Why Knowledge Distillation Amplifies Gender Bias and How to Mitigate - from the Perspective of DistilBERT

Jaimeen Ahn[*,†] Danggeun Market Inc.
Hwaran Lee Naver AI LAB
Jinhwa Kim Naver AI LAB
Alice Oh KAIST

Abstract

Knowledge distillation is widely used to transfer the language understanding of a large model to a smaller model. However, after knowledge distillation, it was found that the smaller model is more biased by gender compared to the source large model. This paper studies what causes gender bias to increase after the knowledge distillation process. Moreover, we suggest applying a variant of the mixup on knowledge distillation, which is used to increase generalizability during the distillation process, not for augmentation. By doing so, we can significantly reduce the gender bias amplification after knowledge distillation. We also conduct an experiment on the GLUE benchmark to demonstrate that even if the mixup is applied, it does not have a significant adverse effect on the model's performance.

1 Introduction

Knowledge distillation (Hinton et al., 2015) is one way to use the knowledge of a large language model under the limited resources by transferring the knowledge of a larger model to a smaller model. Under the supervision of the teacher model, the small model is trained to produce the same result as that of the teacher model. By doing so, small models can leverage the knowledge of larger models (Sanh et al., 2019).

To maintain the performance of the model trained by knowledge distillation, the distilled model focuses more on the majority appearing in the data (Hooker et al., 2020). Recent studies have described that pre-trained language model also results in a more biased representation when distillation proceeds (Silva et al., 2021). However, only the issue is reported, and what part of knowledge distillation causes an increase in bias is not explored, and no solution is provided.

This paper studies which part of knowledge distillation causes the increase of social bias and how to alleviate the problem in terms of DistilBERT (Sanh et al., 2019). We first examine what part that contributes to knowledge distillation brings social bias amplification. There is no difference between the distilled and original models except for size and training loss. Thus, we check from two perspectives: (1) the capacity of the model being distilled and (2) the loss used in knowledge distillation. Then we suggest leveraging *mixup* (Zhang et al., 2018) on the knowledge distillation loss to mitigate this amplification by giving generalizability during the training.

We conduct the experiments from two measurements: social bias with the Sentence Embedding Test (SEAT) (May et al., 2019) and downstream task performance with the GLUE Benchmark (Wang et al., 2019). We report that the factors that increase the social bias are the student model's limited capacity and the cross-entropy loss term between the logit distribution of the student model and that of the teacher model. We also demonstrate that applying the *mixup* to knowledge distillation can reduce this increase without significant effect on the downstream task performance.

Our contributions can be summarized as follows:

- We reveal the capacity of the model and cross-entropy loss in knowledge distillation have a negative effect on social bias.
- We suggest mixup as a mitigation technique if it is applied during the knowledge distillation proceeds.

2 Background

Knowledge distillation is trained so that a student model outputs the same output as a teacher model's for one input. It makes the student model have the problem-solving ability of the large model, even though the student model has a smaller structure.

[*] jaime@daangn.com
[†] This is work done during an internship in Naver CLOVA AI LAB.

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 266 - 272

DistilBERT, the model this study is mainly about, is trained with three loss terms. First, cross-entropy loss (L_{ce}) forces the logit distribution between the student model and the teacher model to be similar. Next, the student model learns language understanding itself with masked language modeling loss (L_{mlm}). Lastly, cosine loss between two model's output (L_{cos}) makes the direction of output embeddings between the student model and the teacher model closer (Sanh et al., 2019). In total, the loss term of DistilBERT is as follows:

$$\text{Loss} = L_{ce} + L_{mlm} + L_{cos}.$$

3 Bias Statement

In this paper, we investigate stereotypical associations between male and female gender and attribute pairs, particularly from the perspective of sentence embeddings in knowledge distillation language models. For the attribute pairs, we consider Careers and Family, Math and Arts, and Science and Arts. If there exists a correlation between a certain gender and an attribute, the language model intrinsically and perpetually causes representational harm (Blodgett et al., 2020) through improper preconceptions. Additionally, when the language model is trained for other downstream tasks, such as occupation prediction (De-Arteaga et al., 2019; McGuire et al., 2021), it may lead to an additional risk of gender-stereotyped biases.

Since knowledge distillation (KD) has become a prevalent technique to efficiently train smaller models, it is vital to figure out to what extent the gender biases are amplified after knowledge distillations and which loss terms exacerbate the biases during the training. Our work firstly conducts the in-depth analysis and then proposes mitigation methods for the gender bias amplification during the KD process.

