Uncertainty-Aware Natural Language Inference with Stochastic Weight Averaging

Anonymous Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymouser Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymousest Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

This paper introduces Bayesian uncertainty modeling using Stochastic Weight Averaging-Gaussian (SWAG) in Natural Language Understanding (NLU) tasks. We apply the approach to standard tasks in natural language inference (NLI) and demonstrate the effectiveness of the method in terms of prediction accuracy and correlation with human annotation disagreements. We argue that the uncertainty representations in SWAG better reflect subjective interpretation and the natural variation that is also present in human language understanding. The results reveal the importance of uncertainty modeling, an often neglected aspect of neural language modeling, in NLU tasks.

1 Introduction

Arguably, human language understanding is not objective nor deterministic. The same utterance or text can be interpreted in different ways by different people depending on their language standards, background knowledge and world views, the linguistic context, as well as the situation in which the utterance or text appears. This uncertainty about potential readings is typically not modeled in Natural Language Understanding (NLU) research and is often ignored in NLU benchmarks and datasets. Instead, they usually assign a single interpretation as a gold standard to be predicted by an artificial system ignoring the inherent ambiguity of language and potential disagreements that humans arrive at.

Some datasets like SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) do, however, contain information about different readings in the form of annotation disagreement. These datasets include the labels from five different rounds of annotation which show in some cases clear disagreement about the correct label for the sentence pair. Those labeling discrepancies can certainly be a result of annotation mistakes but more commonly they arise from differences in understanding the task, the given information and how it relates to world knowledge and personal experience.

Moving towards uncertainty-aware neural language models, we present our initial results using Stochastic Weight Averaging (SWA) (Izmailov et al., 2018) and SWA-Gaussian (SWAG) (Maddox et al., 2019) on the task of Natural Language Inference. SWAG provides a scalable approach to calibrate neural networks and to model uncertainty presentations and is straightforward to apply with standard neural architectures. Our study addresses the two main questions:

- How does uncertainty modeling using SWAG influence prediction performance and generalization in NLI tasks?
- How well does the calibrated model reflect human disagreement and annotation variance?

In this paper, we first test the performance of SWA and SWAG in SNLI and MNLI tasks. We then study if adding weight averaging improves the generalization power of NLI models as tested through cross-dataset experiments. Finally, we analyse the probability distributions from SWA and SWAG to test how well the model uncertainty corresponds to annotator disagreements.

2 Background and Related Work

2.1 Uncertainty in human annotations

In a recent position paper Plank (2022) argue that instead of taking human label variation as a problem, we should embrace it as an opportunity and

take it into consideration in all the steps of the ML pipeline: data, modeling and evaluation. The paper provides a comprehensive survey of research on (i) reasons for human label variation, (ii) modeling human label variation, and (iii) evaluating with human label variation.

Pavlick and Kwiatkowski (2019) studied human disagreements in NLI tasks and argue that we should move to an evaluation objective that more closely corresponds to the natural interpretation variance that exists in data. Such a move would require that NLU models be properly calibrated to reflect the distribution we can expect and, hence, move to a more natural inference engine.

Chen et al. (2020) propose Uncertain NLI (UNLI), a task that moves away from categorical labels into probabilistic values. They use a scalar regression model and show that the model predictions correlate with human judgement.

2.2 Representing Model Uncertainty

The approach to uncertainty modeling that we consider is related to the well-established technique of model ensembling. Stochastic optimization procedures applied in training deep neural networks are non-deterministic and depend on hyper-parameters and initial seeds. Ensembles have been used as a pragmatic solution to average over several solutions, and the positive impact on model performance pushed ensembling into the standard toolbox of deep learning. Related to ensembling is the technique of checkpoint averaging (refer to e.g. Gao et al., 2022), which is also known to improve performance.

Intuitively, ensembles and checkpoint averages also reflect the idea of different views and interpretations of the data and, therefore, provide a framework for uncertainty modeling. SWA and SWAG build on that idea, and SWAG provides a generic and efficient approach for approximating Bayesian uncertainty and model calibration.

SWA (Izmailov et al., 2018) is a checkpoint averaging method that tracks the optimization trajectory for a model during training, using the average of encountered values as the eventual parameters:

$$\theta_{\text{SWA}} = \frac{1}{T}\sum_{i=1}^{T}\theta_i, \tag{1}$$

with θ_{SWA} denoting the SWA solution for parameter θ after T epochs of training.

