Smoothing with Fake Label

Ziyang Luo[*][†]
Hong Kong Baptist University
Hong Kong, China
chiyeunglaw1@gmail.com

Yadong Xi[*]
Fuxi AI Lab, NetEase Inc.
Hangzhou, China
xiyadong@corp.netease.com

Xiaoxi Mao
Fuxi AI Lab, NetEase Inc.
Hangzhou, China
maoxiaoxi@corp.netease.com

ABSTRACT

Label Smoothing is a widely used technique in many areas. It can prevent the network from being over-confident. However, it hypotheses that the prior distribution of all classes is uniform. Here, we decide to abandon this hypothesis and propose a new smoothing method, called *Smoothing with Fake Label*. It shares a part of the prediction probability to a new fake class. Our experiment results show that the method can increase the performance of the models on most tasks and outperform the *Label Smoothing* on text classification and cross-lingual transfer tasks.

CCS CONCEPTS

• **Computing methodologies → Regularization**; *Natural language processing*; *Neural networks*.

KEYWORDS

label smoothing, neural networks, text classification, cross-lingual, machine translation

ACM Reference Format:
Ziyang Luo, Yadong Xi, and Xiaoxi Mao. 2021. Smoothing with Fake Label. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482184

1 INTRODUCTION

Text classification [5, 10, 12, 21] is a widely studied task in natural language processing and has wide applications. For example, given a movie review, judge whether its attitude is positive or negative. Given a Twitter, judge whether it is a rumor. In recent years, we usually use the deep neural networks to achieve such goal, including the convolution neural networks [11], the recurrent neural networks [2, 8] and the Transformer [23]. To train these models, we usually use a large amount of training data and hope that they can generalize well on the test data.

The previous work [22] shows that a neural network is overfitting when it places all probability on a single class in the training set, which will hurt models' generalization ability. To alleviate this problem, they introduce a simple smoothing method, called *Label Smoothing*. It shares a part of the true label's probability to all classes. People use it widely to improve the performance of neural models across different tasks, including machine translation [23] and image recognition [20, 29]. However, this method relies on a hypothesis that the prior distribution of all classes is uniform. This may not be satisfied by all classification tasks. Therefore, one can find that in text classification or some multilingual tasks, people seldom use this method.

To fill this gap, we decide to abandon this hypothesis and not to share the probability to all classes, but only a new fake class. We call this smoothing method as *Smoothing with Fake Label (FLS)*. We use multiple NLP tasks to prove the effectiveness of our smoothing method, including semantics analysis, natural language inference, sentence meaning similarity and machine translation. Our experiment results show that it can increase our models' performance across a wide range of tasks.

2 SMOOTHING WITH FAKE LABEL

2.1 Motivation

We first give a brief introduction of the *Label Smoothing*. [22] first introduces this method in their work, which shares a part of probability to all classes. Suppose that we have a K-class classification task, a training sample can be denoted as $(x^{(n)}, y^{(n)})$ for $n = 1, ..., N$ and $y^{(n)} \in \{1, 2, ..., K\}$. θ is the parameter of a model. Their method is described as follows:

$$\begin{aligned}\mathcal{L}(\theta) = &- \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}(y^{(n)} = k) \log P(k|x^{(n)}) \\ &- \lambda \sum_{n=1}^{N} \frac{1}{K} \sum_{k=1}^{K} \log P(k|x^{(n)}).\end{aligned} \tag{1}$$

The first term is the cross-entropy loss and the second term is the smoothing term. This method relies on a hypothesis that the prior distribution of all classes is uniform, $\frac{1}{K}$, which cannot be satisfied by all tasks. In order to alleviate this problem, we remove such prior and propose our method in the following section.

2.2 Our Method

In this part, we illustrate our smoothing method and call it as *Smoothing with Fake Label (FLS)*. We manually create a new fake label $K + 1$ for the K-class classification task. The loss function

[*]Both authors contributed equally to this research.
[†]Work was done during internship at NetEase Inc..