We measure the streotypical associations with the Sentence Embedding Association Test (SEAT) (May et al., 2019) [1]. The SEAT uses semantically bleached sentence templates such as "This is a [attribute-word]" or "Here is [gender-word]". Then the associations between a gender and an attribute are calculated by cosine similarities of sentence encoded embeddings. We leave the detailed equations to calculate the SEAT scores in Appendix B.

There are several tests in SEAT. This study focuses on C6, C7, and C8 categories related to gender bias. C6 tests similarity between embedding of Male/Female Names, and Career/Family attribute words. C7 and C8 measure the similarity between embeddings of male and female pronouns and embeddings of Math/Arts related words and Math/Science related words, respectively.

[1] https://github.com/W4ngatang/sent-bias/

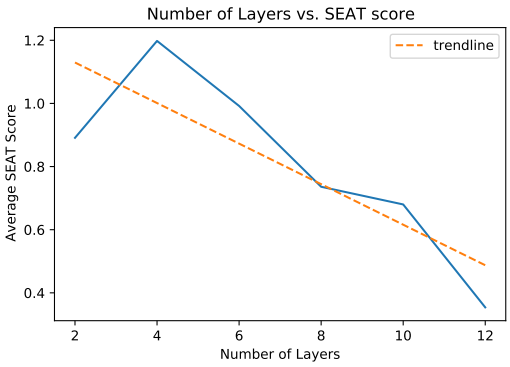

Figure 1: SEAT score by adjusting the number of layers of DistillBERT. The SEAT score and the number of layers in DistillBERT are negatively correlated (Pearson $r = -0.82$).

4 Gender Bias Amplification after KD

In this section, we conduct in-depth analyses about what brings gender bias amplification after knowledge distillation from the perspective of (1) the student model's capacity and (2) the loss used in the knowledge distillation process.

4.1 Experimental Setup

We use 30% of the corpus constructed by two datasets, the Wikipedia dataset and Bookcorpus (Zhu et al., 2015) dataset that were used to create DistillBERT[2]. The distillation is trained for three epochs using four V100 GPUs. All other settings remain the same following the way DistilBERT is trained. We list the settings in Appendix D.

4.2 Does the capacity of the student model matter?

To figure out whether and to what extent the student model's parameter capacity affects the gender biases, we varied the number of layers of the student model (DistilBERT). Note that BERT and DistilBERT have the same architecture parameters except the number of layers. Figure 1 shows

[2] We check the DistilBERT with 30% of the corpus preserves 98.73% of the performance of DistilBERT with the entire dataset on GLUE.

SEAT	Loss Term		
	$L_{\text{mlm}} + L_{\text{cos}} + L_{\text{ce}}$	$L_{\text{mlm}} + L_{\text{ce}}$	$L_{\text{mlm}} + L_{\text{cos}}$
C6	1.236	1.137	**1.093**
C6b	0.499	0.557	**0.292**
C7	**0.907**	1.041	1.153
C7b	1.428	1.316	**0.139**
C8	0.534	**0.475**	0.852
C8b	1.347	1.237	**0.653**
Avg.	0.992	0.960	**0.670**
GLUE Avg.	**76.7**	76.3	75.2

Table 1: SEAT and GLUE scores obtained by ablation of each part in distillation loss. C6 is tested with the names and C7 and C8 are gender pronouns. Thus, for each test, C6b is tested with a gender pronoun, and C7 and C8 are also tested with names.

that the average SEAT scores are increasing as the number of layers is decreasing. Quantitatively, the number of layers has a strong negative correlation with the SEAT score (Pearson $r = -0.82$), which means that the smaller the capacity, the more severe the gender bias. This result also aligns with the previous study that reveals the models with limited capacity tend to exploit the biases in the dataset (Sanh et al., 2021).

4.3 Does the knowledge distillation process matter itself?

To ascertain how each loss term contributes to the increase in SEAT scores in the knowledge distillation process, we conducted an ablation study against each loss term. As shown in Table 1, the model trained without the distillation loss L_{ce} results in the lowest average SEAT score (0.670) among the three loss functions. However, this model shows the lowest performance (75.2%) in the GLUE benchmark, whereas the model trained with all loss terms results the best with 76.7%. This implies that the transfer of the teacher's knowledge is helpful for general language understanding tasks while exacerbating gender bias simultaneously. Consequently, it can be concluded that the current knowledge distillation technique itself is also a factor in increasing gender biases.

