Knowledge Distillation with Noisy Labels for Natural Language Understanding

Shivendra Bhardwaj[1]* **Abbas Ghaddar**[1] **Ahmad Rashid**[1] **Khalil Bibi**[1]
Chengyang Li[1,2]† **Ali Ghodsi**[2] **Philippe Langlais**[3] **Mehdi Rezagholizadeh**[1]
[1]Huawei Noah's Ark Lab
[2]David R. Cheriton School of Computer Science, University of Waterloo
[3] RALI/DIRO, Université de Montréal, Canada
{abbas.ghaddar, ahmad.rashid,khalil.bibi, mehdi.rezagholizadeh}@huawei.com
ali.ghodsi@uwaterloo.ca, felipe@iro.umontreal.ca

Abstract

Knowledge Distillation (KD) is extensively used to compress and deploy large pre-trained language models on edge devices for real-world applications. However, one neglected area of research is the impact of noisy (corrupted) labels on KD. We present, to the best of our knowledge, the first study on KD with noisy labels in Natural Language Understanding (NLU). We document the scope of the problem and present two methods to mitigate the impact of label noise. Experiments on the GLUE benchmark show that our methods are effective even under high noise levels. Nevertheless, our results indicate that more research is necessary to cope with label noise under the KD.

1 Introduction

Large-scale pre-trained language models (Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020) have shown remarkable abilities to match and even surpass human performances on many Natural Languages Understanding (NLU) tasks (Rajpurkar et al., 2018; Wang et al., 2018, 2019a). However, the deployment of these models in dynamic commercial environments come with challenges, including: large model size, and low training data quality.

Knowledge Distillation (Hinton et al., 2015; Turc et al., 2019) is a compression technique of choice that has proven to be effective to fit a cumbersome NLU model on edge devices (Sanh et al., 2019; Jiao et al., 2020; Sun et al., 2020). Meanwhile, numerous methods were developed to combat noisy (corrupted) labels, mainly for computer vision (Frénay and Verleysen, 2013; Jiang et al., 2018; Thulasidasan et al., 2019; Han et al., 2020) and more recently for NLU (Ardehaly and Culotta, 2018; Jindal et al., 2019; Garg et al., 2021; Ghaddar et al., 2021a,b; Jafari et al., 2021).

Despite its success, KD has mostly been studied with the availability of massive amount of high quality labeled data. In practice, however, it is costly and impractical to produce such data (Ghaddar and Langlais, 2019), and noisy labels are commonly encountered. In this paper, we consider the problem of KD when noisy labels are provided for training the main (teacher) and compressed (student) models. To our knowledge, this is the first time KD is studied under a noisy setting in NLU.

We conduct experiments on 7 tasks from the GLUE benchmark (Wang et al., 2018) and observe a drastic drop of performance of distilled models when we increase the level of noise. In response, we propose 2 distillation training methods, namely Co-Distill and Label Refining, that are specifically designed to handle noise. Experiments show that our methods lead to improvements over fair baselines, and that it combination also performs the best. Yet, our analysis indicates that the problem is far from solved, and that there is much room for research.

2 Related Work

The vanilla KD framework (Buciluǎ et al., 2006; Hinton et al., 2015) consists in training a small *student* model to mimic the output of a large *teacher* model. Recent years have seen a wide array of methods that leverage intermediate layer matching (Ji et al., 2021; Wu et al., 2020; Passban et al., 2021; Wang et al., 2020), data augmentation (Fu et al., 2020; Li et al., 2021; Jiao et al., 2020; Kamalloo et al., 2021), or adversarial training (Zaharia et al., 2021; Rashid et al., 2020, 2021) in order to reduce the teacher-student performance gap. Instead, our proposed methods are designed to handle label noise during KD. Nevertheless, they can be easily fused with the aforementioned methods to further boost performance.

*This work has been done while Shivendra Bhardwaj was at Huawei.

† This work has been done while Chengyang Li was an intern at Huawei.

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 297–303

Label noise (corruption) is a common problem in real-world datasets, and it has been well studied in the literature (Frénay and Verleysen, 2013; Li et al., 2017; Han et al., 2020). Methods to combat noise build on the idea that samples with small training loss at early epochs are more likely to be clean (Dehghani et al., 2018; Wang et al., 2019b).

In co-teaching (Han et al., 2018), two networks of different capacity teach each other to reject wrong labels. At each forward pass, each network keeps only small-loss samples and sends them to its peer network for updating the parameters. The main idea is that the error flow can be reduced, as networks of different learning abilities have different views on the data.