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11...$15.00
https://doi.org/10.1145/3459637.3482184

becomes:

$$\mathcal{L}(\theta) = -\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{1}(y^{(n)} = k)\log P(k|x^{(n)}) - \lambda\sum_{n=1}^{N}\log P(K+1|x^{(n)}), \tag{2}$$

where λ controls the amount of probability on the fake label. The first term is also the cross-entropy loss. For the second term, we require our models to share some probabilities on the fake label for every sample. Our method can avoid the prior assumption of uniform distribution and also play a smoothing role.

3 EXPERIMENTS SETUPS

To examine the effectiveness of our method, we train different models with our method on different tasks.

3.1 Tasks

We first employ different text classification tasks of GLUE [24] to evaluate our method.

- Sentiment analysis task
  - SST-2: the Stanford Sentiment Treebank [21]
- Natural language inference tasks
  - QNLI: Question-answering NLI based on the Stanford Question Answering Dataset [19]
  - MNLI: the Multi-Genre Natural Language Inference Corpus [25]
- Sentence meaning similarity tasks
  - RTE: the Recognizing Textual Entailment datasets[1]
  - QQP: the Quora Question Pairs dataset[2]
  - MRPC: the Microsoft Research Paraphrase Corpus [6]

We report the scores on the validation, rather than test data, so the results are different from the original Roberta paper [13].

We also include a cross-lingual task, XNLI [4] to evaluate how our method affects the cross-lingual transfer ability of the model. This task is a multilingual version MNLI task. Its test set is translated into 15 different languages. We train our models with English training data and evaluate them with 15 different languages' test data.

Apart from this, we also evaluate our method on machine translation task with IWSLT2014 German-English parallel dataset [1]. Since *Label Smoothing* is widely used in machine translation [7, 23], we also want to analyze whether our method is useful in this area.

3.2 Models

For the English text classification tasks and the cross-lingual task, our models are two state-of-the-art pre-trained language models, RoBERTa [13] and XLM-R [3]. We choose to use the based version models, which contain 12 layers Transformer Encoder block, 768 hidden size and 12 attention heads. For the machine translation task, we use the Transformer Encoder-Decoder Seq2seq structure. Both the Encoder and Decoder have 6 layers, 512 hidden size and 4 attention heads.

Since training such models requires a large amount of computational resources, it is difficult (and environmentally costly) for individual researchers to do so independently. Luckily, one can download these pre-trained models from various community resources. FairSeq [18] and Huggingface Transformers [26] are two well-known package for pre-trained language models. We download the RoBERTa-base model from FairSeq[3] and XLM-R-base model from Huggingface.[4] Then we fine-tune these models with our scripts.

	RoBERTa	+LS	Ours
RTE	78.38	78.36	**78.52**
SST-2	94.78	94.76	**94.98**
QNLI	92.74	**92.78**	**92.78**
MNLI	87.58	87.60	**87.62**
QQP	**91.60**	91.40	91.48
MRPC	87.44	87.36	**87.64**
Avg.	88.75	88.71	**88.84**

Table 1: Accuracy results for English tasks. All results are averaged over five different seeds. Bold indicates the best result of every task. (LS = *Label Smoothing*)

3.3 λ Design

The value of λ is important for our method. For the text classification, if λ is too large, models can only learn to predict the fake class. We carefully design the λ value for each text classification task. We find that λ being smaller than $\frac{1}{K+1}$ is fine for a K-class classification problem. All λ values are in Section 8. For machine translation, we surprisingly find that λ can be larger. We test ten different values in our experiments from 0.1 to 1.0. For *Label Smoothing*, λ is set to 0.1 for all tasks.

3.4 Training Details

We fine-tune the RoBERTa-base model and training the machine translation models with Fairseq. We directly use the hyper-parameters which are recommended by Fairseq[5]. To avoid overstating, we average all results with five different random seeds (1,2,3,4 and 5) for the English text classification tasks.

