Interspeech 2024
1-5 September 2024, Kos, Greece

Learning from Multiple Annotator Biased Labels in Multimodal Conversation

Kazutoshi Shinoda, Nobukatsu Hojo, Saki Mizuno, Keita Suzuki, Satoshi Kobashikawa, Ryo Masumura

NTT Corporation, Japan

kazutoshi.shinoda@ntt.com

Abstract

In multimodal conversation analysis, annotating social signals such as speakers' communication skills is inherently subjective and prone to individual annotator bias, which is annotator's tendency to assign labels based on their values. These biases can contribute to label distributions biased towards specific speakers and classes that match annotators' values, leading to degraded classification performance for minority classes and speakers. Existing methods for addressing class imbalance and dataset bias often overlook the variable biases introduced by multiple annotators, which can lead to overfitting to the majority. Thus, we propose a novel two-stage debiasing method, MAD-LM, that first learns the typical label distribution for each annotator and then promotes the learning of untypical labels. MAD-LM effectively mitigates performance degradation for the minority in a multimodal conversation dataset with multiple annotator labels, while maintaining the performance for the majority.

Index Terms: Multimodal Conversation, Annotator Bias, Debias

1. Introduction

Understanding social signals, such as emotion, attitudes, and skills, from verbal and non-verbal information in multimodal conversation is an essential research area for developing human-machine interaction systems [1, 2]. Many datasets have been constructed for multimodal conversation understanding, where conversation videos are annotated by human annotators [3–7].

Annotating social signals [1] in multi-modal conversation tends to involve subjective processes [8], which are susceptible to annotator bias, i.e., annotators often assign labels to speakers based on their values (e.g., confirmation bias [9], stereotypes [10], and various cognitive biases [11]). Existing datasets often only provide the result of "majority voting" to aggregates multiple annotators' labels because they often disagree due to annotators' own values and biases.

In this study, we conjecture that biases unique to each annotator, which were ignored in majority voting, can contribute to label distributions biased towards speakers and classes that match annotators' values in multimodal conversation datasets. See Figure 1 for example. Different annotators assign different distributions of labels across classes and speakers in a recently proposed multimodal conversation dataset [12], which provides multiple annotators' labels. This label imbalance across classes and speakers results in degraded performance in both the minority classes and speakers as shown in Figure 3.

For class imbalance, various methods have been proposed [13]. However, these methods have not considered multiple annotator biases or speaker imbalance as seen in multimodal conversation datasets. On the other hand, existing methods for addressing dataset bias [14], such as ensemble-based debiasing

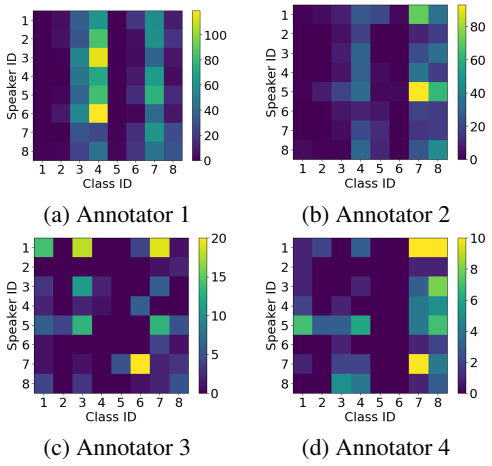

Figure 1: *Distributions of labels annotated by four different annotators in the training set of a dataset for estimating multiple classes of communication skills [12]. These figures indicate that there is a significant variation among different annotators in the label distribution over classes and speakers.*

illustrated in Figure 2 (b), have mainly focused on mitigating over-reliance on a single modality [15, 16]. These methods have not taken into account multiple annotator biases either, which can lead to overfitting to annotator biases.

Therefore, we propose a novel two-stage debiasing approach designed to mitigate the risk of overfitting to biased labels provided by multiple annotators. We tailor ensemble-based debiasing [17–19] to multiple annotator biases in multimodal conversation. In our method, we first train a bias-only model to learn a solution that performs well on typical labels provided by each annotator but fails on untypical ones. Namely, the model intentionally takes only shallow features and annotator IDs as input and predicts labels annotated by the respective annotator. Then, a robust model is trained in an ensemble with the bias-only model to learn untypical labels from multiple annotators that the bias-only model fails to predict.

