ICML 2020 Workshop on

Uncertainty & Robustness in Deep Learning

Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification

Jordy Van Landeghem

17-07-2020

Jordy Van Landeghem, Matthew Blaschko, Bertrand Anckaert, and Marie-Francine Moens:
Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification.
ICML Workshop on Uncertainty and Robustness in Deep Learning, 2020.

Predictive uncertainty is a key enabler for reliable ML

Decision-making under Predictive Uncertainty

In-distribution

From: jack.dunn@gmail.com
To: customer_admin@insurco.com
Subject: stop car policy 12-3456-789

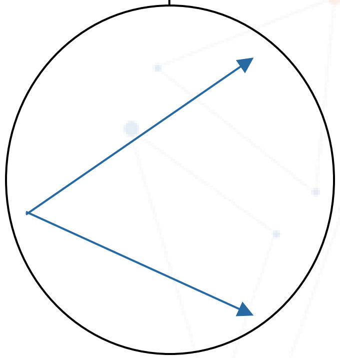
Dear,

I would like to end the policy for my vehicle 1-CHA-123. The car has been sold and the license plate returned to the authorities (proof attached).

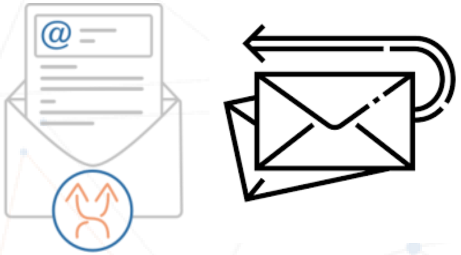
Kindly refund the unused portion of my premium.

Best,
Jack

label	prediction	confidence
process	car policy cancellation	99%
policy number	12-3456-789	95%
license plate	1CHA123	98%



Automate action



Novel class

From : jeff.smith@gmail.com
To : customer_admin@insurco.com
Subject : Jeff Smith hobby drone coverage

Dear,

My car is already covered with your insurance company under the policy with number 23-4567-890.
Now, I would like to buy insurance coverage for my new hobby drone "DJI Mavic 2LBZ-548".
See attached the purchase receipt and below an illustration of the device and components.

Will you send me a proposal with monthly coverage fees?
Kind regards,

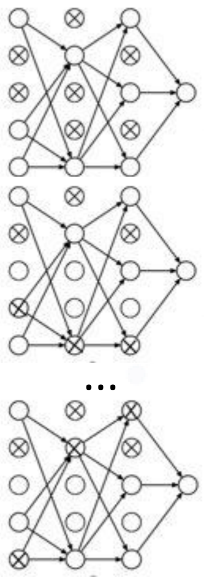
Jeff

label	prediction	confidence
process	car policy contract start	98%
policy number	23-4567-890	95%
license plate	2LBZ-548	75%

Manual review

! **Catastrophically overconfident**

Predictive Uncertainty in practice



Monte Carlo Dropout *(Gal & Ghahramani 2016)*

Algorithm 1: MCdropout

Input: data x^*, encoder $g(\cdot)$, prediction network $h(\cdot)$, dropout probability p, number of iterations B

Output: prediction $\hat{y}^*_{mc}$, uncertainty η_1

1: **for** $b = 1$ to B **do**

2: $e^*_{(b)} \leftarrow \mathit{VariationalDropout}(g(x^*), p)$

3: $z^*_{(b)} \leftarrow \text{Concatenate}(e^*_{(b)}, \text{extFeatures})$

4: $\hat{y}^*_{(b)} \leftarrow \mathit{Dropout}\ (h(z^*_{(b)}), p)$

5: **end for**

// prediction

6: $\hat{y}^*_{mc} \leftarrow \frac{1}{B}\sum_{b=1}^{B} \hat{y}^*_{(b)}$

// model uncertainty and misspecification

7: $\eta_1^2 \leftarrow \frac{1}{B}\sum_{b=1}^{B} (\hat{y}^*_{(b)} - \hat{y}^*)^2$

8: **return** $\hat{y}^*_{mc}, \eta_1$

I. **Regularization** (dropout, L2 weight decay)
II. **Stochastic output layer** $\mathcal{N}(f_{\hat{\theta}}(x), diag(\boldsymbol{\sigma}(x)^2))$
III. **Attenuated learned loss**

