国立国語研究所学術情報リポジトリ

Adversarial Training for Commonsense Inference

Adversarial Training for Commonsense Inference

Lis Pereira[1], Xiaodong Liu[2], Fei Cheng[3], Masayuki Asahara[4], Ichiro Kobayashi[1]
[1] Ochanomizu University [2] Microsoft Research [3] Kyoto University
[4] The National Institute for Japanese Language and Linguistics (NINJAL)
kanashiro.pereira@ocha.ac.jp, xiaodl@microsoft.com, feicheng@i.kyoto-u.ac.jp
masayu-a@ninjal.ac.jp, koba@is.ocha.ac.jp

Abstract

We propose an **A**dversaria**L** training algorithm for commonsense **I**nferen**CE** (ALICE). We apply small perturbations to word embeddings and minimize the resultant adversarial risk to regularize the model. We exploit a novel combination of two different approaches to estimate these perturbations: 1) using the true label and 2) using the model prediction. Without relying on any human-crafted features, knowledge bases or additional datasets other than the target datasets, our model boosts the fine-tuning performance of RoBERTa, achieving competitive results on multiple reading comprehension datasets that require commonsense inference.

1 Introduction

Commonsense knowledge is often necessary for natural language understanding. As shown in Table 1, we can understand that the writer needs help to get dressed and seems upset with this situation, indicating that he or she is probably not a child. Thus, we can infer that a possible reason that the writer needs to be dressed by other people is that he or she may have a physical disability (Huang et al., 2019). Although a simple task for humans, it is still challenging for computers to understand and reason about commonsense.

Commonsense inference in natural language processing (NLP) is generally evaluated via machine reading comprehension task, in the format of selecting plausible responses with respect to natural language queries. Recent approaches are based on the use of pre-trained Transformer-based language models such as BERT (Devlin et al., 2019). Some approaches rely solely on these models by adopting either a single or multi-stage fine-tuning approach (by fine-tuning using additional datasets in a stepwise manner) (Li and Xie, 2019; Sharma and Roychowdhury, 2019; Liu and Yu, 2019; Huang et al.,

Paragraph: It's a very humbling experience when you need someone to dress you every morning, tie your shoes, and put your hair up. Every menial task takes an unprecedented amount of effort. It made me appreciate Dan even more. But anyway I shan't dwell on this (I'm not dying after all) and not let it detract from my lovely 5 days with my friends visiting from Jersey.

Question: What's a possible reason the writer needed someone to dress him every morning?

Option1: The writer doesn't like putting effort into these tasks.
Option2: **The writer has a physical disability.**
Option3: The writer is bad at doing his own hair.
Option4: None of the above choices.

Table 1: Example from the CosmosQA dataset (Huang et al., 2019). The task is to identify the correct answer option. The correct answer is in **bold.**

2019; Zhou et al., 2019), while others further enhance their word representations with knowledge bases such as ConceptNet (Jain and Singh, 2019; Da, 2019; Wang et al., 2020). However, due to the often limited data from the downstream tasks and the extremely high complexity of the pre-trained model, aggressive fine-tuning can easily make the adapted model overfit the data of the target task, making it unable to generalize well on unseen data (Jiang et al., 2019). Moreover, some researchers have shown that such pre-trained models are vulnerable to adversarial attacks (Jin et al., 2020).

Inspired by the recent success of adversarial training in NLP (Zhu et al., 2020; Jiang et al., 2019), our **A**dversaria**L** training algorithm for commonsense **I**nferen**CE** (ALICE) focuses on improving the generalization of pre-trained language mod-

els on downstream tasks by enhancing their robustness in the embedding space. More specifically, during the fine-tuning stage of Transformer-based models, e.g. RoBERTa (Liu et al., 2019b), random perturbations are added to the embedding layer to regularize the model by updating the parameters on these adversarial embeddings. ALICE exploits a novel way of combining two different approaches to estimate these perturbations: 1) using the true label and 2) using the model prediction. Experiments show that we were able to boost the performance of RoBERTa on multiple reading comprehension datasets that require commonsense inference, achieving competitive results with state-of-the-art approaches.