We conduct experiments on praise estimation, a recently proposed task for estimating speaker's communication skills using speech and video modalities [12]. To the best of our knowledge, this is the only multimodal conversation dataset that provides labels from multiple annotators on each instance, while others provide only the results of majority voting [3–7]. Our experiments show that the proposed method successfully mitigates the performance degradation for the minority classes and speakers. Moreover, our method maintains the performance in the majority

10.21437/Interspeech.2024-1197

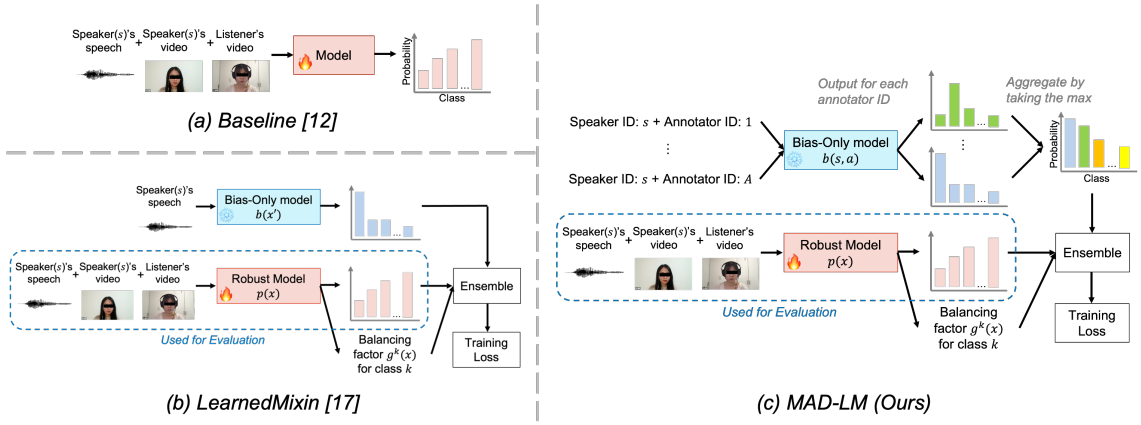

Figure 2: *Overview of the following methods: (a) Baseline [12], (b) LearnedMixin [17], and (c) MAD-LM (Ours). Models with fire marks are trainable and those with snow marks are frozen when training robust models.*

classes and speakers. Comparison to conventional methods for class imbalance, label noise, dataset bias, and ablation study supports the utility of our method. Given the effectiveness of our approach, we suggest that future multimodal conversation datasets should provide multiple annotators' labels to achieve robust classifier performance.

In summary, our contributions are as follows:

- We propose a novel two-stage debiasing approach to prevent the learning of multiple annotator biases, based on our hypothesis that these biases are responsible for data imbalance and performance degradation in the minority classes and speakers.
- Our proposed method effectively mitigates the performance degradation for the minority while maintaining the performance for the majority. Comparison to existing approaches to data imbalance that do not consider multiple annotator biases also supports our hypothesis.

2. Method

In this section, we describe our novel two-stage approach for multiple annotator biases, named Multiple-Annotator Debiasing (MAD). Our MAD first 1) trains a bias-only model (§2.2) and then 2) trains a robust model in an ensemble with the bias-only model (§2.3). We propose and compare two variants of MAD, MAD-PoE and MAD-LM, as explained in §2.3. The overview of training a robust model in MAD-LM is given in Figure 2.

2.1. Notation

Let b be a bias-only model and p be a robust model. Let x be the input to a robust model p, which includes a speaker's speech, speaker's video, and listener's video following [12]. Let $l_a^k \in \{0, 1\}$ be a binary label annotated by annotator $a \in \{1, ..., A\}$ for class $k \in \{1, ..., K\}$, where A and K is the number of annotators and classes, respectively. The ground-truth label $l^k \in \{0, 1\}$ for class k is defined as the logical sum of multiple annotator labels, $l_1^k \vee ... \vee l_A^k$, following [12]. Let $s \in \{1, ..., S\}$ be the speaker identifier for input x, where S denotes the number of speakers.