5 Mitigation of Bias Amplification

5.1 Proposed method

This section describes how to improve the distillation process to make gender bias not amplified even after knowledge distillation. We found two causes (capacity, loss term) in the previous section. Among them, we decide to modify the loss term because this study is targeting the fixed size model, DistillBERT.

According to the ablation study in Section 4.3, we ascertain distillation loss (L_{ce}) hurts gender bias scores in a huge portion. Our intuition to alleviate this amplification is to give supervision as fair as possible during the knowledge distillation is proceeded. One way is to reduce the SEAT score of the teacher model first and give its supervision to the student model. However, most of the existing methods (Liang et al., 2020b; Cheng et al., 2021) for the teacher are designed to work only on the special token ([CLS]). It is not suitable for knowledge distillation that is trained with logits and embeddings on a token-by-token basis.

In this paper, we use mixup (Zhang et al., 2018) on knowledge distillation to increase gender-related generalization ability by using mixup. Specifically, when a gender-related word appears, we use the values generalized by a mixup in the knowledge distillation process. First, we employ the pre-defined gender word pair (D) set (w_{male} : w_{female}) from the previous work (Bolukbasi et al., 2016)[3]. We next make the *teacher's output logit* (y) and *student's input embedding* (x) same or similar between two corresponding gendered terms with λ drawn from Beta(α, α) when words in D appear:

$$\bar{x} = \lambda x_{w_{\text{male}}} + (1-\lambda) x_{w_{\text{female}}}$$
$$\bar{y} = \lambda y_{w_{\text{male}}} + (1-\lambda) y_{w_{\text{female}}},$$

. We train DistilBERT with the mixup applied instances $(\bar{x}, \bar{y})$ for words in D and with the original instances (x, y) for the rest of words. Notice that we do not use mixup as a data augmentation technique but rather employ its idea in the knowledge distillation.

We view the *mixup* as being worked as a regularizer rather than as a learning objective when knowledge distillation takes place (Chuang and Mroueh, 2021; Liang et al., 2020a). Because the student model learns masked language modeling itself, the generalized gender information by the mixup will act as a regularizer not to be trapped in the information commonly appearing in the pre-training corpus.

5.2 Experimental setup

Dataset We only use the same dataset in knowledge distillation used in Section 4. Also, we lever-

[3]We list the pairs in Appendix C

Supervision		C6	C6b	C7	C7b	C8	C8b	Avg.
Original Supervision	Original Teacher	1.236	0.499	0.907	1.428	0.534	1.347	0.992
	Debiased Teacher (Kaneko and Bollegala, 2021)	**0.889**	0.294	0.509	1.192	0.838	1.292	0.836
Mixup Supervision	Output embeddings	1.215	0.460	0.761	1.541	0.650	1.420	1.008
	Input embeddings	1.305	**0.049**	0.460	1.334	0.465	1.342	0.830
	Logits + Output embeddings	1.310	0.397	1.325	**0.989**	0.863	1.321	1.034
	Logits + Output embeddings + Input embeddings	1.246	**0.049**	0.566	1.367	0.407	**1.144**	0.796
	Logits + Input embeddings (*proposed*)	1.176	0.062	**0.447**	1.218	**0.310**	1.211	**0.738**

Table 2: The result of applying mixup on distillation process in terms of SEAT score (lower scores indicate less social bias). The lowest score on each tests are marked in **bold**.

Task	Original Teacher	*Mixup* in distillation
MNLI	80.6	80.4
QQP	85.9	85.3
QNLI	86.5	86.2
SST-2	90.4	90.7
CoLA	44.8	43.6
STS-B	83.2	83.2
MRPC	82.2	81.7
RTE	59.9	62.1
Avg.	76.7	76.7

Table 3: The performance on the GLUE benchmark after applying the proposed mixup (Logits + Input Embeddings) in the knowledge distillation.

age GLUE Benchmark to assess model performance.

Baseline We set a baseline as the distilled model from a teacher model that was trained with a debiasing method (Kaneko and Bollegala, 2021).

5.3 Experimental Results

In Table 2, we report the scores for each SEAT test and the average. It shows that mixup (Zhang et al., 2018) applied in the distillation process outperforms in terms of the average SEAT score. Compared to the baseline, distilled model under the supervision of the debiased teacher, *mixup* scores lower in four out of six tests (C6b, C7, C8, C8b).

Table 2 also shows the results according to the part where the mixup is applied. We experimented with applying *mixup* to many different levels of representations in the distillation process: logits, teacher's output embeddings, and student's input embeddings. The proposed method that applies the mixup to inputs (input embeddings) and labels (logits) showed the best results.