Self-distillation was proposed by Dong et al. (2019), where the model is trained to mimic its own prediction from the previous training epoch. The goal is to prevent the model from memorizing wrong labels, as the model has less tendency to fit noise at early epochs. In addition, Bagherinezhad et al. (2018) showed improvements when distillation at early epochs is used to refine noisy labels.

Another line of works is the learning to weight approach (Ren et al., 2018; Li et al., 2019; Zhang et al., 2020; Fan et al., 2020) that aims to learn per-sample loss weights in order to discount noisy samples. The proposed methods use an auxiliary meta-learner to re-weight training samples of the main model. However, all aforementioned works mainly focus on computer vision. Recently, Garg et al. (2021) utilize a noise detection model to cluster, then score the training samples for text classification in an attempt to guide the main model to focus on samples that are most likely to be correct.

3 Methodology

We first introduce our method, Co-Distill (CD), which jointly trains the teacher and the student. Next, we incorporate Label Refinement (LR) which is motivated by the algorithms of Jiang et al. (2018), Arazo et al. (2019) and Garg et al. (2021) for noise mitigation in regular (no KD) training framework.

3.1 Co-Distill (CD)

The key feature of our method is that the teacher and the student are trained together, but unlike traditional KD, the teacher also learns from the student. Figure 1 showcases the complete architecture. We train the student model $S_{\theta^{\mathrm{S}}}(\cdot)$ with the following loss function $\mathcal{L}^{\mathrm{S}}$:

$$\mathcal{L}^{\mathrm{S}} = \frac{1}{N}\sum_{i=1}^{N}[\alpha \cdot \mathcal{L}_{CE}(y_i, S_{\theta^{\mathrm{S}}}(x_i)) + (1-\alpha)\cdot \mathcal{L}_{KD}(T_{\theta^{\mathrm{T}}}(x_i), S_{\theta^{\mathrm{S}}}(x_i))] \quad (1)$$

where θ^{T} and θ^{S} are the teacher and student parameters respectively, α is the KD weight parameter, $\mathcal{L}_{\mathrm{CE}}$ is the Cross Entropy (CE) loss, y_i is the label, N is the total number of training samples and $\mathcal{L}_{KD}$ is the symmetric Kullback-Leibler (KL) divergence (Kullback, 1997) between the teacher and the student logits, i.e. we sum both the forward and reverse KL.

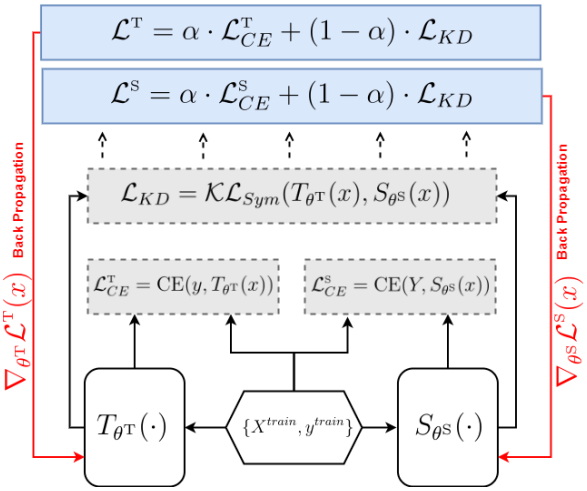

Figure 1: The Co-Distill architecture

In addition to CE loss, the teacher "learns" from the student and is trained to minimize the $\mathcal{L}_{KD}$ loss. It is worth mentioning that we always train the teacher at the first epoch with an α value of 1. We do so to avoid propagating low confident information to the teacher at the beginning of the training. After the first epoch, the feedback of $\mathcal{L}_{KD}$ improves the overall performance of both teacher and student models.