2 ALICE

Given a dataset D of N training examples, $D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$, the objective of supervised learning is to learn a function $f(x; \theta)$ that minimizes the empirical risk, which is defined by $\min_\theta E_{(x,y) \sim D}[l(f(x; \theta), y)]$. Here, the function $f(x; \theta)$ maps input sentences x to an output space y, and θ are learnable parameters. While this objective is effective to train a neural network, it usually suffers from overfitting and poor generalization to unseen cases (Goodfellow et al., 2015; Madry et al., 2018). To alleviate these issues, one can use adversarial training, which has been primarily explored in computer vision (Goodfellow et al., 2015; Madry et al., 2018). The idea is to perturb the data distribution in the embedding space by performing adversarial attacks. Specifically, its objective is defined by:

$$\min_\theta E_{(x,y) \sim D}[\max_\delta l(f(x + \delta; \theta), y)], \quad (1)$$

where δ is the perturbation added to the embeddings. One challenge of adversarial training is how to estimate this perturbation δ, which is to solve the inner maximization, $\max_\delta l(f(x + \delta; \theta), y)$. A feasible solution is to approximate it by a fixed number of steps of a gradient-based optimization approach (Madry et al., 2018).

Based on recent successful cases that applied adversarial training to NLP (Jiang et al., 2019; Miyato et al., 2018), the approaches to estimate δ can be divided into two categories: adversarial training that uses the label y (Zhu et al., 2020) and adversarial training that uses the model prediction $f(x; \theta)$, i.e. a "virtual" label (Miyato et al., 2018; Jiang et al., 2019). We hypothesize that these two categories

complement each other: the first one is to improve the robustness of our target label, by avoiding an increase in the error of the unperturbed inputs, while the second term enforces the smoothness of the model, encouraging the output of the model not to change much, when injecting a small perturbation to the input. Thus, ALICE proposes a novel algorithm by combining these two approaches, which is defined by:

$$\min_\theta E_{(x,y) \sim D}[\max_{\delta_1} l(f(x + \delta_1; \theta), y) + \\ \alpha \max_{\delta_2} l(f(x + \delta_2; \theta), f(x; \theta))], \quad (2)$$

where δ_1 and δ_2 are two perturbations, bounded by a general l_p norm ball, estimated by a fixed K steps of the gradient-based optimization approach. In our experiments, we set $p = \infty$. It has been shown that a larger K can lead to a better estimation of δ (Qin et al., 2019; Madry et al., 2018). However, this can be expensive, especially in large models, e.g. BERT and RoBERTa. Thus, K is set to 1 for a better trade-off between speed and performance. Note that α is a hyperparameter balancing these two loss terms. In our experiments, we set α to 1.

3 Experiments

3.1 Datasets and Evaluation Metrics
We evaluate ALICE on three reading comprehension benchmarks that require commonsense inference:

CosmosQA (Huang et al., 2019): a large-scale dataset that focuses on people's everyday narratives, asking questions about the likely causes or effects of events that require reasoning beyond the exact text spans in the context. It has 35,888 questions on 21,886 distinct contexts taken from blogs of personal narratives. Each question has four answer candidates, one of which is correct. 93.8% of the dataset requires contextual commonsense reasoning.

MCScript2.0 (Ostermann et al., 2019b): a dataset focused on short narrations on different everyday activities (e.g. baking a cake, taking a bus, etc.). It has 19,821 questions on 3,487 texts. Each question has two answer candidates, one of which is correct. Roughly half of the questions require inferences over commonsense knowledge.

MC-TACO (Zhou et al., 2019): a dataset that entirely focuses on a specific reasoning capablity: temporal commonsense. It considers five temporal properties, (1) duration (how long an event takes), (2) temporal ordering (typical order of events), (3)

Dataset	#Train	#Dev	#Test	#Label	Task	Metrics
CosmosQA	25,262	2,985	6,963	4	Relevance Ranking	Accuracy
MCScript2.0	14,191	2,020	3,610	2	Relevance Ranking	Accuracy
MCTACO	-	3,783	9,442	2	Pairwise Text Classification	Exact Match (EM)/F1

Table 2: Summary of the three datasets: CosmosQA, MCScript2.0 and MCTACO.

typical time (when an event occurs), (4) frequency (how often an event occurs), and (5) stationarity (whether a state is maintained for a very long time or indefinitely). It contains 13k tuples, each consisting of a sentence, a question, and a candidate answer, that should be judged as plausible or not. The sentences are taken from different sources such as news, Wikipedia and textbooks.

The summary of the datasets is in Table 2. For the MCTACO dataset, no training set is available. Following (Zhou et al., 2019), we use the dev set for fine-tuning the model. We perform 5-fold cross-validation for fine-tuning the parameters.

We evaluate CosmosQA and MCScript2.0 in terms of accuracy. Following (Ostermann et al., 2019a), we also report for the MCScript2.0 accuracy on the commonsense based questions and accuracy on the questions that are not commonsense based. For the MCTACO, we report the exact match (EM) and F1 scores, following (Zhou et al., 2019). EM measures how many questions a system correctly labeled all candidate answers, while F1 measures the average overlap between one's predictions and the ground truth. Our implementation for pairwise text classification and relevance ranking tasks are based on the MT-DNN framework[1] (Liu et al., 2019a, 2020).