For the cross-lingual task, we fine-tune XLM-R-base model with Huggingface Transformers. We use a batch size of 8 and train for 3 epochs, optimized by AdamW [14]. The max length of each sentence is 128. If the length of a sentence exceeds 128, we clip this sentence.

4 RESULTS AND ANALYSIS

English Text Classification Tasks Table 1 shows that RoBERTa-base cannot benefit from *Label Smoothing*, which corroborates our claims in the introduction that people seldom use it in text classification tasks. Its uniform distribution hypothesis prior is not suitable for most tasks. After removing this prior, our method improves the model's performance on most tasks, which reveals the effectiveness of our method.

[1] https://aclweb.org/aclwiki/Recognizing_Textual_Entailment
[2] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs
[3] http://dl.fbaipublicfiles.com/fairseq/models/roberta.base.tar.gz
[4] https://huggingface.co/xlm-roberta-base
[5] https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.glue.md

Models	en	fr	es	bg	zh	de	ru	ar	th	sw	el	tr	vi	hi	ur	Avg.
Fine-tuning Models on English Training Set																
XLM-R	83.9	77.5	77.9	76.8	72.9	76.1	74.7	71.2	71.8	64.3	75.4	72.5	73.3	68.6	66.1	73.5
+LS	84.4	77.5	77.9	77.0	72.9	75.7	74.9	71.3	70.3	62.8	75.0	72.3	73.6	69.2	65.3	73.3
Ours	**84.9**	**79.0**	**79.3**	**78.1**	**74.7**	**77.3**	**76.8**	**72.4**	**72.7**	**64.7**	**76.2**	**73.0**	**75.8**	**71.0**	66.0	**74.8**

Table 2: Accuracy results for XNLI task. Bold indicates the best result of every language. (LS = *Label Smoothing*)

	De-En		En-De	
	Valid	Test	Valid	Test
Transformer	34.37	33.52	29.18	27.41
+LS	**35.53**	**34.67**	**30.00**	**28.48**
Ours (λ = 0.5)	35.03	**34.32**	29.61	28.37
Ours (λ = 0.6)	**35.25**	34.31	29.75	28.31
Ours (λ = 0.7)	35.06	34.14	**29.76**	**28.46**

Table 3: BLEU results for machine translation task. Bold indicates the best two results. (LS = *Label Smoothing*)

Cross-lingual Transfer Task Table 2 illustrates that our method outperforms *Label Smoothing* on every language. Some languages' results increase more than 1%. For example, the Thai(th) score of *Label Smoothing* is 70.3, which is 2.4 lower than our method. Although *Label Smoothing* can increase the accuracy score of English, it harms the model's cross-lingual transfer ability. Model's performance on some languages is degenerated, like German(de), Thai(th) and Swedish(sw). This indicates that the uniform distribution prior of *Label Smoothing* is not suitable for the cross-lingual transfer learning scenario.

Comparing our method with the raw model's results, we can find that removing the uniform distribution prior, the performance of XLM-R on all languages except ur increases. This indicates that our method can improve the model's cross-lingual transfer ability.

Machine Translation This task is distinguishable from the previous tasks. We consider it as a classification task with thousands of classes. We add a fake token into the vocabulary during training our models. When we are translating a new sentence or calculating the perplexity, we manually set the logit of the fake label to be $-\infty$. This ensures that the fake token will not appear in the translation results and affect the perplexity.

Table 3 illustrates that models with *Label Smoothing* have the highest BLEU scores on German-English translation validation and test sets. The BLEU scores increase more than one point. This corroborates the previous research's results [23] that *Label Smoothing* can improve model's performance on machine translation. Though our method lag behind *Label Smoothing*, it still outperforms the original models, which proves the effectiveness of our method. For example, when λ = 0.6, the BLEU score of De-En validation set increases about 0.9 and only 0.28 point lower than *Label Smoothing*. From Figure 1, we surprisingly find that increasing the value of λ will not degenerate the model's performance, which indicates that the machine translation tasks are quite different from the text classification.

