Targeted Adversarial Training for Natural Language Understanding

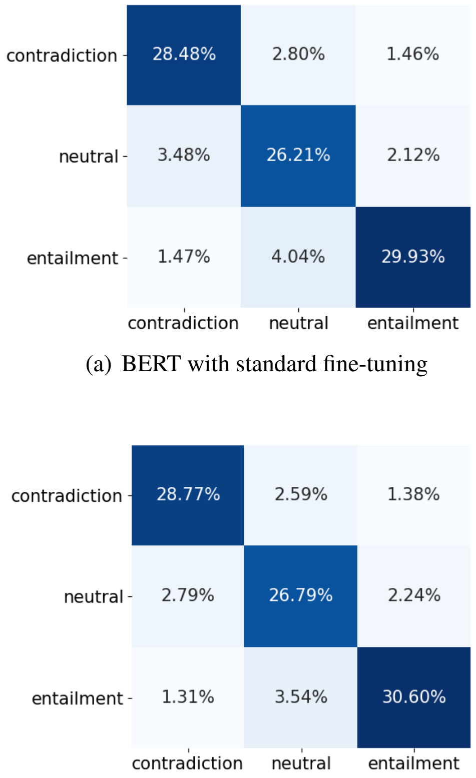
Lis Pereira[1];[*] Xiaodong Liu[2*], Hao Cheng[2],
Hoifung Poon[2], Jianfeng Gao[2], Ichiro Kobayashi[1]
[1] Ochanomizu University [2] Microsoft Research
`{kanashiro.pereira,kobayashi.ichiro}@ocha.ac.jp`
`{xiaodl,chehao,hoifung,jfgao}@microsoft.com`

Abstract

We present a simple yet effective **T**argeted **A**dversarial **T**raining (**TAT**) algorithm to improve adversarial training for natural language understanding. The key idea is to introspect current mistakes and prioritize adversarial training steps to where the model errs the most. Experiments show that TAT can significantly improve accuracy over standard adversarial training on GLUE and attain new state-of-the-art zero-shot results on XNLI. Our code will be released at: `https://github.com/namisan/mt-dnn`.

1 Introduction

Adversarial training has proven effective in improving model generalization and robustness in computer vision (Madry et al., 2017; Goodfellow et al., 2014) and natural language processing (NLP) (Zhu et al., 2019; Jiang et al., 2019; Cheng et al., 2019; Liu et al., 2020a; Pereira et al., 2020; Cheng et al., 2020). It works by augmenting the input with a small perturbation to steer the current model prediction away from the correct label, thus forcing subsequent training to make the model more robust and generalizable. Aside from some prior work in computer vision (Dong et al., 2018; Tramèr et al., 2017), most adversarial training approaches adopt *non-targeted* attacks, where the model prediction is not driven towards a specific incorrect label. In NLP, the cutting-edge research in adversarial training tends to focus on making adversarial training less expensive (e.g., by reusing backward steps in FreeLB (Zhu et al., 2019)) or regularizing rather than replacing the standard training objective (e.g., in virtual adversarial training (VAT) (Jiang et al., 2019)).



(a) BERT with standard fine-tuning

(b) BERT with TAT fine-tuning

Figure 1: Comparison of confusion matrices on MNLI development set (in-domain). X-axis and Y-axis represent the predicted and gold labels, respectively. TAT produces an accuracy gain of 1.7 absolute points.

By contrast, in this paper, we investigate an orthogonal direction by augmenting adversarial training with introspection capability and adopting *targeted* attacks to focus on where the model errs the most. We observe that in many NLP applications, the error patterns are non-uniform. For example, in the MNLI development set (in-domain), standard fine-tuned BERT model tends to misclassify a non-neutral instance as "neutral" more often than the opposite label (Figure 1 top). We thus propose *Targeted Adversarial Training* (TAT), a simple yet effective algorithm for adversarial training. For each instance, instead of taking adversarial steps *away* from the gold label, TAT samples an incorrect label proportional to how often the current

[*] Equal contribution.