3.2 Implementation Details

The RoBERTa$_{\text{LARGE}}$ model (Liu et al., 2019b) was used as the text encoder. We used ADAM (Kingma and Ba, 2015) as our optimizer with a learning rate in the range $\in \{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}, 5 \times 10^{-5}\}$ and a batch size $\in \{16, 32, 64\}$. The maximum number of epochs was set to 10. A linear learning rate decay schedule with warm-up over 0.1 was used, unless stated otherwise. We also set the dropout rate of all the task specific layers as 0.1, except 0.3 for MCTACO. To avoid gradient exploding, we clipped the gradient norm within 1. All the texts were tokenized using wordpieces and were chopped to spans no longer than 512 tokens.

3.3 Baselines

We compare ALICE to a list of state-of-the-art models, as shown in Table 3. **BERT + unit normalization** (Zhou et al., 2019) is the BERT base model. The authors further add unit normalization to temporal expressions in candidate answers and fine-tune on the MC-TACO dataset. **RoBERTa$_{\text{LARGE}}$** is our re-implementation of the large RoBERTa model by (Liu et al., 2019b). **PSH-SJTU** (Li and Xie, 2019) is based on multi-stage fine-tuning XL-NET (Yang et al., 2019) on RACE (Lai et al., 2017), SWAG (Zellers et al., 2018) and MC-Script2.0 datasets. **K-ADAPTER** (Wang et al., 2020) further enhances RoBERTa word representations with multiple knowledge sources, such as factual knowledge obtained through Wikipedia and Wikidata and linguistic knowledge obtained through dependency parsing web texts. **SMART** (Jiang et al., 2019) is an adversarial training model for fine-tuning pre-trained language models through regularization. SMART uses the model prediction, $f(x; \theta)$, for estimating the perturbation δ. This model recently obtained state-of-the-art results on a bunch of NLP tasks on the GLUE benchmark (Wang et al., 2018). We also compare ALICE with a baseline that uses only the label y for estimating the perturbation δ (called model **ADV** hereafter) (Madry et al., 2018).

3.4 Results

The results are summarized in Table 3. Overall, we observed that adversarial methods, i.e. ADV, SMART and ALICE, were able to achieve competitive results over the baselines, without using any additional knowledge source, and without using any additional dataset other than the target task datasets. These results suggest that adversarial training lead to a more robust model and help generalize better on unseen data.

ALICE consistently outperformed SMART (which overall outperformed ADV) across all three datasets on both dev and test sets, indicating that adversarial training that uses the label y and adversarial training that uses the model prediction

---

[1]https://github.com/namisan/mt-dnn

57

	CosmosQA	MCScript2.0			MCTACO	
Model	Acc	Acc	Acc_{cs}	Acc_{ood}	EM	F1
Development Set Results						
RoBERTa$_{LARGE}$ (Liu et al., 2019b)	80.60	90.0	87.7	92.1	44.12	64.85
ADV	81.14	92.1	89.8	94.3	52.70	78.12
SMART (Jiang et al., 2019)	82.00	93.6	**91.8**	95.2	53.79	78.31
ALICE	**83.60**	**93.8**	91.7	**95.7**	**58.02**	**78.64**
Test Set Results						
Human Performance	94.00	97.0	-	-	75.80	87.10
BERT + unit normalization (Zhou et al., 2019)	-	-	-	-	42.70	69.90
T5-3B fine-tuned + number normalization[*]	-	-	-	-	**59.08**	79.46
PSH-SJTU (Li and Xie, 2019)	-	90.6	90.3	91.5	-	-
K-ADAPTER (Wang et al., 2020)	81.83	-	-	-	-	-
GB-KSI(v2)[*]	83.97	-	-	-	-	-
RoBERTa$_{LARGE}$ (Liu et al., 2019b)	-	88.8	87.0	90.7	51.05	76.85
ADV	-	91.1	90.3	92.0	54.27	77.23
SMART (Jiang et al., 2019)	81.90	91.8	90.5	93.4	54.80	78.03
ALICE	**84.57**	**92.5**	**91.6**	**93.5**	56.45	**79.50**

Table 3: Development and test results of CosmosQA, MCScript 2.0 and MCTACO. The best results are in **bold**. Note that RoBERTa$_{LARGE}$, SMART and ALICE models use RoBERTa$_{LARGE}$ as the text encoder, and for a fair comparison, all these results are produced by ourselves. On the test results, note that CosmosQA and MCTACO are scored by using the official evaluation server (https://leaderboard.allenai.org/). * denotes unpublished work and scores were obtained from the evaluation server on April 16, 2020. Acc_{cs} denotes the accuracy on commonsense based questions and Acc_{ood} denotes the accuracy on questions that are not commonsense based, i.e. out-of-domain questions.