2.2. Building a Bias-Only Model

We first train a bias-only model to learn prototypical distributions of labels assigned to speaker s by annotator a. Namely, the bias-only model learns who (annotator) labels whom (speaker) in what class. To this end, the model takes only annotator ID a and speaker ID s as input and predicts the corresponding annotator's labels $l_a = \{l_a^k\}_{k=1}^K$. The output of the bias-only model is denoted as $b(a, s) = \{b^k(a, s)\}_{k=1}^K$, where $b^k(a, s)$ is the predicted probability for class k.

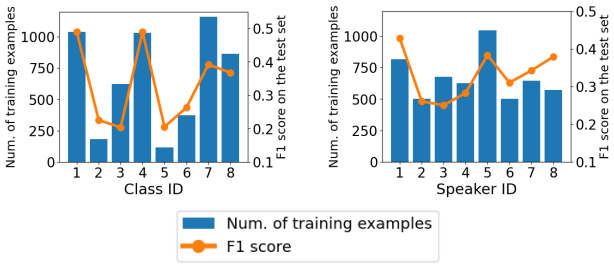

Figure 3: *Frequency of positive utterances in the training set (3,632 utterances in total) and F1 score on the test set for each speaker and class in the praise estimation dataset [12].*

Note that, in existing ensemble-based debiasing methods [17], partial inputs such as only question in visual question answering were taken as input for training a bias-only model. To promote our bias-only model learning the biases on who annotates whom with what class, we originally design the inputs to the bias-only model. We also examined the usefulness of superficial features other than speaker IDs, including word per second, speech duration, speaker's gender, words in utterances, but these failed to improve the performance of the robust model.

2.3. Building a Robust Model

After a bias-only model b is built, we train a robust model by minimizing a binary cross-entropy loss calculated with an ensemble of the two models. We describe two variants of our methods, MAD-PoE and MAD-LM, which are inspired by ensemble-based debiasing methods [17], Product-of-Experts (PoE) and LearnedMixin (LM).

Namely, MAD-PoE is formulated as:

$$\hat{p}^k = \text{sigmoid}(\log p^k(x) + (\log \max_a b^k(a, s))), \quad (1)$$

and MAD-LM is formulated as:

$$\hat{p}^k = \text{sigmoid}(\log p^k(x) + g^k(x)(\log \max_a b^k(a, s))). \quad (2)$$

The difference between the two is the learnable function $g(\geq 0)$. Intuitively, function g balances how much the predictions of the bias-only model should be considered to train the robust model. The predictions of the bias-only model given multiple annotator IDs are aggregated by taking the maximum in Eq. (1) and (2) because the ground-truth labels are also determined by the logical sum of multiple annotator labels as described in §2.1. The loss for training a robust model is defined as follows:

$$\mathcal{L} = \sum_k -l^k \log \hat{p}^k - (1 - l^k) \log(1 - \hat{p}^k) \quad (3)$$

The parameters of the bias-only model are frozen when minimizing $\mathcal{L}$ to train the robust model. The robust model, which takes speech and video as input, makes predictions during evaluation. Note that neither annotator nor speaker IDs are used for testing.

3. Experiments

3.1. Experimental Setups

Dataset: We used the praise estimation dataset constructed in the recent study [12][1] because it provides multiple annotator labels. To the best of our knowledge, there are no other multimodal conversation datasets that provide multiple annotators' labels and annotator identifiers. The praise estimation dataset contains multimodal conversations of business negotiation in Japanese for four topics (chat tools, insurance, TVs, and cell phones). The recorded videos were segmented with a voice activity detection method to obtain the speakers' utterances, which are used as inputs to the robust model as described in §2.1. The dataset contains eight speakers, four of which were women and others were men. Eight classes of preferable behaviors are defined as described in Table 1.

We split the dataset so that the eight speakers' speeches are contained in training, validation, and test sets, while the test set includes only unseen conversations. We did that because our motivation is to see if the performance degradation in the minority classes and speakers (Figure 3) is mitigated by our proposed method. In addition, we showed that the standard training baseline failed to perform well in minority speakers even seen during training. Thus, mitigating the speaker imbalance issue for seen speakers was regarded as the first step in this study. The sizes of the training, validation, and test sets were 3632, 1070, and 2090, respectively. We ensured that there were 458 instances evenly for the eight speakers in the training set.