Proceedings of the 2021 Conference of the North American Chapter of the
Association for Computational Linguistics: Human Language Technologies, pages 5385–5393
June 6–11, 2021. ©2021 Association for Computational Linguistics

model makes the same error in general, and takes adversarial steps *towards* the chosen incorrect label. To our knowledge, this is the first attempt to apply targeted adversarial training to NLP tasks. In our experiments, this leads to significant improvement over standard non-adversarial and adversarial training alike. For example, in the MNLI development set, TAT produced an accuracy gain of 1.7 absolute points (Figure 1 bottom). On the overall GLUE benchmark, TAT outperforms state-of-the-art non-targeted adversarial training methods such as FreeLB and VAT, and enables the BERT$_{\text{BASE}}$ model to perform comparably to the BERT$_{\text{LARGE}}$ model with standard training. The benefit of TAT is particularly pronounced in out-domain settings, such as in zero-shot learning in natural language inference, attaining new state-of-the-art cross-lingual results on XNLI.

2 Targeted Adversarial Training (TAT)

In this paper, we focus on fine-tuning BERT models (Devlin et al., 2018) in our investigation of targeted adversarial training, as this approach has proven very effective for a wide range of NLP tasks.

The training algorithm seeks to learn a function $f(x;\theta) : x \to C$ as parametrized by θ, where C is the class label set. Given a training dataset D of input-output pairs (x, y) and the loss function $l(.,.)$ (e.g., cross entropy), the standard training objective would minimize the empirical risk:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[l(f(x;\theta), y)].$$

By contrast, in adversarial training, as pioneered in computer vision (Goodfellow et al., 2014; Hsieh et al., 2019; Madry et al., 2017; Jin et al., 2019), the input would be augmented with a small perturbation that maximize the adversarial loss:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}[\max_{\delta} l(f(x+\delta;\theta), y)],$$

where the inner maximization can be solved by projected gradient descent (Madry et al., 2017).

Recently, adversarial training has been successfully applied to NLP as well (Zhu et al., 2019; Jiang et al., 2019; Pereira et al., 2020). In particular, FreeLB (Zhu et al., 2019) leverages the *free adversarial training* idea (Shafahi et al., 2019) by reusing the backward pass in gradient computation to carry out inner ascent and outer descent steps simultaneously. SMART (Jiang et al., 2019) instead regularizes the standard training objective using *virtual adversarial training* (Miyato et al., 2018):

Algorithm 1 TAT

```
Input: $T$: the total number of iterations, $\mathcal{X} = \{(x_1, y_1), ..., (x_n, y_n)\}$: the dataset, $f(x;\theta)$: the machine learning model parametrized by $\theta$, $\sigma^2$: the variance of the random initialization of perturbation $\delta$, $\epsilon$: perturbation bound, $K$: the number of iterations for perturbation estimation, $\eta$: the step size for updating perturbation, $\tau$: the global learning rate, $\alpha$: the smoothing proportion of adversarial training in the augmented learning objective, $\Pi$: the projection operation and $C$: the classes.
for $t = 1, .., T$ do
    for $(x, y) \in \mathcal{X}$ do
        $\delta \sim \mathcal{N}(0, \sigma^2 I)$
        $y_t = sample(C_{\setminus y})$
        for $m = 1, .., K$ do
            $g_{adv} \leftarrow \nabla_{\delta} l(f(x+\delta;\theta), y_t)$
            $\delta \leftarrow \Pi_{\|\delta\|_\infty \le \epsilon}(\delta - \eta g_{adv})$
        end for
        $g_\theta \leftarrow \nabla_\theta l(f(x;\theta), y)$
            $+\alpha \nabla_\theta l(f(x;\theta), f(x+\delta;\theta))$
        $\theta \leftarrow \theta - \tau g_\theta$
    end for
end for
Output: $\theta$
```

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

Effectively, the adversarial term encourages smoothness in the input neighborhood, and α is a hyperparameter that controls the trade-off between standard errors and adversarial errors.