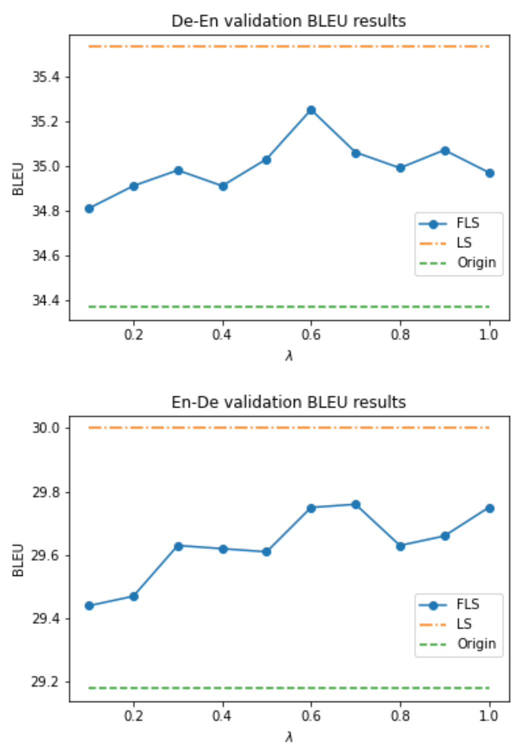

Figure 1: Validation BLEU results for De-En and En-De validation sets. The λ value of *Label Smoothing* is fixed to 0.1. (FLS = *Smoothing with Fake Label*, LS = *Label Smoothing*)

We also analyze how these smoothing methods affect the perplexity. Figure 2 illustrates that the larger λ of our method leads to lower perplexity. When λ = 1.0, it has the lowest PPL on the De-En validation set, which is lower than *Label Smoothing*. All PPL scores of the two smoothing methods are lower than the original model by a large margin. These results reveal that the smoothing method can reduce the PPL, which is different from previous work [23].

5 DISCUSSION

The analyses and experiments in this work point out that the uniform distribution prior hypothesis *Label Smoothing* is not suitable for all tasks, especially for the cross-lingual transfer learning. Our work abandon this prior and smooth with a fake label, which outperform *Label Smoothing* on some tasks. However, we still can find that it is less useful in machine translation than *LS*, which indicates that some tasks can benefit from the inductive bias of the uniform

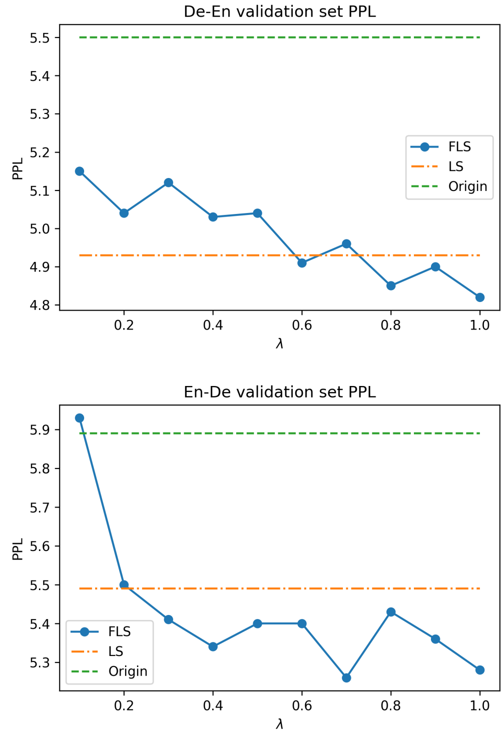

Figure 2: PPL results for De-En and En-De validation sets. The λ value of *Label Smoothing* is fixed to 0.1. (FLS = *Fake Label Smoothing*, LS = *Label Smoothing*)

distribution prior. It is worthy to analyze how to adjust the prior distribution for different tasks in the future works.