SWAG (Maddox et al., 2019) extends this method to estimate Gaussian posteriors for model parameters, by also estimating a covariance matrix for the parameters. For computational feasibility, a low-rank plus diagonal approximation to the covariance matrix is used:

$$\Sigma_{\text{low-rank}} \approx \frac{1}{T-1}\sum_{i=1}^{T}(\theta_i - \hat{\theta}_i)(\theta_i - \hat{\theta}_i)^T \tag{2}$$

$$\Sigma_{\text{diag}} = \text{diag}(\frac{1}{T}\sum_{i=1}^{T}\theta_i^2 - \theta_{\text{SWA}}^2) \tag{3}$$

where $\hat{\theta}_i$ in (2) is the running estimate of the parameters' mean obtained from the first i samples. The resulting posterior approximations are given by

$$\theta_{\text{SWAG}} \sim \mathcal{N}(\theta_{\text{SWA}}, \frac{1}{2}(\Sigma_{\text{diag}} + \Sigma_{\text{low-rank}})). \tag{4}$$

Once the posteriors are thus approximated, in test time, the model is utilized by sampling from the approximated posteriors for N times, and taking the average of the predicted distributions from these samples as the answer of the model.

One of the advantages of SWAG is the possibility to seamlessly start with any pre-trained solution. Approximating the posterior is then done during fine-tuning without the need to change the underlying model.

3 Experiments

We test the performance of SWA and SWAG on the natural language inference task using three NLI datasets, including cross-dataset experiments, and study the effect on both hard and soft labeling.

3.1 Datasets

We use Stanford Natural Language Inference corpus (SNLI) (Bowman et al., 2015) and Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2018) as the datasets in our experiments. We also study cross-dataset generalisation capability of the model with and without weight averaging. For those experiments we also include SICK (Marelli et al., 2014) as a test set. In cross-dataset generalization experiments we first fine-tune the model with a training data from one NLI dataset (e.g. SNLI) and then test with a test set from another NLI dataset (e.g. MNLI-mm).

SNLI is a dataset of 570k sentence pairs which have been manually labeled with entailment, contradiction, and neutral labels. The source for the premise sentences in SNLI were image captions from the Flickr30k corpus (Young et al., 2014).

MNLI is made of 433k sentence pairs labeled with entailment, contradiction and neutral, containing examples from ten genres of written and spoken English. Five of the genres are included in the training set. The development and test sets have been split into matched (MNLI-m) and mismatched (MNLI-mm) sets, where the former includes only sentences from the same genres as the training data, and the latter includes genres not present in the training data.[1]

SICK includes 9,840 examples with logical inference (negation, conjunction, disjunction, apposition, relative clauses, etc.). The dataset was constructed automatically by taking pairs of sentences from a random subset of the 8K ImageFlickr (Young et al., 2014) and the SemEval 2012 STS MSRVideo Description (Agirre et al., 2012) datasets by using rule-based approach to construct examples for the different logical inference types.

3.2 Methods

In all the experiments we fine tune a pre-trained RoBERTa-base model (Liu et al., 2019) from the Hugging Face Transformers library (Wolf et al., 2020). As a common practice in the NLI tasks, we use the majority-vote gold labels for training even if multiple annotations are available.

We add stochastic weight averaging to the RoBERTa model by using the SWA implementation from PyTorch 1.12[2] and the SWAG implementation by (Maddox et al., 2019)[3]. To study how well SWA and SWAG perform in NLI as compared to a baseline model, we ran the same fine-tuning with SNLI and MNLI datasets utilizing SWA and SWAG for weight averaging.

[1] As the test data for MNLI have not been made publicly available, we use the development sets when reporting the results for MNLI.

[2] https://pytorch.org/docs/1.12/optim.html#stochastic-weight-averaging

[3] https://github.com/wjmaddox/swa_gaussian

Dataset	Method	Acc (%)	SD	Δ
SNLI	base	90.80	0.26	-
SNLI	SWA	91.47	0.24	+0.67
SNLI	SWAG	**91.59**	0.14	**+0.79**
MNLI-m	base	86.53	0.20	-
MNLI-m	SWA	87.60	0.19	+1.07
MNLI-m	SWAG	**87.76**	0.12	**+1.23**
MNLI-mm	base	86.31	0.26	-
MNLI-mm	SWA	87.34	0.29	+1.03
MNLI-mm	SWAG	**87.51**	0.19	**+1.20**

Table 1: Comparison of SWA and SWAG performance on NLI benchmarks (mean accuracy and standard deviation over 5 runs). Δ is the difference to the baseline result (base) with no weight averaging.

3.3 Results

The standard evaluation for the NLI task is the accuracy on aggregated gold labels. However, as two of the test data sets (from SNLI and MNLI) also contains multiple human annotations, we also use those for measuring the cross entropy of the predicted distribution on the human label distribution (soft labeling, e.g. Peterson et al., 2019; Pavlick and Kwiatkowski, 2019).