$f(x; \theta)$ are complementary, leading to better results. For example, on the CosmosQA dataset, we obtained a dev-set accuracy of 83.6% with ALICE, a 1.6% and 3.0% absolute gains over SMART and RoBERTa$_{LARGE}$, respectively. On the blind test-set, ALICE outperforms by a large margin K-ADAPTER, a model that enhances RoBERTa word representations with multiple knowledge sources. Our submission to the CosmosQA leaderboard achieved a test-set accuracy of 84.57%, ranking first place among all submissions (as of April 16, 2020). On the MCScript2.0 dataset, ALICE obtained a dev-set accuracy of 93.8% in total, a 0.2% and 3.8% absolute gains over SMART and RoBERTa$_{LARGE}$, respectively. On the commonsense based questions, ALICE underperformed SMART by 0.1% and outperformed RoBERTa$_{LARGE}$ by 4.0%. On the out-of-domain questions, ALICE obtained 0.5% and 3.6% absolute gains over SMART and RoBERTa$_{LARGE}$, respectively. On the MCScript2.0 test-set, ALICE outperformed all baselines (on all types of questions), including SMART, indicating that it can generalize better to unseen cases. Moreover, ALICE outperformed PSH-SJTU, which is based on

XLNET, and used additional datasets other than MCScript2.0, while ALICE does not use any additional dataset. On the MCTACO dataset, ALICE obtained on the dev-set 56.20% EM score, a 2.41% and 12.8% absolute gains over SMART and RoBERTa$_{LARGE}$, respectively, and 79.06% F1 score, a 0.75% and 14.21% absolute gains over SMART and RoBERTa$_{LARGE}$, respectively. On the test-set, ALICE outperformed SMART obtaining absolute gains of 1.65% and 1.47% on EM and F1 scores, respectively. Compared to the T5-3B fine-tuned + number normalization model, which uses T5, a much larger model (with 3B parameters) than RoBERTa (300M parameters), ALICE obtained competitive results, outperforming by 0.04 on F1 score and obtaining 2.63% lower score on EM. Regarding the training time, ALICE takes on average 4X more time to train compared to standard fine-tuning.

4 Conclusion

We proposed ALICE, a simple and efficient adversarial training algorithm for fine-tuning large scale pre-trained language models. Our experiments demonstrated that it achieves competitive

results on multiple machine reading comprehension datasets, without relying on any additional resource other than the target task dataset. Although in this paper we focused on the machine reading comprehension task, ALICE can be generalized to solve other downstream tasks as well, and we will explore this direction as future work.

Acknowledgments

References

Jeff Da. 2019. Jeff da at coin-shared task. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 85–92.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. *ICLR 2015*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2391–2401.

Yash Jain and Chinmay Singh. 2019. Karna at coin shared task 1: Bidirectional encoder representations from transformers with relational knowledge for machine comprehension with common sense. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 75–79.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? natural language attack on text classification and entailment. *AAAI 2020*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR (Poster) 2015*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

Zhexi Zhang Wei Zhu Zheng Li Yuan Ni Peng Gao Junchi Yan Li, Xiepeng and Guotong Xie. 2019. Pingan smart health and sjtu at coin-shared task: utilizing pre-trained language models and common-sense knowledge in machine reading tasks. In *In Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 93–98, Hong Kong.

Chunhua Liu and Dong Yu. 2019. Blcu-nlp at coin-shared task1: Stagewise fine-tuning bert for commonsense inference in everyday narrations. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 99–103.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao. 2020. The microsoft toolkit of multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:2002.07972*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. *ICLR 2018*.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019a. Mcscript2. 0: A machine comprehension corpus focused on script events and participants. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117.

Simon Ostermann, Sheng Zhang, Michael Roth, and Peter Clark. 2019b. Commonsense inference in natural language processing (coin)-shared task report. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 66–74.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Alhussein Fawzi, Soham De, Robert Stanforth, Pushmeet Kohli, et al. 2019. Adversarial robustness through local linearization. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*.

Prakhar Sharma and Sumegh Roychowdhury. 2019. Iitkgp at coin 2019: Using pre-trained language models for modeling machine comprehension. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 80–84.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

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

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than" going for a walk": A study of temporal commonsense understanding. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for language understanding. *ICLR 2020*.