3.2 CD plus Label Refinement (CD+LR)

We further enhance CD by refining the training labels based on loss values at early epochs. In LR, an auxiliary classifier is trained to flag noisy samples, which in turn are re-labeled by the main model. In prior work on noisy labels (Arpit et al., 2017; Dehghani et al., 2018; Wang et al., 2019b) it has been observed that small training losses at early

Model	CoLA	SST-2	MRPC	RTE	QNLI	QQP	MNLI	Avg.
				0%				
BERT-base	61.9	93.1	90.9	68.6	91.6	91.6	85.0	83.2
w/o KD	51.3	91.3	87.5	59.9	89.2	88.5	82.1	78.5
Vanilla	56.4	92.0	90.0	68.6	90.3	90.6	85.0	81.8
				25%				
BERT-base	46.7	91.5	75.7	61.0	87.9	72.4	81.8	73.9
with CD	47.5	93.2	78.7	62.5	88.2	71.3	82.8	74.9
with CD+LR	48.6	92.8	78.7	60.3	88.6	74.0	82.3	75.0
w/o KD	39.1	90.4	79.9	61.0	84.5	67.3	79.3	71.6
Self-DSTL	43.5	90.5	79.7	60.6	84.1	69.3	80.0	72.5
Vanilla	40.2	90.9	79.2	61.7	85.9	72.9	80.3	73.0
CD	45.1	90.6	79.2	63.2	85.5	70.5	80.7	73.5
CD+LR	**46.1**	**91.3**	**80.9**	**63.5**	**86.9**	**73.8**	**81.0**	**74.8**
				50%				
BERT-base	17.7	56.7	68.4	59.2	64.7	63.6	76.5	58.1
with CD	16.0	56.5	70.8	55.6	62.7	71.3	76.8	58.5
with CD+LR	17.2	61.7	71.8	56.3	62.7	71.3	77.0	59.7
w/o KD	8.3	55.0	66.6	57.0	56.5	67.3	72.2	54.7
Self-DSTL	8.8	57.6	68.1	58.4	57.3	69.3	73.2	56.1
Vanilla	11.9	60.0	68.6	**58.5**	60.3	66.1	75.0	57.2
CD	13.6	60.3	69.1	57.4	60.0	70.5	75.1	58.0
CD+LR	**17.7**	**64.1**	**71.1**	57.4	**60.9**	**73.8**	**76.6**	**60.2**

Table 1: Performances on GLUE dev sets of models trained on 0%, 25%, and 50% of noisy labels. Dash lines separate teacher (up) and student models.

epochs are more likely to indicate that a sample is clean.

Instead, we assume that we have access to a small subset of validation data where noisy and clean samples are known a priori (see Section 4.1). We train both teacher and student with Co-Distill for 2 epochs,[1] and then calculate $\mathcal{L}_{CE}^{T}$ and $\mathcal{L}_{CE}^{S}$ for each sample in the validation set.

We use these values as features for a discriminator model $D(.)$ trained to predict whether a sample is noisy. Once it is trained, $D(.)$ is used to flag noisy training samples, so that the teacher re-labels them. Finally, we resume the co-distillation for the remaining epochs while calculating the CE loss using the new labels.

4 Experiments

4.1 Dataset and Evaluation

We experiment on 7 tasks from the GLUE benchmark (Wang et al., 2018): 2 single-sentence (CoLA and SST-2) and 5 sentence-pair (MRPC, RTE, QQP, QNLI, and MNLI) classification tasks. Following prior work, we report Matthews correlation on CoLA and accuracy for the other tasks. Since GLUE test sets are hidden and the number of submissions to leaderboard is limited, we held-out 10% of the training set for validation and used the rest for training. We used this validation set to train the discriminator as well as for hyper-parameter tuning, while official GLUE dev sets are used to evaluate the models.

We test our methods on training sets with 25% and 50% noisy labels[2]. We introduce the same level of noise for the validation sets. Following prior

[1] Empirically, we found that it works well on most of the tasks we experimented on.

[2] We do not evaluate beyond 50% of noise because many GLUE tasks are binary classification.

works (Jiang et al., 2018; Dong et al., 2019; Garg et al., 2021), we inject artificial noise by randomly changing the original labels of the training samples.

4.2 Baselines

We compare our noise mitigation methods with 3 popular baselines:

- **w/o KD** In this setting, only the CE loss is used. This baseline is used as a witness.
- **Vanilla-KD** Here, we select the best performing α value for each task.
- **Self-DSTL** In Self-Distillation (Dong et al., 2019), the student is first trained for few epochs on hard labels only, and the best checkpoint is used to generate logits on the training data. For the rest of the epochs the student is trained on both hard and its own soft labels.

4.3 Implementation

We use as our teacher the 12-layer BERT-base-uncased model (Devlin et al., 2019), and the pre-trained 6-layer distillBERT (Sanh et al., 2019) to initialize all student models. We use `scikit-learn` (Pedregosa et al., 2011) to train a Random Forest discriminator (Breiman, 2001) as our auxiliary classifier. For all models, we perform hyper-parameter tuning and best model selection based on early stopping on noisy validation sets. We report average results over 3 random seeds.