We also measure SEAT after applying the teacher's output embeddings. It is because, although not included in the original distillation, the cosine loss for embedding is included in the learning process of DistilBERT. However, Table 2 reports that the mixup on output embeddings increases the SEAT score in most tests and is even higher than the original distillation process.

We also checked the performance on downstream tasks when *mixup* is applied in knowledge distillation. Table 3 summarizes the results on GLUE benchmark. Compared to the model using the original distillation, the average performance remains the same.

6 Conclusion

In this paper, we study what causes gender bias amplification in the knowledge distillation process and how to alleviate the amplification by applying mixup in the knowledge distillation process. We confirmed that both the cross-entropy loss between the logits and the model capacity affects the increase of gender bias. Since this study focused on the DistilBERT, we alleviated the problem by modifying the knowledge distillation loss. We reported that the SEAT score decreased when the mixup was applied to the student's input embedding and the teacher's output logit in the distillation method when gender-related words appeared. We also showed that this method does not have a significant adverse effect on downstream tasks.

There are limitations in this study. First, we used sub-samples of the pre-training corpus. Although we checked that there was no significant differences when trained with a fraction of data in terms of the SEAT score and the GLUE score, the experimental results for the entire data should be explored. Second, we do not yet know why the SEAT score increases when the mixup is applied to the output embedding. The embeddings between the two genders are expected to be close, but we do not yet figure out why the scores are reversed contrary to expectations. We leave these as our future work.

Acknowledgement

This work has been financially supported by KAIST-NAVER Hypercreative AI Center.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Ching-Yao Chuang and Youssef Mroueh. 2021. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. 2021. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. 2020a. Mixkd: Towards efficient distillation of large-scale language models. *arXiv preprint arXiv:2011.00593*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020b. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Luke McGuire, Tina Monzavi, Adam J. Hoffman, Fidelia Law, Matthew J. Irvin, Mark Winterbottom, Adam Hartstone-Rose, Adam Rutland, Karen P. Burns, Laurence Butler, Marc Drews, Grace E. Fields, and Kelly Lynn Mulvey. 2021. Science and math interest and gender stereotypes: The role of educator gender in informal science learning sites. *Frontiers in Psychology*, 12.

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*.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *International Conference on Learning Representations*.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies

and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Related Work

There were several attempts to apply mixup in knowledge distillation. Du et al. (2021) uses a fair representation created by the medium of the embeddings of two sensitive attributes (the neutralization) in distillation. Students are trained with the neutralized embeddings created in this way so that the student's input is dependent on the teacher's output. MixKD (Liang et al., 2020a) applies mixup during knowledge distillation to get better performance on the GLUE benchmark. Notably, MixKD takes the method of training the teacher model as well as the student model when distillation proceeds. Our suggestion guarantees independence between student and teacher model inputs in this work, as DistilBERT is trained. Moreover, we train a task-agnostic model by applying a mixup to distillation.

B Sentence Embedding Association Test (SEAT)

Let X and Y be target embeddings, the embedding of sentence template with gender word in our case, and A and B as attribute words. The SEAT basically measures similarity difference between attribute words and target word w. So the similarity difference on word w is

$$s(w, A, B) = [\text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)].$$

The SEAT score (d) is the Cohen's d on s. The Cohen's d is calculated as follows:

$$d = \frac{[\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)]}{\text{std\_dev}_{w \in X \bigcup Y} s(w, A, B)}.$$

C Gender Word Pairs

[["woman", "man"], ["girl", "boy"], ["she", "he"], ["mother", "father"], ["daughter", "son"], ["gal", "guy"], ["female", "male"], ["her", "his"], ["herself", "himself"], ["Mary", "John"]]

D Experiment settings: hyperparameters

D.1 Knowledge Distillaton Hyperparameters

- temperature = 2.0
- mlm_mask_prop = 0.15
- word_mask = 0.8
- word_keep = 0.1
- word_rand = 0.1
- mlm_smoothing = 0.7
- n_epoch = 3
- batch_size = 8
- warmup_prop = 0.05
- weight_decay = 0
- learning_rate = 5e-4
- max_grad_norm = 5
- adam_epsilon= 1e-6
- initializer_range= 0.02
- α = 0.4

D.2 GLUE Experiment Hyperparameters

- max_seq_length = 128
- batch_size = 32
- learning_rate = 2e-5
- n_epochs = 3