Residual heteroscedastic loss & extensions *(Kendall & Gal 2017; Xiao & Wang 2019)*

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

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.

```python
def attenuated_learned_loss(y_true, y_pred, T=10):
    sampled = y_pred.sample(T)
    sampled_loss = tf.stack(
        [tf.nn.softmax_cross_entropy_with_logits(y_true, sampled[t]) for t in range(T)]
    )  # T x batch_size x maxlen (x labels)
    batch_losses = logsumexp(-sampled_loss)
    likelihood_loss = -tf.reduce_mean(batch_losses) + tf.math.log(tf.cast(T, "float32"))
    return likelihood_loss
```

Uncertainty quantification

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

Methodology & experiments

<u>Research question</u>:

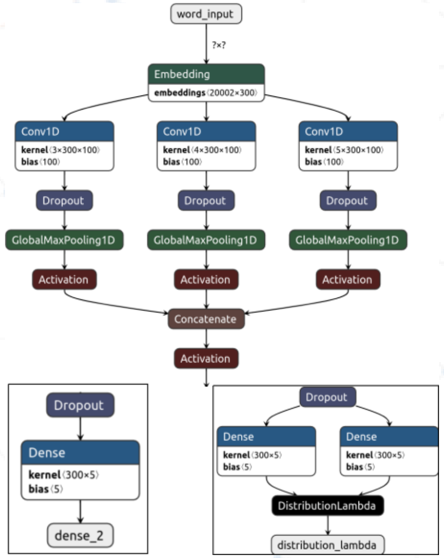
How **reliable** are Monte Carlo Dropout-based uncertainty estimates for unsupervised detection of novel class data in text classification?

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

<u>Methodology</u>:

- 3 real-world multi-class news and sentiment classification datasets
- 1-D ConvNets for text classification (*Kim 2014*)
- 5 uncertainty quantification model setups
- Robustness protocol of *leave-one-class-out*

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



Key findings (1/2)

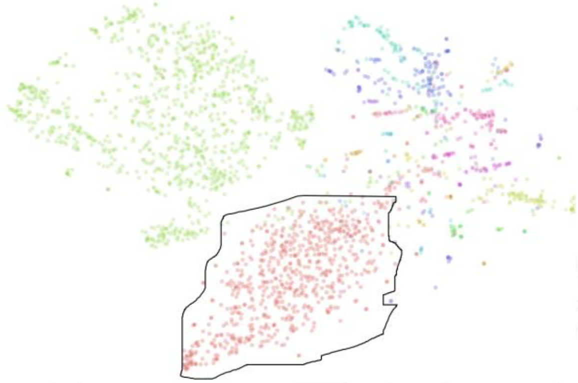
- Necessary regularization for uncertainty estimation proves to **not always guarantee increase in model performance.**