In standard adversarial training, the algorithm simply tries to perturb the input x away from the gold label y given the current parameters θ. It is agnostic to which incorrect label $f(x)$ might be steered towards. By contrast, in Targeted Adversarial Training (TAT), we would explicitly pick a target $y_t \neq y$ and try to steer the model towards y_t. Intuitively, we would like to focus training on where the model currently errs the most. We accomplish this by keeping a running tally of $e(y, y_t)$, which is the current expected error of predicting y_t when the gold label is y, and sample y_t from $C_{\setminus y} = C - \{y\}$ in proportion to $e(y, y_t)$. See Algorithm 1 for details. TAT can be applied to the

Methods	MNLI-m/mm	QQP	RTE	QNLI	MRPC	CoLA	SST	STS-B	Average
	Acc	Acc/F1	Acc	Acc	Acc/F1	Mcc	Acc	P/S Corr	Score
Standard (BERT$_{\text{LARGE}}$)$^{\text{dev}}$	86.3/86.2	91.3/88.4	71.1	92.4	85.8/89.5	61.8	93.5	89.6/89.3	84.0
Standard (BERT$_{\text{LARGE}}$)$^{\text{test}}$	86.7/85.9	72.1/89.3	70.1	92.7	85.4/89.3	60.5	94.9	87.6/86.5	82.4
Standard$^{\text{dev}}$	84.5/84.4	90.9/88.3	63.5	91.1	84.1/89.0	54.7	92.9	89.2/88.8	81.5
FreeLB$^{\text{dev}}$	85.4/85.5	91.4/88.4	70.4	91.5	86.2/90.3	**59.1**	93.2	89.7/89.1	83.5
VAT$^{\text{dev}}$	85.5/85.7	91.5/88.5	71.2	91.7	87.7/91.3	58.2	93.3	90.0/89.4	83.7
TAT$^{\text{dev}}$	**86.2/85.9**	**91.8/89.1**	**72.6**	**92.2**	**88.2/91.5**	58.5	**93.6**	**90.8/89.6**	**84.2**
Standard$^{\text{test}}$	84.6/83.4	71.2/89.2	66.4	90.5	84.8/88.9	52.1	93.5	87.1/85.8	80.0
TAT$^{\text{test}}$	**85.8/84.8**	**72.8/89.6**	**69.7**	**92.4**	**88.2/91.1**	**59.8**	**94.5**	**89.7/89.0**	**82.8**

Table 1: Comparison of standard and adversarial training methods on GLUE. All rows except the top two use standard BERT$_{\text{BASE}}$ model. The GLUE test results are scored using the GLUE evaluation server. Note that the test results of Standard including BERT$_{\text{BASE}}$ and BERT$_{\text{LARGE}}$ are taken from https://gluebenchmark.com/leaderboard.

original adversarial training or virtual adversarial training alike. In this paper, we focus on adapting virtual adversarial training (VAT) (Jiang et al., 2019). The two lines in blue color are the only change from VAT. We initialize $e(y, y_t)$ with uniform distribution and update them in each epoch. We conducted an oracle experiment where $e(y, y_t)$ was taken from the confusion matrix from standard training and found that it performed similarly as our online version.

It is more challenging to apply TAT to regression tasks, as we would need to keep track of a continuous error distribution. To address this problem, we quantize the value range into ten bins and apply TAT similarly as in the classification setting (once a bin is chosen, a value is sampled uniformly within).

3 Experiments

We compare targeted adversarial training (TAT) with standard training and state-of-the-art adversarial training methods such as FreeLB (Zhu et al., 2019) and VAT (Miyato et al., 2018; Jiang et al., 2019). We use the standard uncased BERT$_{\text{BASE}}$ model (Devlin et al., 2018), unless noted otherwise. Due to the additional overhead incurred during training, adversarial methods are somewhat slower than standard training. Like VAT, TAT requires an additional K adversarial steps compared to standard training. In practice, $K = 1$ suffices for TAT and VAT, so they are just slightly slower (roughly 2 times compared to standard training). FreeLB, by contrast, typically requires 2-5 steps to attain good performance, so is significantly slower.