4.4 Results

Table 1 shows performances on GLUE dev sets of 3 teachers and 5 student models trained on clean (0%), 25% and 50% of noisy training sets. As expected, the performance of all models drops drastically with noise. For instance, the teacher and vanilla student average performances drop by 25.1% and 24.7% respectively when we train with 50% of noisy labels. Among all baselines, training the student solely on hard labels (w/o KD) performs the worst under all levels of noise.

Performing distillation with the student logits itself (Self-DSTL) slightly improves the performances by 0.5% and 1.5% on 25% and 50% noise level respectively. However, using teacher logits (Vanilla) for distillation always performs better than using that of the student by 1% on average. This indicates that the teacher knowledge remains crucial even under a noisy label setting.

Overall, our method CD leads to an average gain of 0.5% and 0.8% on top of the Vanilla baseline at 25% and 50% noise level respectively. Moreover, enhancing the CD methods with label refinement (CD+LR) significantly boosts these scores by 1.3% and 2.2% respectively. CD+LR consistently outperforms Vanilla KD across all tasks and noise levels, except at 50% noise for MRPC. It is worth noting that our methods are more effective under extreme noise level, since the gap with Vanilla KD gets larger at 50% noise level (the max for binary classification).

On the teacher side, we observe that the teachers obtained with our methods outperform the other teachers. The CD+LR teacher is better than its naive counterpart by 1.1% and 1.6% on 25% and 50% noise level respectively. This observation is inline with Han et al. (2018) who find that in Co-teaching, the two networks communicating with each other get improved. More interestingly, results show that CD+LR students outperform significantly (>2%) the naive and slightly (0.2%) their respective teachers. This is mainly due to the tendency of over-parameterized neural networks (teachers) to fit noisy labels (Han et al., 2018; Jiang et al., 2018), compared to smaller models (students in our cases). This suggests that in a high noise setting, training a robust teacher is important as much as training the student.

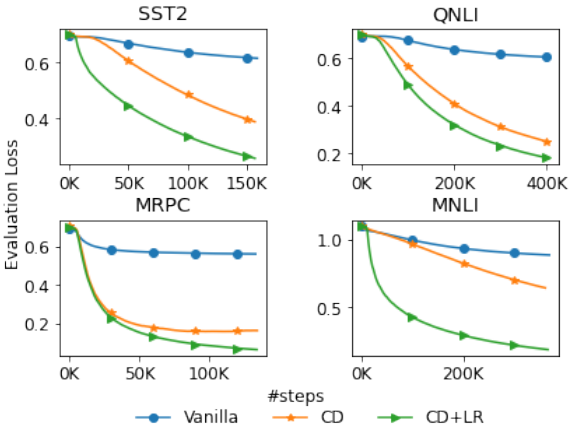

Figure 2: Validation (on GLUE dev sets) curve for the 3 student models trained on 50% of noisy labels.

We plot the losses on dev sets at early steps to better understand how our methods combat noisy labels. Figure 2 shows dev loss values on 4 GLUE tasks[3] for Vanilla KD, CD, and CD+LR methods. First, we observe that the loss curve of Vanilla KD

[3]Similar figures are observed on the remaining 3 tasks.

flattens at early stages. We investigated the training loss and noticed that it rather decreases, mainly due over-fitting the noise labels.

Co-Distillation (CD) shows better signs of mitigating noise, as the loss decreases slowly on MNLI and sharply on QNLI and MRPC. Adding LR leads to a sharp drop, followed by a steady decrease of loss values. The drop happens immediately after refining the training set labels, which seems crucial for large datasets like MNLI and SST-2.

5 Conclusion

We present the first study on Knowledge Distillation when learning from noisy labels in NLU, and show that the problem is extremely challenging. Future work involves conducting a comparative study on the robustness of state-of-the-art KD techniques against noisy labels, and merging them within our methods. We hope that our study will encourage future research on KD in the noisy label setting, a genuine setting in real world applications.

Acknowledgments

We thank Mindspore[4] for the partial support of this work, which is a new deep learning computing framework.

[normalem]ulem

References

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pages 312–321. PMLR.

Ehsan Mohammady Ardehaly and Aron Culotta. 2018. Learning from noisy label proportions for classifying online social data. *Social Network Analysis and Mining*, 8(1):1–18.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks.

Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. 2018. Label refinery: Improving imagenet classification through label progression. *arXiv preprint arXiv:1805.02641*.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Mostafa Dehghani, Arash Mehrjou, Stephan Gouws, Jaap Kamps, and Bernhard Schölkopf. 2018. Fidelity-weighted learning. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bin Dong, Jikai Hou, Yiping Lu, and Zhihua Zhang. 2019. Distillation = early stopping? harvesting dark knowledge utilizing anisotropic information retrieval for overparameterized neural network. *stat*, 1050:2.

Yang Fan, Yingce Xia, Lijun Wu, Shufang Xie, Weiqing Liu, Jiang Bian, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2020. Learning to teach with deep interactions. *arXiv preprint arXiv:2007.04649*.

Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

Jie Fu, Xue Geng, Zhijian Duan, Bohan Zhuang, Xingdi Yuan, Adam Trischler, Jie Lin, Chris Pal, and Hao Dong. 2020. Role-wise data augmentation for knowledge distillation. *arXiv preprint arXiv:2004.08861*.

Siddhant Garg, Goutham Ramakrishnan, and Varun Thumbe. 2021. Towards robustness to label noise in text classification via noise modeling. *arXiv preprint arXiv:2101.11214*.

Abbas Ghaddar and Philippe Langlais. 2019. Contextualized word representations from distant supervision with and for ner. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 101–108.

Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, and Mehdi Rezagholizadeh. 2021a. Context-aware adversarial training for name regularity bias in named entity recognition. *Trans. Assoc. Comput. Linguistics*, 9:586–604.

Abbas Ghaddar, Philippe Langlais, Mehdi Rezagholizadeh, and Ahmad Rashid. 2021b. End-to-end self-debiasing framework for robust NLU training. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event,*

[4] https://www.mindspore.cn/

August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1923–1929. Association for Computational Linguistics.

B Han, Q Yao, X Yu, G Niu, M Xu, W Hu, IW Tsang, and M Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *32nd Conference on Neural Information Processing Systems (NIPS)*. NEURAL INFORMATION PROCESSING SYSTEMS (NIPS).

Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. 2020. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Aref Jafari, Mehdi Rezagholizadeh, Pranav Sharma, and Ali Ghodsi. 2021. Annealing knowledge distillation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2493–2504, Online. Association for Computational Linguistics.

Mingi Ji, Byeongho Heo, and Sungrae Park. 2021. Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313. PMLR.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4163–4174.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. 2019. An effective label noise model for dnn text classification. 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 3246–3256.

Ehsan Kamalloo, Mehdi Rezagholizadeh, Peyman Passban, and Ali Ghodsi. 2021. Not far away, not so close: Sample efficient nearest neighbour data augmentation via minimax. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3522–3533. Association for Computational Linguistics.

Solomon Kullback. 1997. *Information theory and statistics*. Courier Corporation.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2019. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5051–5059.

Tianda Li, Ahmad Rashid, Aref Jafari, Pranav Sharma, Ali Ghodsi, and Mehdi Rezagholizadeh. 2021. How to select one among all? an extensive empirical study towards the robustness of knowledge distillation in natural language understanding. *arXiv preprint arXiv:2109.05696*.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. 2021. ALP-KD: attention-based layer projection for knowledge distillation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13657–13665. AAAI Press.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Ahmad Rashid, Vasileios Lioutas, Abbas Ghaddar, and Mehdi Rezagholizadeh. 2020. Towards zero-shot knowledge distillation for natural language processing. *CoRR*, abs/2012.15495.

Ahmad Rashid, Vasileios Lioutas, and Mehdi Rezagholizadeh. 2021. MATE-KD: Masked adversarial TExt, a companion to knowledge distillation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1062–1071, Online. Association for Computational Linguistics.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170.

Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. 2019. Combating label noise in deep learning using abstention. *arXiv preprint arXiv:1905.10964*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

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*, pages 353–355.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019b. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 322–330.

Yimeng Wu, Peyman Passban, Mehdi Rezagholizadeh, and Qun Liu. 2020. Why skip if you can combine: A simple knowledge distillation technique for intermediate layers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1016–1021. Association for Computational Linguistics.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2021. Dialect identification through adversarial learning and knowledge distillation on romanian BERT. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial@EACL 2021, Kiyv, Ukraine, April 20, 2021*, pages 113–119. Association for Computational Linguistics.

Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. 2020. Distilling effective supervision from severe label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9294–9303.