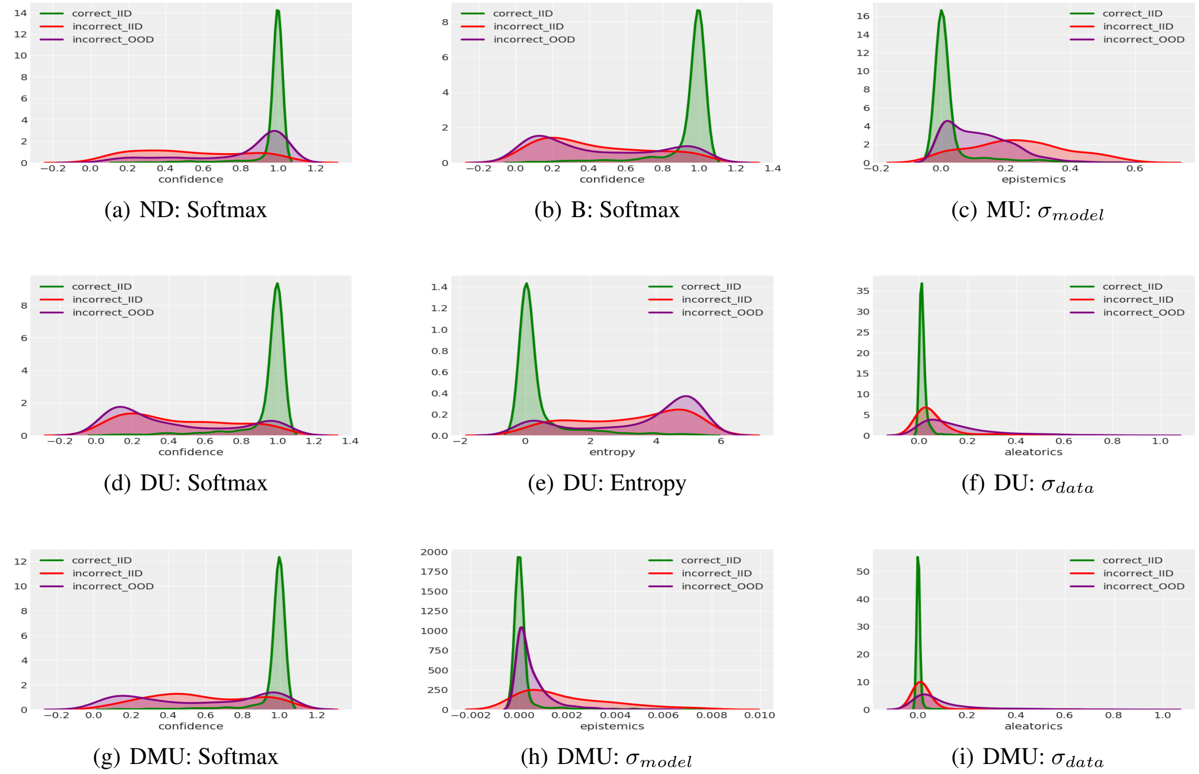


Figure 1: Visualization of the two architectures used with appropriate hyperparametrization (K=5). The left architecture denotes the standard *softmax* model, and on the right, the *heteroscedastic* model outputs a Normal distribution $\mathcal{N}(\boldsymbol{\mu}(x), diag(\boldsymbol{\sigma}(x)^2)$ parametrizing mean and variance by the logits coming from two separate preceding feedforward layers. Visualization source: Netron v 4.01 (Lutz Roeder)



(a) ND: Softmax

(b) B: Softmax

(c) MU: σ_{model}

(d) DU: Softmax

(e) DU: Entropy

(f) DU: σ_{data}

(g) DMU: Softmax

(h) DMU: σ_{model}

(i) DMU: σ_{data}

Figure 2: A selection of most interesting Gaussian kernel density plots over (abbreviated) model setup metrics evaluated on Reuters. Each plot captures probabilistic density over correct IID (green), incorrect IID (red) and OOD (purple).
(a) demonstrates "overconfidence" due to no regularization, which already improves in (b). Model uncertainty (c) seemingly captures OOD samples, yet they are unrecognizable from wrong IID. Predictive entropy (e) demonstrates a clearer separation, at least with respect to correct IID samples, with the same trend to a lesser degree in (d). Data uncertainty (f) and (i) show no awareness on novel class samples. The combined methods and quantities (g), (h) and (i) perform worse overall.