3.1 Implementation Details

Our implementation is based on the MT-DNN toolkit (Liu et al., 2020b). We follow the default hyperparameters used for fine-tuning the uncased BERT base model (Devlin et al., 2018; Liu et al., 2020b). Specifically, we use 0.1 for the dropout rate except 0.2 for MNLI, 0.01 for the weight decay rate and the Adamax (Kingma and Ba, 2014) optimizer with the default Lookahead (Zhang et al., 2019) to stabilize training. We select the learning rate from $\{5\text{e}{-5}, 1\text{e}{-4}\}$ for all the models. The maximum training epoch is set to 6, and the we follow (Jiang et al., 2019) to set adversarial training hyperparameters: $\epsilon = 1\text{e}{-5}$ and $\eta = 1\text{e}{-4}$. In our experiments, we simply set $\alpha = 1$ in Eq 1.

3.2 Standard GLUE Evaluation

We first compare adversarial training methods on the standard GLUE benchmark (Wang et al., 2018). See Table 1 for the results [1]. TAT consistently outperforms both standard training and the state-of-the-art adversarial training methods of FreeLB and VAT. Remarkably, BERT$_{\text{BASE}}$ with targeted adversarial training performs on par with BERT$_{\text{LARGE}}$ with standard training overall, and outperforms the latter by a large margin on tasks with smaller datasets such as RTE, MRPC and STS-B, which illustrates the benefit of TAT in improving model generalizability.

[1] Due to restriction on the number of submissions by the GLUE organizers, we only compared TAT with the published results from (Devlin et al., 2018) on the test set.

Method	HANS Acc	SNLI Acc	SciTail Acc	MedNLI Acc
Standard	55.4	80.1	77.3	43.2
FreeLB	62.0	80.5	78.6	56.8
VAT	62.5	80.8	78.5	58.1
TAT	**65.8**	**81.0**	**78.8**	**60.6**

Table 2: Comparison of standard and adversarial training in zero-shot evaluation on various natural language inference datasets, where the standard BERT$_{\text{BASE}}$ model is fine-tuned on the MNLI training data.

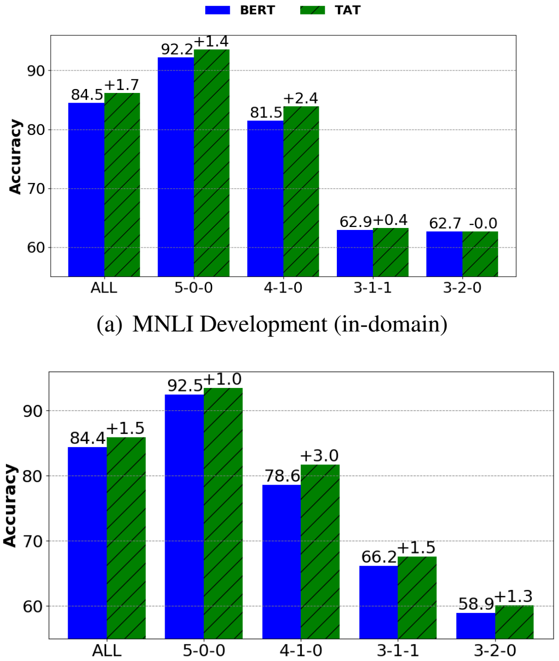
3.3 Zero-Shot Learning on Natural Language Inference

Next, we compare standard and adversarial training in generalizability to out-domain datasets. Specifically, we fine-tune BERT$_{\text{BASE}}$ on the MNLI training data and evaluate it on various natural language inference test sets: HANS (McCoy et al., 2019), SNLI (Bowman et al., 2015), SciTail (Khot et al., 2018), MeNLI (Romanov and Shivade, 2018). See Table 2 for the results. TAT substantially outperforms standard training and state-of-the-art adversarial training methods. Interestingly, the gains are particularly pronounced on the two hardest datasets, HANS and MedNLI. HANS used heuristic rules to identify easy instances for MNLI-trained BERT models and introduced modifications to make them harder. MedNLI is from the biomedical domain, which is substantially different from the general domain of MNLI. This provides additional evidence that targeted adversarial training is especially effective in enhancing generalizability in out domains.

