Analysis methods in NLP: Adversarial training (and testing)

Christopher Potts

Stanford Linguistics

CS224u: Natural language understanding

Overview

Behavioral evaluations

- Adversarial testing
- **Adversarial training and testing**

SWAG: Situations With Adversarial Generations

SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference

Rowan Zellers[♠] **Yonatan Bisk**[♠] **Roy Schwartz**[♠♡] **Yejin Choi**[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
{rowanz,ybisk,roysch,yejin}@cs.washington.edu
https://rowanzellers.com/swag

HellaSwag: Can a Machine *Really* Finish Your Sentence?

Rowan Zellers[♠] **Ari Holtzman**[♠] **Yonatan Bisk**[♠] **Ali Farhadi**[♠♡] **Yejin Choi**[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
https://rowanzellers.com/hellaswag

SWAG examples

Example

- Context (given): He is throwing darts at a target.
- Sentence start (given): Another man
- Continuation (predicted): throws a dart at the target board.
- Distractors:
  1. comes running in and shoots an arrow at a target.
  2. is shown on the side of men.
  3. throws darts at a disk.

Sources

- ActivityNet: 51,439 exs; 203 activity types
- Large Scale Movie Description Challenge: 62,118 exs

Zellers et al. 2018;
https://rowanzellers.com/swag/

Adversarial filtering for SWAG

For each example *i*:

i The mixture creams the butter. Sugar

a. is added.
b. is sprinkled on top. [Model incorrect; keep this sample]
c. is in many foods.

Repeat for some number of iterations.

Zellers et al. 2018

Model accuracies under adversarial filtering

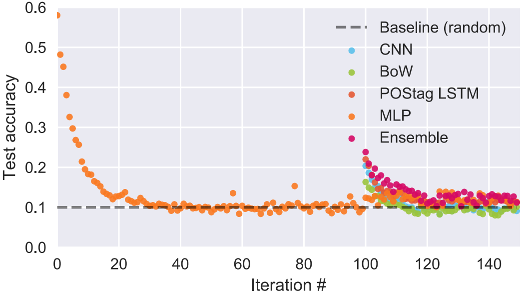

Figure 2: Test accuracy by AF iteration, under the negatives given by $\mathcal{A}$. The accuracy drops from around 60% to close to random chance. For efficiency, the first 100 iterations only use the MLP.

Ensembling begins at iteration 1000
Zellers et al. 2018

SWAG in the original BERT paper

System	Dev	Test
ESIM+GloVe	51.9	52.7
ESIM+ELMo	59.1	59.2
$\text{BERT}_{\text{BASE}}$	81.6	-
$\text{BERT}_{\text{LARGE}}$	**86.6**	**86.3**
Human (expert)†	-	85.0
Human (5 annotations)†	-	88.0

Table 4: SWAG Dev and Test accuracies. Test results were scored against the hidden labels by the SWAG authors. †Human performance is measure with 100 samples, as reported in the SWAG paper.

HellaSWAG

1. ActivityNet retained
2. Large Scale Movie Description Challenge dropped
3. WikiHow data added
4. Adversarial filtering as before, now with more powerful generators and discriminators
5. Human agreement at 94%

Zellers et al. 2019;
https://rowanzellers.com/hellaswag/

HellaSWAG

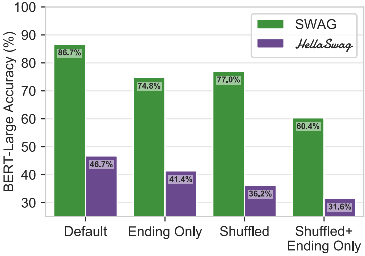

Figure 4: BERT validation accuracy when trained and evaluated under several versions of SWAG, with the new dataset *HellaSwag* as comparison. We compare:

`Ending Only`	No context is provided; just the endings.
`Shuffled`	Endings that are indidivually tokenized, shuffled, and then detokenized.
`Shuffled+ Ending Only`	No context is provided *and* each ending is shuffled.