3.3.1 Accuracy

The basic classification results are in Table 1. We report average accuracies and standard deviation over 5 runs with different random seeds.

Both SWA and SWAG provide significant improvements over the baseline without weight averaging. SWAG performs slightly better than SWA across all the three experiments.

In order to test if weight averaging improves the generalization capability of NLI models, we further performed cross-dataset generalization tests following (Talman and Chatzikyriakidis, 2019). The results are reported in Table 2.

The results of cross-dataset experiments are slightly mixed: We do not notice a clear advantage of SWAG over SWA, but with the exception of training with MNLI and testing with SICK, we do notice improvement for weight averaging approaches as compared to the baseline. The performance on SICK drops significantly in all cases and the difference between the approaches is minimal, showing that the NLI training data is not a good fit for that benchmark.

The other cross-dataset results highlight the advantage of weight averaging, indicating that the

Dataset	Method	Acc (%)	SD	Δ
SNLI → MNLI-m	base	77.31	0.57	
SNLI → MNLI-m	SWA	**79.67**	0.37	**2.37**
SNLI → MNLI-m	SWAG	79.33	0.21	2.03
SNLI → MNLI-mm	base	77.40	0.78	
SNLI → MNLI-mm	SWA	**79.44**	0.19	**2.04**
SNLI → MNLI-mm	SWAG	79.24	0.29	1.84
SNLI → SICK	base	57.08	0.77	
SNLI → SICK	SWA	57.09	0.32	0.01
SNLI → SICK	SWAG	**57.17**	0.37	**0.08**
MNLI → SNLI	base	82.84	0.74	
MNLI → SNLI	SWA	84.15	0.35	1.31
MNLI → SNLI	SWAG	**84.45**	0.27	**1.61**
MNLI → SICK	base	**56.63**	0.94	
MNLI → SICK	SWA	56.17	0.60	-0.46
MNLI → SICK	SWAG	56.53	0.91	-0.10

Table 2: Cross-dataset experiments with and without weight averaging (mean accuracy and standard deviation over 5 runs with different random seeds), where the left hand side of the arrow is the training set and the right hand side is the testing set.

improved modeling of uncertainty can lead to better generalizations.

3.3.2 Cross Entropy

We also test how well weight averaging approaches can be used to model annotator disagreement and annotation uncertainty in the NLI testsets of SNLI and MNLI. These two datasets come with five different annotation labels for every data point, often with high disagreement between human annotators indicating inherently *confusing* data points with high aleatoric uncertainty (Der Kiureghian and Ditlevsen, 2009). For quantifying the goodness of fit of the model predictions, we calculate the cross entropy between the predicted and annotation distributions.[4]

Table 3 depicts the resulting cross entropy values, with lower values denoting more faithful predictions. SWA and SWAG result in consistently more similar distributions with that of annotations, complementing their overall better accuracy results (Section 3.3). In contrast to the accuracy results, here SWAG outperforms SWA in all cases, indicating that the Gaussian posterior helps to model the data uncertainty more accurately. The results also carry over to the cross-dataset experiments as shown on the table.

The comparison between system predictions

[4]Note that for the Baseline and SWA models, we consider the output from the eventual softmax function as the predicted distribution, while for the SWAG model, we use the average output distribution from $N = 20$ sampled models.

Dataset	Method	Cross Entropy	Δ
SNLI	base	0.83	
SNLI	SWA	0.75	-0.08
SNLI	SWAG	**0.69**	**-0.14**
MNLI-m	base	0.87	
MNLI-m	SWA	0.80	-0.07
MNLI-m	SWAG	**0.73**	**-0.14**
MNLI-mm	base	0.84	
MNLI-mm	SWA	0.77	-0.07
MNLI-mm	SWAG	**0.69**	**-0.15**
SNLI → MNLI-m	base	1.13	
SNLI → MNLI-m	SWA	0.90	-0.23
SNLI → MNLI-m	SWAG	**0.80**	**-0.33**
SNLI → MNLI-mm	base	1.12	
SNLI → MNLI-mm	SWA	0.88	-0.24
SNLI → MNLI-mm	SWAG	**0.79**	**-0.33**
MNLI → SNLI	base	1.04	
MNLI → SNLI	SWA	0.97	-0.07
MNLI → SNLI	SWAG	**0.89**	**-0.15**

Table 3: Comparison of cross entropies between data annotation distributions using base, SWA and SWAG methods. Δ is the difference to the baseline cross entropy values.

and annotator variation deserves some further analysis. Preliminary study (see examples in Appendix A) indicates that the prediction uncertainty in SWAG for individual instances very well follows human annotation confusion. Furthermore, we identified cases with a larger mismatch between system predictions and human disagreement where the latter is mainly caused by erroneous or at least questionable decisions. This points to the use of SWAG in an active learning scenario, where annotation noise can be identified using a well calibrated prediction model.