Method \ Measure	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

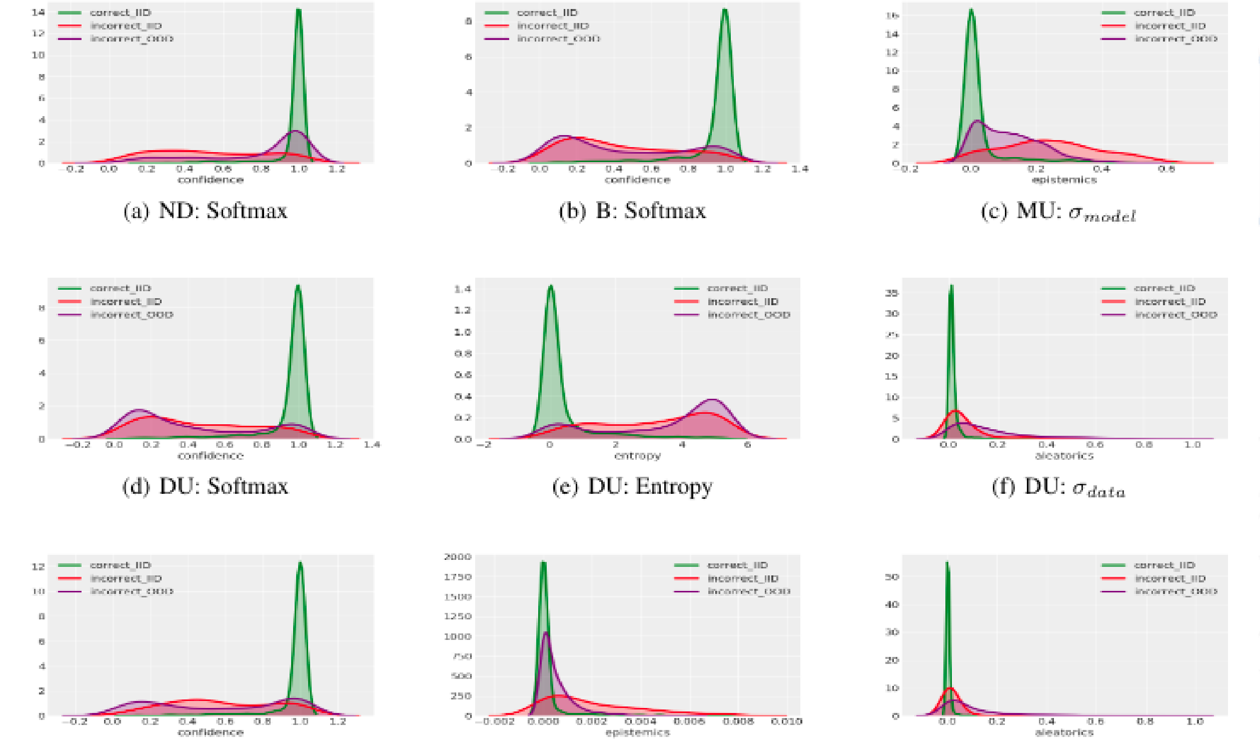
Table 3: This table reports on the effectiveness of the text classification using the 3 datasets.We report all metrics on the test data, respectively *classification* scores: Accuracy, Mean-Squared Error, weighted and macro F1; *calibration* metrics: Negative Log Likelihood, Expected Calibration Error (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.

Key findings (2/2)

- MC Dropout and extensions **underestimate uncertainty**
- Predictive entropy demonstrates superior performance



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



(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}

Taking this forward

a. underrepresentation of NLP in Bayesian Deep Learning research
b. embedding in a theoretical framework
   - What does uncertainty represent in an NLP task context?
   - How does uncertainty manifest?
   - What forms of uncertainty require capture?
   - What architectures in combination with regularization methods are best suited?
c. extended uncertainty protocols for benchmarking
d. extended uncertainty quantification methods

Thank you! Questions?

Poster #133

Predictive Uncertainty for Probabilistic Novelty Detection in Text Classification

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

Jordy Van Landeghem
jordy@contract.fit

Bibtex

```
@inproceedings{VanLandeghem2020a,
  TITLE = {Predictive Uncertainty for Probabilistic Novelty
Detection in Text Classification},
  AUTHOR = {Van Landeghem, Jordy and Blaschko, Matthew B. and
Anckaert, Bertrand and Moens, Marie-Francine},
  BOOKTITLE = {ICML Workshop on Uncertainty and Robustness in Deep
Learning},
  YEAR = {2020},
}
```

Keywords: Predictive Uncertainty, Text Classification, Unsupervised Novelty Detection, Monte Carlo Dropout

Backup slides

References

- 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.
- 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.
- 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.

Questions to be expected

Q	A
Why so little datasets?	three real-world text corpora with differing number of classes and size of documents. Additional focus was the classification complexity.
Why Reuters?	For novelty detection, shows easy separable classes. Multi-label annotations ensure class separability information. Then why would uncertainty not be able to communicate on it? Would expect it to go wrong for classification in a spectrum/ordinal scale. -> going for semi-synthetic experiment not as good as an “MNIST” for NLP ☺
Softmax thresholding and calibration	-> perfect calibration requires less uncertainty quantification. Yes, but data uncertainty can communicate on label noise, which might be captured less by calibration.
Little fine-tuning of regularization parameters	Admittedly, we would have done more fine-tuning on these parameters. However, striving for SoTa was not the goal here. We tried to find an OK out-of-the-box setting.
Why no MI?	

Research Question & Contributions

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.

- 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.

Backup pictures