3.4 Zero-Shot Learning on Cross-Lingual Natural Language Inference

We also conducted zero-shot evaluation in the cross-lingual setting by comparing standard and adversarial training on XNLI (Conneau et al., 2018). Specifically, a cross-lingual language model is fine-tuned using the English NLI dataset and then tested on datasets of other languages. Following Conneau et al. (2019), we used the pre-trained XLM-R large model in our experiments, and compare targeted adversarial training (XLM-R+TAT) with state-of-the-art systems that use standard training (XLM-R) and adversarial training (XLM-R+R3F/R4F) (Aghajanyan et al., 2020), as well as another state-of-the-art language model InfoXLM (Chi et al., 2020). To ensure fair comparison, we also report the results from our reimplementation of XLM-R (Conneau et al., 2018) (XLM-R$_{\text{Reprod}}$). See Table 3 for the results. Targeted adversarial training (TAT) demonstrates a clear advantage in improving zero-shot transfer learning across languages, especially for languages most different from English, such as Urdu. Overall, TAT produces a new state-of-the-art result of 81.7% over 15 languages on XNLI.

3.5 Analysis



(a) MNLI Development (in-domain)

(b) MNLI Development (out-domain)

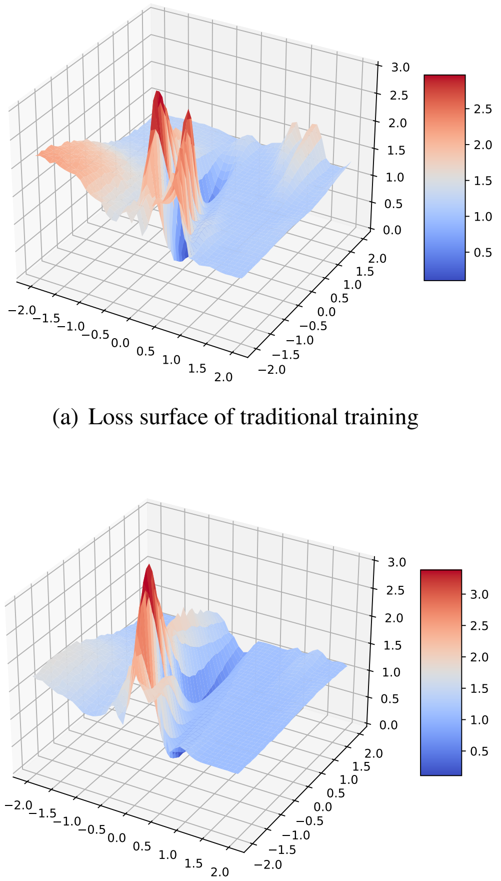
Figure 2: Comparison of standard and targeted adversarial training on MNLI, subdivided per agreement.

As we have seen in Figure 1 earlier, TAT reduces the errors across the board on MNLI development set. To understand how TAT improves performance, we conducted a more detailed analysis by subdividing the dataset based on the degree of human agreement. Here, there are three label classes and each sample instance has 5 human annotations. The samples can be divided into four categories: 5-0-0, 4-1-0, 3-2-0, 3-1-1. E.g., 3-1-1 signifies that there are three votes for one label and one for each of the other two labels. In Figure 2, we see that TAT outperforms the baseline consistently over all categories, with higher improvement on the more ambiguous samples, especially for out-domain samples. This suggests that TAT is most helpful for the challenging instances that exhibit higher ambiguity and are more different from training examples.