In our machine translation tasks, we also surprisingly find that *Label Smoothing* and our method do not make the perplexity worse. Since in most of the previous works [7, 23], they find that *LS* will increase the perplexity. Though we do not attempt to dispute these claims with our findings, we do hope our experiments will figure out the role of different smoothing methods.

6 RELATED WORK

Label Smoothing is first proposed by Szegedy et al. [22] and widely used in computer vision area. Much works focus on understanding this method [15–17, 27, 28]. We find that the uniform distribution prior is not suitable for all tasks and propose a new smoothing method.

Most similar to our work, [9] uses the Pseudo-Labels in the image classification task. For un-labeled data, they just pick up the class which has the maximum predicted probability and use it as the true labels. However, our method does not need the un-labeled data.

7 CONCLUSION

In this work, we propose a new label smoothing method, called *Smoothing with Fake Label*, which outperforms *Label Smoothing* on text classification and cross-lingual transfer tasks. For machine translation tasks, using our method is comparable to *Label Smoothing*.

Our experiment results show that both our method and *Label Smoothing* promote the performance poorly on the GLUE benchmark while promote the performance with a relatively remarkable margin on the Machine Translation. Future works will explore when label smoothing can bring in improvement for a task. In addition, we will explore how to adaptively adjust the prior distribution for a specific task.

Tasks	λ Value
RTE	0.30
SST-2	0.26
QNLI	0.26
MNLI	0.10
QQP	0.22
MRPC	0.12
XNLI	0.25

Table 4: Different λ values for different tasks.

8 APPENDIX

Table 4 shows the values of λ for different tasks.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their excellent feedback.

REFERENCES

[1] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and Marcello Federico. 2015. Report on the 11 th IWSLT Evaluation Campaign , IWSLT 2014.

[2] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555 [cs.NE]

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[4] Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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* (Brussels, Belgium). Association for Computational Linguistics.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. https://www.aclweb.org/anthology/I05-5002

[7] Yingbo Gao, Weiyue Wang, Christian Herold, Zijian Yang, and Hermann Ney. 2020. Towards a Better Understanding of Label Smoothing in Neural Machine Translation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 212–223. https://www.aclweb.org/anthology/2020.aacl-main.25

[8] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[9] Dong hyun Lee. [n.d.]. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks.

[10] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[11] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. https://doi.org/10.1109/5.726791

[12] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent Neural Network for Text Classification with Multi-Task Learning. arXiv:1605.05101 [cs.CL]

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[14] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG]

[15] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. 2020. Does label smoothing mitigate label noise?. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 6448–6458. http://proceedings.mlr.press/v119/lukasik20a.html

[16] Clara Meister, Elizabeth Salesky, and Ryan Cotterell. 2020. Generalized Entropy Regularization or: There's Nothing Special about Label Smoothing. arXiv:2005.00820 [cs.CL]

[17] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2020. When Does Label Smoothing Help? arXiv:1906.02629 [cs.LG]

[18] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 48–53. https://doi.org/10.18653/v1/N19-4009

[19] 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*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[20] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. 2019. Regularized Evolution for Image Classifier Architecture Search. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 4780–4789. https://doi.org/10.1609/aaai.v33i01.33014780

[21] 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*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://www.aclweb.org/anthology/D13-1170

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs.CV]

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[24] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Brussels, Belgium, 353–355. https://doi.org/10.18653/v1/W18-5446

[25] 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

[26] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[27] Yi Xu, Yuanhong Xu, Qi Qian, Hao Li, and Rong Jin. 2020. Towards Understanding Label Smoothing. arXiv:2006.11653 [cs.LG]

[28] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. 2020. Delving Deep into Label Smoothing. arXiv:2011.12562 [cs.CV]

[29] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2018. Learning Transferable Architectures for Scalable Image Recognition. arXiv:1707.07012 [cs.CV]