Model Architecture: For the bias-only model, speaker and annotator IDs were fed to embedding layers to get d-dimensional vectors. For the robust model, the raw inputs (speaker's speech, speaker's video, and listener's video) were fed to modality-specific pretrained encoders to obtain a set of d-dimensional vectors as many as the corresponding time lengths. d was set to 256. The obtained vectors were concatenated across time axis and fed to a 1-layer multi-modal transformer [20] with four attention heads, and then to an attention-pooling layer. The output vector h was of length d. The output probability p^k or b^k was computed by applying a fully connected layer and sigmoid function to h for each class k. For the robust model, the learnable function g^k was computed as $\texttt{softplus}(\text{FC}^k(h))$ where FC^k was also a fully connected layer for class k, and $\texttt{softplus}(z) = \log(1 + e^z)$. For pretrained encoders and preprocessing of raw inputs, we followed the settings in [12].

Training Details: We trained each model for ten epochs. The batch size was 100.[2] We used Adam [27] for optimization with the initial learning rate set to 0.001. We trained each model for five random seeds and reported the average.

Evaluation Metrics: For evaluation metrics, we first compute the F1 score for each pair of speakers and classes and then average the scores to obtain the overall score. This metric weights all the pairs equally, thereby reflecting the robustness to the minority. We split the classes and speakers into three groups based on the number of positive instances in the training set.[3] We also report the averaged F1 scores on the groups to see the effect of our method on the majority and minority.

Baselines: We used the following baselines for comparison: (1) Empirical risk minimization (ERM): standard training with empirical risk minimization on the whole training set [12]. (2) Re-weighting (RW): We used weights that were inversely proportional to the square root of class frequency. (3) focal loss [21]: We used $\alpha = 0.25$ and $\gamma = 2.0$ following the original paper. (4) Class-Balanced (CB) loss [22]: class-level re-weighting considering frequencies of positive instances. (5) CB-focal [22]: CB with focal loss. (6) Class-aware sampling (CAS) [23]: This method samples a fixed number of samples equally from each class. We sampled 321 positive instances from each class for a fair comparison to other methods in terms of the training steps. (7) Rebalanced-BCE (R-BCE) [24]: After applying CAS, R-BCE re-weights training losses by considering the co-occurences of labels. (8) Distribution Balanced (DB) loss [24]: After applying R-BCE, DB loss introduces penalty terms to avoid overfitting to negative instances. (9) DB-focal [24]: DB with focal loss. (10) Co-teaching [25] and (11) Co-teaching+ [26]: These were initially introduced to handle label noise, applied to multimodal conversation tasks by [8]. (12) Product-of-Experts (PoE) [17] and (13) LearnedMixin (LM) [17]: Existing methods for mitigating dataset biases proposed in vision and language tasks. In these methods, uni-modal inputs such as text are used to train bias-only models. We directly applied PoE and LM to multimodal conversation by constructing two variants of uni-modal inputs: speech-only and video-only.

Ablation Study: We conducted an ablation study with regard to bias-only models. We tested two additional variants of bias-only models that take as inputs only speaker or annotator ID. The

Table 1: *Class IDs and their descriptions in the praise estimation dataset [12]. For each class, one binary label is annotated to each utterance by ten annotators.*

Class ID	Description
1	Good consideration for relationship
2	Good ice break
3	Good topic development
4	Good change of topic to suit the customer
5	Good use of honorifics
6	Sufficient nods
7	Sufficient expression of sympathy
8	Sufficient persuasion efforts

[1]This dataset is not publicly available because the participants did not consent to publicly publish the recorded videos.

[2]We used a gradient accumulation technique to increase the batch size because a larger batch size achieved better performance.

[3]Class IDs are grouped into Head (1, 4, 7), Mid (3, 6, 8), and Tail (2, 5). Speaker IDs are grouped into Head (1, 5), Mid(3, 4, 7), and Tail (2, 6, 8).