We also visualize the loss landscape of both the standard training and TAT, shown in Figure 3. TAT

Model	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg.
XLM-R	89.1	84.1	85.1	83.9	82.9	84.0	81.2	79.6	79.8	80.8	78.1	80.2	76.9	73.9	73.8	80.9
XLM-R$_{\text{Reprod}}$	88.1	83.6	84.1	83.0	82.6	83.8	81.7	80.7	80.4	80.7	78.9	80.1	77.8	74.2	74.0	80.9
XLM-R+R3F	89.4	84.2	85.1	83.7	83.6	84.6	82.3	80.7	80.6	81.1	79.4	80.1	77.3	72.6	74.2	81.2
XLM-R+R4F	89.6	**84.7**	85.2	**84.2**	83.6	84.6	**82.5**	80.3	80.5	80.9	79.2	80.6	78.2	72.7	73.9	81.4
InfoXLM	**89.7**	84.5	85.5	84.1	83.4	84.2	81.3	80.9	80.4	80.8	78.9	80.9	77.9	**74.8**	73.7	81.4
XLM-R+TAT	89.3	84.2	**85.7**	83.9	**83.7**	**85.0**	82.1	**81.0**	**80.7**	**81.3**	**79.7**	**81.0**	**78.4**	74.1	**75.1**	**81.7**

Table 3: Comparison of targeted adversarial training (TAT) and prior state of the art in zero-shot cross-lingual learning on the XNLI test set.



(a) Loss surface of traditional training

(b) Loss surface of TAT

Figure 3: Training loss surfaces of traditional training vs TAT on MNLI.

has a wider and flatter loss surface, which generally indicates better generalization (Hochreiter and Schmidhuber, 1997; Hao et al., 2019; Li et al., 2018).

4 Conclusion

We present the first study to apply targeted attacks in adversarial training for natural language understanding. Our TAT algorithm is simple yet effective in improving model generalizability for various NLP tasks, especially in zero-shot learning and for out-domain data. Future directions include: applying TAT in pretraining and other NLP tasks e.g., sequence labeling, exploring alternative approaches for target sampling.

Acknowledgments

We thank Microsoft Research Technology Engineering team for setting up GPU machines. We also thank the anonymous reviewers for valuable discussions.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, and Danilo Giampiccolo. 2006. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth pascal recognizing textual entailment challenge. In *In Proc Text Analysis Conference (TAC'09*.

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 (EMNLP)*. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Hao Cheng, Xiaodong Liu, Lis Pereira, Yaoliang Yu, and Jianfeng Gao. 2020. Posterior differential regularization with f-divergence for improving model robustness. *arXiv preprint arXiv:2010.12638*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Berlin, Heidelberg. Springer-Verlag.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

William B Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of bert. In *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 4134–4143.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Flat minima. *Neural Computation*, 9(1):1–42.

Yu-Lun Hsieh, Minhao Cheng, Da-Cheng Juan, Wei Wei, Wen-Lian Hsu, and Cho-Jui Hsieh. 2019. On the robustness of self-attentive models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1520–1529.

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. 2019. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, pages 6389–6399.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020a. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Xiaodong Liu, Yu Wang, Jianshu Ji, Hao Cheng, Xueyun Zhu, Emmanuel Awa, Pengcheng He, Weizhu Chen, Hoifung Poon, Guihong Cao, and Jianfeng Gao. 2020b. The Microsoft toolkit of multitask deep neural networks for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

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.

Lis Pereira, Xiaodong Liu, Fei Cheng, Masayuki Asahara, and Ichiro Kobayashi. 2020. Adversarial training for commonsense inference. *arXiv preprint arXiv:2005.08156*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.

Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free! *arXiv preprint arXiv:1904.12843*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. 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.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.

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. *arXiv preprint arXiv:1804.07461*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2018. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*.

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. Association for Computational Linguistics.

Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. 2019. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pages 9597–9608.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Thomas Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv:1909.11764*.

A NLU Benchmarks

The NLU benchmarks used in our experiments, i.e. GLUE benchmark (Wang et al., 2018), SNLI (Bowman et al., 2015), SciTail (Khot et al., 2018), HANS (McCoy et al., 2019), MedNLI (Romanov and Shivade, 2018) and XNLI (Conneau et al., 2018), are briefly introduced in the following sections. Table 4 summarizes the information of these tasks. In the experiments, GLUE is used for the normal setting, while the other datasets are used for the zero-shot setting.

