How Emotionally Stable is ALBERT? Testing Robustness with Stochastic Weight Averaging on a Sentiment Analysis Task

Urja Khurana[1] Eric Nalisnick[2] Antske Fokkens[1,3]

[1]Vrije Universiteit Amsterdam [2]Universiteit van Amsterdam [3]Eindhoven University of Technology

1. Problem definition

- Current deep language models are fragile.
- **Underspecification** [1]: different predictors can achieve similar results on a specific evaluation set, but exhibit diverging performance on other data due to different induced biases.
- Look at two aspects of robustness: CheckList [2] and Stochastic Weight Averaging (SWA) [3].
- **CheckList**: test for linguistic phenomena captured by model.
- **SWA**: cheap way to ensemble.

2. Our Work

- **Research question: Does SWA provide more stability for a BERT-based model on a sentiment analysis task?**
- Hypothesis: Due to ensembling nature of SWA, expected to bring more robustness and stability.
- Stability $\longrightarrow$ similar model behavior.
- Investigate behavior of models with different random seeds when trained with SWA on CheckList tests.

3. Technical Details

- Train ALBERT-large v2 [4] on SST-2 [5].
- 10 vanilla models and 10 SWA models: each 10^a random seeds.
- Evaluate with 18 CheckList capability tests.
- Quantify agreement of mistakes by **overlap ratio** and **Fleiss' Kappa** [6]

[a]Original experiments conducted with five random seeds.

4. Results

- SWA models: similar or better results on development set of SST-2.
- Random Seed 0 appears to be an outlier.
- CheckList error rates globally reduced with SWA.
- Overlap ratios on the lower side.

	Vanilla	SWA
Random Seed 0	**0.9083**	0.8991
Random Seed 1	0.9507	**0.9541**
Random Seed 2	0.9450	**0.9495**
Random Seed 3	0.9507	**0.9541**
Random Seed 4	0.9450	**0.9461**
Random Seed 5	0.9495	**0.9507**
Random Seed 6	0.9450	**0.9472**
Random Seed 7	**0.9438**	0.9392
Random Seed 8	**0.9461**	0.9450
Random Seed 9	0.9415	**0.9461**

Table: Accuracy on the validation set of SST-2 for the vanilla and SWA models of the different random seeds.

	Vanilla	SWA	Difference
Negation of Positive Sentences	0.029640	0.020448	-0.009192
Negation of Positive, neutral words in the middle	0.107637	0.142219	0.034582
Movie Genre Specific Sentiments	0.581853	0.660138	0.078285
Temporal Sentiment Change	0.248653	0.290926	0.042273
Change Names	-0.091694	-0.084096	0.007598
Negative Names - Positive Instances	0.006975	0.006021	-0.000954
Positive Names - Negative Instances	-0.069162	-0.076226	-0.007064
Negative Names - Negative Instances	-0.082486	-0.069141	0.013346
Positive Names - Positive Instances	0.012704	0.035196	0.022492
Change Movie Industries	-0.072503	-0.052239	0.020264
Change Neutral Words	0.087306	0.135759	0.048453
Add Negative Phrases	-0.031328	-0.062053	-0.030724

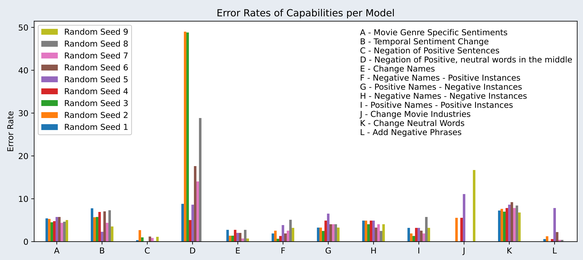
Table: Fleiss' Kappa values of the vanilla and SWA models on the agreement on CheckList mistakes per capability.

- Fleiss' Kappa: increases with first four random seeds on development set. Increase minor with all nine seeds.
- CheckList capabilities: minor increase and decrease with SWA.

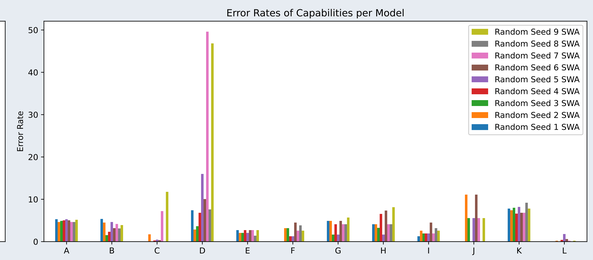
	Vanilla	SWA	Difference
With Random Seed 0	0.205964	0.247299	0.041335
Without Random Seed 0	0.226725	0.360317	0.133592
With Random Seed 0	0.3984	0.4381	0.03967
Without Random Seed 0	0.3881	0.4106	0.0225

Table: Fleiss' Kappa values of the vanilla and SWA models on the agreement on the misclassifications on the development set. The upper block is with the first five random seeds and the lower is with all ten.

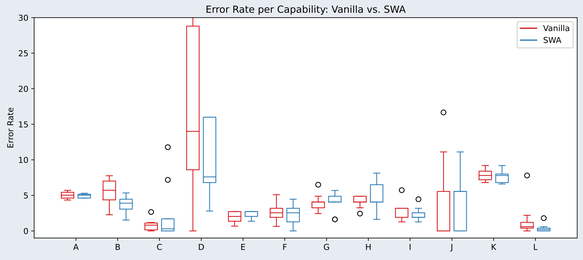
CheckList Error Rates and Overlap Ratios of Mistakes



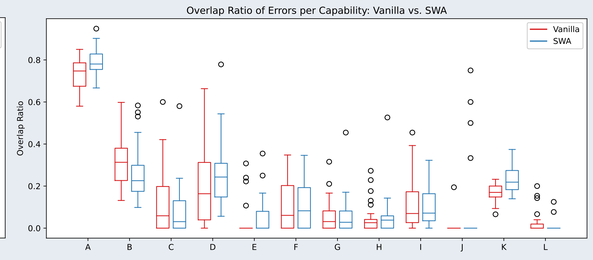
(a) Error rates for vanilla models per capability.



(b) Error rates for SWA models per capability.



(c) Comparison of variation in error rates between vanilla and SWA models per capability.



(d) Comparison of variation in overlap ratios between vanilla and SWA models per capability.

5. Conclusion

- SWA reduces error rates in general but agreement is inconclusive $\longrightarrow$ other tasks and/or models?
- Agreement on lower side in general.
- Easy to compare model behavior with overlap ratios and Fleiss' Kappa scores.

References

[1] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[2] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[3] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. 34th Conference on Uncertainty in Artificial Intelligence 2018, 2018.

[4] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[5] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[6] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Contact Information

- Web: urjakh.github.io
- Email: u.khurana@vu.nl