Table 2: *Experimental results. Macro F1 scores averaged over five random seeds are reported. See §3.1 for the setup.*

Method		Class			Speaker			ALL
		Head	Mid	Tail	Head	Mid	Tail	
ERM [12]		0.4455	0.2816	0.1513	0.3923	0.2719	0.2945	0.3105
Existing Methods for Class Imbalance								
RW		0.4533	0.3941	0.2085	0.4282	0.3502	0.3507	0.3699
focal [21]		0.4027	0.2934	0.2073	0.3541	0.3066	0.2916	0.3129
CB [22]		0.4594	0.3595	0.2112	0.4199	0.3345	0.3453	0.3599
CB-focal [22]		0.4052	0.3308	0.2079	0.3870	0.3107	0.3060	0.3280
CAS [23]		0.4528	0.4221	0.2187	0.4493	0.3718	0.3494	0.3827
R-BCE [24]		0.4542	0.4254	0.2148	0.4473	0.3713	0.3534	0.3836
DB [24]		0.4486	0.3984	0.2310	0.4430	0.3675	0.3382	0.3754
DB-focal [24]		0.3534	0.3167	**0.2344**	0.3449	0.3069	0.2894	0.3099
Existing Methods for Label Noise								
Co-teaching [25]		0.4463	0.3877	0.2187	0.4157	0.3548	0.3478	0.3674
Co-teaching+ [26]		0.4176	0.3864	0.0639	0.3688	0.3119	0.2888	0.3175
Existing Methods for Dataset Bias								
PoE (speech-only) [17]		0.3321	0.2392	0.0670	0.3032	0.2114	0.2025	0.2310
PoE (video-only) [17]		0.3387	0.3009	0.0576	0.3267	0.2375	0.2227	0.2543
LM (speech-only) [17]		0.3389	0.2813	0.1811	0.3622	0.2560	0.2434	0.2779
LM (video-only) [17]		0.3668	0.3434	0.1886	0.3885	0.2943	0.2826	0.3135
Proposed Method								
MAD-PoE		0.4353	0.4200	0.2277	0.4510	0.3613	0.3452	0.3777
MAD-LM		**0.4633**	0.4393	0.2311	**0.4726**	**0.3754**	**0.3662**	**0.3962**
Ablation Study	MAD-LM w/o Annotator ID	0.3887	0.3846	0.2024	0.4228	0.3242	0.3022	0.3406
	MAD-LM w/o Speaker ID	0.4519	**0.4398**	0.2159	0.4625	0.3697	0.3576	0.3884

former predicted l_a while the latter predicted l because the input to the latter is annotator-independent.

3.2. Results

The results are given in Table 2. Our MAD-LM achieved the best scores (ALL), especially for the minority classes and speakers (Tail). Moreover, MAD-LM maintained or even improved the scores for the majority (Head and Mid) compared to existing methods. MAD-PoE slightly underperformed MAD-LM, showing the effectiveness of learnable function g in Eq. (2). This tendency is consistent with previous studies in NLP [28, 29]. The ablation study supports the utility of speaker and especially annotator awareness when constructing the bias-only model.

In addition, our method significantly outperformed existing methods, particularly in enhancing performance for both minority classes and speakers. The existing methods for class imbalance improved the performance for the minority classes (Mid and Tail) but struggled to improve scores on the minority speakers (Tail) compared to our method. This may be because methods for class imbalance do not consider speaker imbalance or annotator biases. The existing methods for label noise and dataset bias also lagged behind our methods in most cases.

4. Analysis of Attention Scores

To gain insights into the effect of our method, we analyzed the attention scores in the multi-modal transformer as in [30]. We computed the attention scores $\texttt{softmax}(QK^T/\sqrt{d})$ averaged over the heads and the test set and then split them according to the input modalities, speech and video. We reported the intra- and inter-modality attention scores for ERM, LM (speech-only) and Ours (MAD-LM).

Table 3: *Intra- and inter-modality attention scores.*

		K					
		speech			video		
		ERM	LM	Ours	ERM	LM	Ours
Q	speech	.91	.90	.87	.09	.10	.13
	video	.56	.47	.52	.44	.53	.48

As shown in Table 3, models trained with ERM may rely on the speech modality to make predictions because the attention scores from speech to speech were larger than others. For LM, attention from video to speech was decreased compared to ERM. Meanwhile, ours reduced the reliance on the speech modality and promoted models to pay attention from speech to video compared to ERM and LM. This analysis suggests that our method improved model robustness by implicitly mitigating the over-reliance on a single modality, referred to as uni-modal biases [15, 16], and enhancing the cross-modal interaction.