• **GLUE**. The General Language Understanding Evaluation (GLUE) benchmark is a collection of nine natural language understanding (NLU) tasks. As shown in Table 4, it includes question answering (Rajpurkar et al., 2016), linguistic acceptability (Warstadt et al., 2018), sentiment analysis (Socher et al., 2013), text similarity (Cer et al., 2017), paraphrase detection (Dolan and Brockett, 2005), and natural language inference (NLI) (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009; Levesque et al., 2012; Williams et al., 2018). The diversity of the tasks makes GLUE very suitable for evaluating the generalization and robustness of NLU models.

• **SNLI**. The Stanford Natural Language Inference (SNLI) dataset contains 570k human annotated sentence pairs, in which the premises are drawn from the captions of the Flickr30 corpus and hypotheses are manually annotated (Bowman et al., 2015). This is the most widely used entailment dataset for NLI.

• **SciTail**. This is a textual entailment dataset derived from a science question answering (SciQ) dataset (Khot et al., 2018). The task involves assessing whether a given premise entails a given hypothesis. In contrast to other entailment datasets mentioned previously, the hypotheses in SciTail are created from science questions while the corresponding answer candidates and premises come from relevant web sentences retrieved from a large corpus. As a result, these sentences are linguistically challenging and the lexical similarity of premise and hypothesis is often high, thus making SciTail particularly difficult.

• **MedNLI**. This is a textual entailment dataset in the clinical domain. It was derived from medical history of patients and annotated by doctors. The task involves assessing whether a given premise entails a given hypothesis. The hypothesis sentences in this dataset were generated by clinicians, while corresponding answer candidates and premises come from MIMIC-III v1.3 (Johnson et al., 2016), a database containing 2,078,705 clinical notes written by healthcare professionals. Its specialized domain nature makes MedNLI a challenging dataset.

• **HANS**. This is an NLI evaluation set that tests three hypotheses about invalid heuristics that NLI models are likely to learn: lexical overlap (assume that a premise entails all hypotheses constructed from words in the premise), subsequence (assume that a premise entails all of its contiguous subsequences), and constituent. HANS is a challenging dataset that aims to test how much models are vulnerable to such heuristics, and standard training often results in models failing catastrophically, even models such as BERT (McCoy et al., 2019).

• **XNLI**. This is a cross-lingual natural language inference dataset built by extending the development and test sets of the Multi-Genre Natural Language Inference Corpus (Williams et al., 2018) to 15 languages, including low-resource languages such as Swahili. This corpus was designed to evaluate cross-language sentence understanding, where models are supposed to be trained in one language and tested in different ones. Validation and test sets are translated from English to 14 languages by professional translators, making results across different languages directly comparable (Artetxe and Schwenk, 2019).

Corpus	Task	#Train	#Dev	#Test	#Label	Metrics
Single-Sentence Classification (GLUE)						
CoLA	Acceptability	8.5k	1k	1k	2	Matthews corr
SST	Sentiment	67k	872	1.8k	2	Accuracy
Pairwise Text Classification (GLUE)						
MNLI	NLI	393k	20k	20k	3	Accuracy
RTE	NLI	2.5k	276	3k	2	Accuracy
WNLI	NLI	634	71	146	2	Accuracy
QQP	Paraphrase	364k	40k	391k	2	Accuracy/F1
MRPC	Paraphrase	3.7k	408	1.7k	2	Accuracy/F1
QNLI	QA/NLI	108k	5.7k	5.7k	2	Accuracy
Text Similarity (GLUE)						
STS-B	Similarity	7k	1.5k	1.4k	1	Pearson/Spearman corr
Pairwise Text Classification for the Zero-shot setting						
SNLI	NLI	-	-	9.8k	3	Accuracy
SciTail	NLI	-	-	2.1k	2	Accuracy
HANS	NLI	-	-	3k	2	Accuracy
MedNLI	NLI	-	-	1.4k	3	Accuracy
XNLI	NLI	-	-	75k	3	Accuracy

Table 4: Summary information of the NLU benchmarks.