Zellers et al. 2019; https://rowanzellers.com/hellaswag/

HellaSWAG

Model	Overall		In-Domain		Zero-Shot		ActivityNet		WikiHow	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
Split Size→	10K	10K	5K	5K	5K	5K	3.2K	3.5K	6.8K	6.5K
Chance	25.0									
fastText	30.9	31.6	33.8	32.9	28.0	30.2	27.7	28.4	32.4	33.3
LSTM+GloVe	31.9	31.7	34.3	32.9	29.5	30.4	34.3	33.8	30.7	30.5
LSTM+ELMo	31.7	31.4	33.2	32.8	30.4	30.0	33.8	33.3	30.8	30.4
LSTM+BERT-Base	35.9	36.2	38.7	38.2	33.2	34.1	40.5	40.5	33.7	33.8
ESIM+ELMo	33.6	33.3	35.7	34.2	31.5	32.3	37.7	36.6	31.6	31.5
OpenAI GPT	41.9	41.7	45.3	44.0	38.6	39.3	46.4	43.8	39.8	40.5
BERT-Base	39.5	40.5	42.9	42.8	36.1	38.3	48.9	45.7	34.9	37.7
BERT-Large	**46.7**	**47.3**	**50.2**	**49.7**	**43.3**	**45.0**	**54.7**	**51.7**	**42.9**	**45.0**
Human	95.7	95.6	95.6	95.6	95.8	95.7	94.0	94.0	96.5	96.5

Table 1: Performance of models, evaluated with accuracy (%).We report results on the full validation and test sets (Overall), as well as results on informative subsets of the data: evaluated on in-domain, versus zero-shot situations, along with performance on the underlying data sources (ActivityNet versus WikiHow). All models substantially underperform humans: the gap is over 45% on in-domain categories, and 50% on zero-shot categories.

Zellers et al. 2019;
https://rowanzellers.com/hellaswag/

Adversarial NLI

Adversarial NLI: A New Benchmark for Natural Language Understanding

Yixin Nie*, Adina Williams†, Emily Dinan†, Mohit Bansal*, Jason Weston†, Douwe Kiela†
*UNC Chapel Hill
†Facebook AI Research

Adversarial NLI: Dataset creation

A direct response to adversarial test failings *NLI datasets:

Adversarial NLI: Dataset creation

A direct response to adversarial test failings *NLI datasets:

1. The annotator is presented with a premise sentence and a condition (entailment, contradiction, neutral).

Adversarial NLI: Dataset creation

A direct response to adversarial test failings *NLI datasets:

1. The annotator is presented with a premise sentence and a condition (entailment, contradiction, neutral).
2. The annotator writes a hypothesis.

Adversarial NLI: Dataset creation

A direct response to adversarial test failings *NLI datasets:

1. The annotator is presented with a premise sentence and a condition (entailment, contradiction, neutral).
2. The annotator writes a hypothesis.
3. A state-of-the-art model makes a prediction about the premise–hypothesis pair.

Adversarial NLI: Dataset creation

A direct response to adversarial test failings *NLI datasets:

1. The annotator is presented with a premise sentence and a condition (entailment, contradiction, neutral).
2. The annotator writes a hypothesis.
3. A state-of-the-art model makes a prediction about the premise–hypothesis pair.
4. If the model's prediction matches the condition, the annotator returns to step 2 to try again.

Adversarial NLI: Dataset creation

A direct response to adversarial test failings *NLI datasets:

1. The annotator is presented with a premise sentence and a condition (entailment, contradiction, neutral).
2. The annotator writes a hypothesis.
3. A state-of-the-art model makes a prediction about the premise–hypothesis pair.
4. If the model's prediction matches the condition, the annotator returns to step 2 to try again.
5. If the model was fooled, the premise–hypothesis pair is independently validated by other annotators.

