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] University of Amsterdam
[3] Eindhoven University of Technology

October 12, 2021

Problem Definition

- Current deep language models are fragile.
- Sensitive to small changes in training settings.
- Deployment in the real world can be problematic due to induced biases.

Underspecification

Underspecification (D'Amour et al., 2020): different predictors can achieve similar results on a specific evaluation set, but exhibit diverging performance on other data due to different induced biases.

CheckList methodology (Ribeiro et al., 2020): to test different linguistic phenomena for investigation of robustness of a model.

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** — Labels: negative, positive, neutral			
Template: `I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** — Same pred. (`inv`) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	`inv`	pos → neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	`inv`	neutral → neg	X
...			
Failure rate = 20.8%			

Figure 1: Ribeiro et al. (2020) illustrate generalization issues with language models when adding negations or changing the name of a place.

Stochastic Weight Averaging

Stochastic Weight Averaging (SWA) is a cheap way of ensembling.

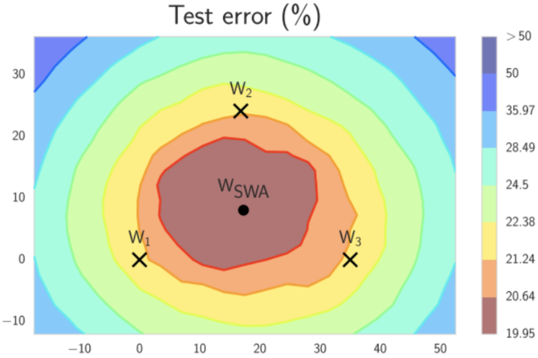

Figure 2: Model Averaging with Stochastic Weight Averaging in Weights Space (Izmailov et al., 2018)

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.
- Test this by training models only differing in random seeds and measuring their agreement on errors.

Our Work

- Model: ALBERT-large version 2 with original hyperparameters (Lan et al., 2020). Both types of models have the same 10[1] different random seeds.
  - 10 Vanilla models
  - 10 SWA models - switches to SWA schedule after 2 epochs
- Data: SST-2 dataset (Socher et al., 2013) and 18 CheckList capability tests.
- Metrics: Accuracy and agreement on errors.
- Measure agreement by calculating overlap ratio and Fleiss' Kappa (Fleiss, 1971).

[1] Original experiments were conducted with five seeds.

Results

Stochastic Weight Averaging

- SWA models achieve similar or better results on the validation set of SST-2.
- *Random Seed 0* appears to be an outlier.

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

CheckList Evaluation

- Error rate goes down for most capabilities with SWA.

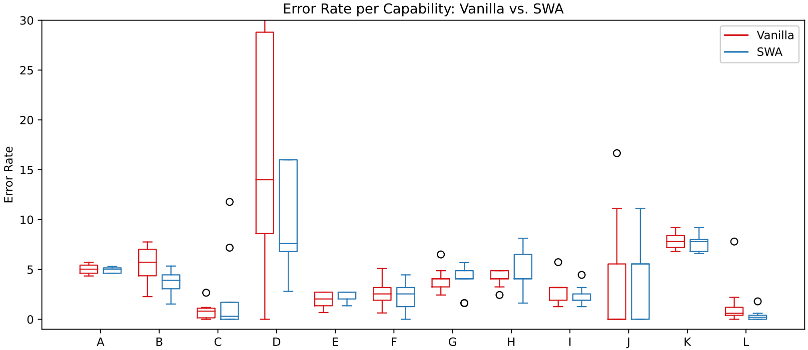

Figure 3: Variation in error rates between vanilla and SWA models per CheckList capability.

- Overlap ratio for most capabilities remains low.
- In comparison to vanilla models, mixed results.

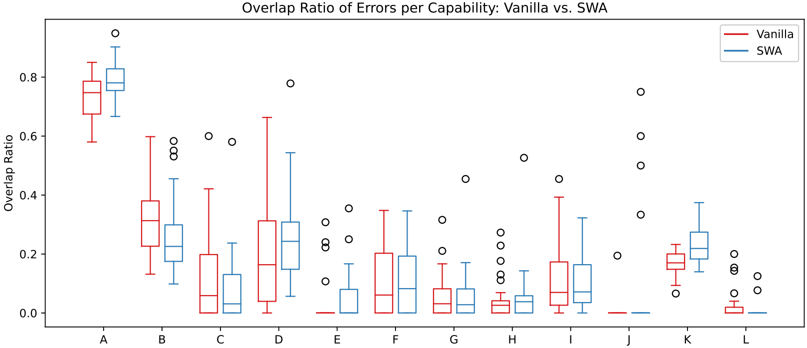

Figure 4: Variation in overlap ratios between vanilla and SWA models per CheckList capability.

Fleiss' Kappa - SST-2 Development Set

	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 2: 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 10.

Fleiss' Kappa - CheckList Tests

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

Conclusion

Summary

Our contributions:

- Explored effects of SWA on stability and robustness of ALBERT-large on sentiment analysis task.
- Combined SWA and CheckList to look at robustness.
- Cheaply quantified agreement between different models: overlap ratio and Fleiss' Kappa scores.

Takeaways

- Current results are inconclusive.
- Outlier random seed and low agreement highlight importance of careful analysis.
- Easy to compare model behavior by looking at overlap ratio and Fleiss' Kappa.
- SWA has potential $\longrightarrow$ explore on other tasks and/or models.

Thank You!

u.khurana@vu.nl

References

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., . . . others (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. (2018). Averaging weights leads to wider optima and better generalization..

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). 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. Retrieved from `https://openreview.net/forum?id=H1eA7AEtvS`

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020, July). Beyond accuracy: Behavioral testing of NLP models with CheckList. , 4902–4912. Retrieved from `https://www.aclweb.org/anthology/2020.acl-main.442` doi: 10.18653/v1/2020.acl-main.442

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642).