4 Conclusions

Our results show that weight averaging provides consistent and significant improvement for both SNLI and MNLI datasets. The cross-dataset results are slightly mixed but also show the trend of improved cross-domain generalization. Finally, we demonstrate a clear increase in the correlation with human annotation variance when comparing SWAG with non-Bayesian approaches.

For future work we consider making use of multiple annotations also during training and extensions of SWAG such as MultiSWAG (Wilson and Izmailov, 2020). We also plan to test the methods on different NLU datasets, especially those with a high number of annotations (e.g. Nie et al., 2020), and compare the annotation variation and system predictions in more detail.

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Yingbo Gao, Christian Herold, Zijian Yang, and Hermann Ney. 2022. Revisiting checkpoint averaging for neural machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 188–196, Online only. Association for Computational Linguistics.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *Uncertainty in Artificial Intelligence (UAI)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. 2019. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics.

Aarne Talman and Stergios Chatzikyriakidis. 2019. Testing the generalization power of neural network models across NLI benchmarks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew G Wilson and Pavel Izmailov. 2020. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, volume 33, pages 4697–4708. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

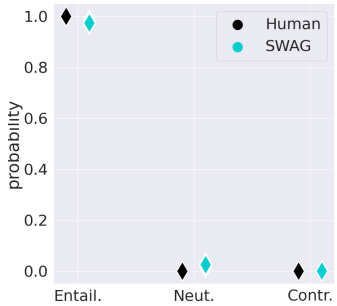
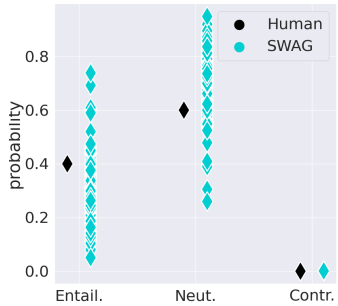
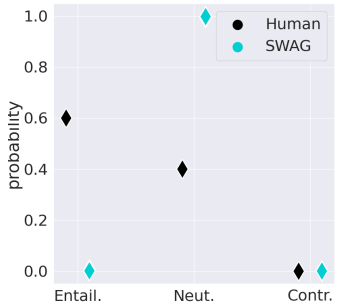
	Correct Prediction		Wrong Prediction
			
Premise	Two female martial artists demonstrate a kick for an audience.	A group of boys playing street soccer.	People shopping at an outside market.
Hypothesis	Two martial artists demonstrate moves for the audience.	A team is playing street soccer.	People are enjoying the sunny day at the market.
Annotations	E-E-E-E-E	E-E-N-N-N	E-E-E-N-N
Cross-Ent	0.02	0.69	3.88

Table 4: Comparison of probability distributions of human annotations vs. SWAG model predictions, for three randomly selected data points from the SNLI dataset. *(Left and middle)* Correctly predicted cases, as indicated by low cross entropy, *(Right)* A wrongly predicted case, as indicated by high cross entropy. SWAG points indicate the outputted probability distributions from $N = 20$ samples.

A Appendix

Here we showcase and discuss three randomly selected data points from the SNLI dataset, and compare the predictions of the $N = 20$ samples from the SWAG model with the annotation distributions for each of these points. Table 4 presents two cases (left and middle) in which the SWAG model makes the correct prediction, and another case (right) in which the model makes the wrong prediction. In the high agreement cases, indicated by lower cross entropies between the annotations and prediction, the SWAG model not only selects the correct label for the instance, but also predicts the annotator disagreement correctly when such a disagreement exists (middle) versus when it does not (left).

The third figure presents a case where the predictions of the SWAG samples are more *certain* than expected: Annotators disagree on whether the hypothesis is Entailment or Neutral, whereas the model predictions place all probability mass to the Neutral class. The corresponding cross entropy is high, which reflects this disagreement. It should be noted that this is also a fairly controversial and difficult data point, and to conclude Entailment requires making some strong assumptions. Ideally, such disagreements between system predictions and annotator distributions may also be used as cues within the training process itself. Two potential venues are (1) using the incongruence between the two distributions as the loss signal to drive the optimization process directly (as opposed to using only the gold label and the predicted class label), and (2) using the incongruence in predictions in an active learning scenario.