Adversarial NLI: Example

Premise	Hypothesis	Reason	Label	Model
A melee weapon is any weapon used in direct hand-to-hand combat; by contrast with ranged weapons which act at a distance. The term “melee” originates in the 1640s from the French word “mělée”, which refers to hand-to-hand combat, a close quarters battle, a brawl, a confused fight, etc. Melee weapons can be broadly divided into three categories	Melee weapons are good for ranged and hand-to-hand combat.	Melee weapons are good for hand to hand combat, but NOT ranged.	E	N

Adversarial NLI results

Model	Data	A1	A2	A3	ANLI	ANLI-E	SNLI	MNLI-m/-mm
BERT	S,M$^{\star 1}$	<u>00.0</u>	28.9	28.8	19.8	19.9	91.3	86.7 / 86.4
	+A1	44.2	32.6	29.3	35.0	34.2	91.3	86.3 / 86.5
	+A1+A2	57.3	45.2	33.4	44.6	43.2	90.9	86.3 / 86.3
	+A1+A2+A3	57.2	49.0	46.1	50.5	46.3	90.9	85.6 / 85.4
	S,M,F,ANLI	57.4	48.3	43.5	49.3	44.2	90.4	86.0 / 85.8
XLNet	S,M,F,ANLI	67.6	50.7	48.3	55.1	52.0	91.8	89.6 / 89.4
RoBERTa	S,M	47.6	25.4	22.1	31.1	31.4	92.6	90.8 / 90.6
	+F	54.0	24.2	22.4	32.8	33.7	92.7	90.6 / 90.5
	+F+A1$^{\star 2}$	68.7	<u>19.3</u>	22.0	35.8	36.8	92.8	90.9 / 90.7
	+F+A1+A2$^{\star 3}$	71.2	44.3	<u>20.4</u>	43.7	41.4	92.9	91.0 / 90.7
	S,M,F,ANLI	73.8	48.9	44.4	53.7	49.7	92.6	91.0 / 90.6

Table 3: Model Performance. 'Data' refers to training dataset ('S' refers to SNLI, 'M' to MNLI dev (-m=matched, -mm=mismatched), and 'F' to FEVER); 'A1–A3' refer to the rounds respectively. '-E' refers to test set examples written by annotators exclusive to the test set. Datasets marked '$\star n$' were used to train the base model for round n, and their performance on that round is <u>underlined</u>.

A vision for future development

Zellers et al. (2019)

"a path for NLP progress going forward: towards benchmarks that adversarially co-evolve with evolving state-of-the-art models."

Nie et al. (2019)

"This process yields a "moving post" dynamic target for NLU systems, rather than a static benchmark that will eventually saturate."

Dynabench

Dyna Bench

Rethinking AI Benchmarking

Dynabench is a research platform for dynamic data collection and benchmarking. Static benchmarks have well-known issues: they saturate quickly, are susceptible to overfitting, contain exploitable annotator artifacts and have unclear or imperfect evaluation metrics.

This platform in essence is a scientific experiment: can we make faster progress if we collect data dynamically, with humans and models in the loop, rather than in the old-fashioned static way?

Read more

Dynabench

1. NLI (see Nie et al. 2020)
2. QA (see Bartolo et al. 2020)
3. Sentiment (DynaSent; Potts et al. 2020)
4. Hate Speech (Vidgen et al. 2020)

Can adversarial training improve systems?

1. Jia and Liang (2017:§4.6): Training on adversarial examples makes them more robust to those examples but not to simple variants.
2. Alzantot et al. (2018:§4.3): “We found that adversarial training provided no additional robustness benefit in our experiments using the test set, despite the fact that the model achieves near 100% accuracy classifying adversarial examples included in the training set.”
3. Liu et al. (2019): Fine-tuning with a few adversarial examples improves systems in some cases (as discussed under ‘inoculation’ just above).
4. Iyyer et al. (2018): Adversarially generated paraphrases improve model robustness to syntactic variation.

References I

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019. Inoculation by fine-tuning: A method for analyzing challenge datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. UNC CHapel Hill and Facebook AI Research.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. DynaSent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv prerint arXiv:2012.15761*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.