5. Conclusion

We proposed a novel two-stage debiasing method that explicitly learned multiple annotators' biased labels for each speaker and class, and then used the information to promote a model to learn a more robust solution. Our experiments showed the utility of multiple annotators' labels to improve the robustness to the minority classes and speakers. Our analysis of attention scores implied that our method enhanced the cross-modal interaction between speech and video to improve the robustness. Future work includes exploring the mechanism behind annotator bias to improve our method.

6. References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli, *Social signal processing*. Cambridge University Press, 2017.

[3] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*, 2006, pp. 28–39.

[4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[5] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *ACL*, 2019, pp. 527–536.

[6] M. Firdaus, H. Chauhan, A. Ekbal, and P. Bhattacharyya, "MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations," in *COLING*, 2020, pp. 4441–4453.

[7] K. Komatani and S. Okada, "Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels," in *ACII*, 2021, pp. 1–8.

[8] Y. Hirano, S. Okada, and K. Komatani, "Recognizing social signals with weakly supervised multitask learning for multimodal dialogue systems," in *ICMI*, 2021, pp. 141–149.

[9] M. A. Gemalmaz and M. Yin, "Accounting for confirmation bias in crowdsourced label aggregation," in *IJCAI*, 2021, pp. 1729–1735.

[10] J. Otterbacher, P. Barlas, S. Kleanthous, and K. Kyriakou, "How do we talk about other people? group (un)fairness in natural language image descriptions," in *HCOMP*, 2019, pp. 106–114.

[11] T. Draws, A. Rieger, O. Inel, U. Gadiraju, and N. Tintarev, "A checklist to combat cognitive biases in crowdsourcing," in *HCOMP*, 2021, pp. 48–59.

[12] N. Hojo, S. Mizuno, S. Kobashikawa, R. Masumura, M. Ihori, H. Sato, and T. Tanaka, "Audio-Visual Praise Estimation for Conversational Video based on Synchronization-Guided Multimodal Transformer," in *INTERSPEECH*, 2023, pp. 2663–2667.

[13] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[14] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011, pp. 1521–1528.

[15] A. Agrawal, D. Batra, and D. Parikh, "Analyzing the behavior of visual question answering models," in *EMNLP*, 2016, pp. 1955–1960.

[16] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "Rubi: Reducing unimodal biases for visual question answering," in *NeurIPS*, 2019.

[17] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases," in *EMNLP-IJCNLP*, 2019, pp. 4069–4082.

[18] H. He, S. Zha, and H. Wang, "Unlearn dataset bias in natural language inference by fitting the residual," in *DeepLo*, 2019, pp. 132–142.

[19] R. Karimi Mahabadi, Y. Belinkov, and J. Henderson, "End-to-end bias mitigation by modelling biases in corpora," in *ACL*, 2020, pp. 8706–8716.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[22] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019, pp. 9260–9269.

[23] L. Shen, Z. Lin, and Q. Huang, "Relay backpropagation for effective learning of deep convolutional neural networks," in *ECCV*, 2016, pp. 467–482.

[24] T. Wu, Q. Huang, Z. Liu, Y. Wang, and D. Lin, "Distribution-balanced loss for multi-label classification in long-tailed datasets," in *ECCV*, 2020, pp. 162–178.

[25] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *NeurIPS*, vol. 31, 2018.

[26] X. Yu, B. Han, J. Yao, G. Niu, I. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *ICML*, 2019, pp. 7164–7173.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] M. Ko, J. Lee, H. Kim, G. Kim, and J. Kang, "Look at the first sentence: Position bias in question answering," in *EMNLP*, 2020, pp. 1109–1121.

[29] K. Shinoda, S. Sugawara, and A. Aizawa, "Look to the right: Mitigating relative position bias in extractive question answering," in *BlackboxNLP*, 2022, pp. 418–425.

[30] H. Xue, Y. Huang, B. Liu, H. Peng, J. Fu, H. Li, and J. Luo, "Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training," in *NeurIPS*, 2021, pp. 4514–4528.