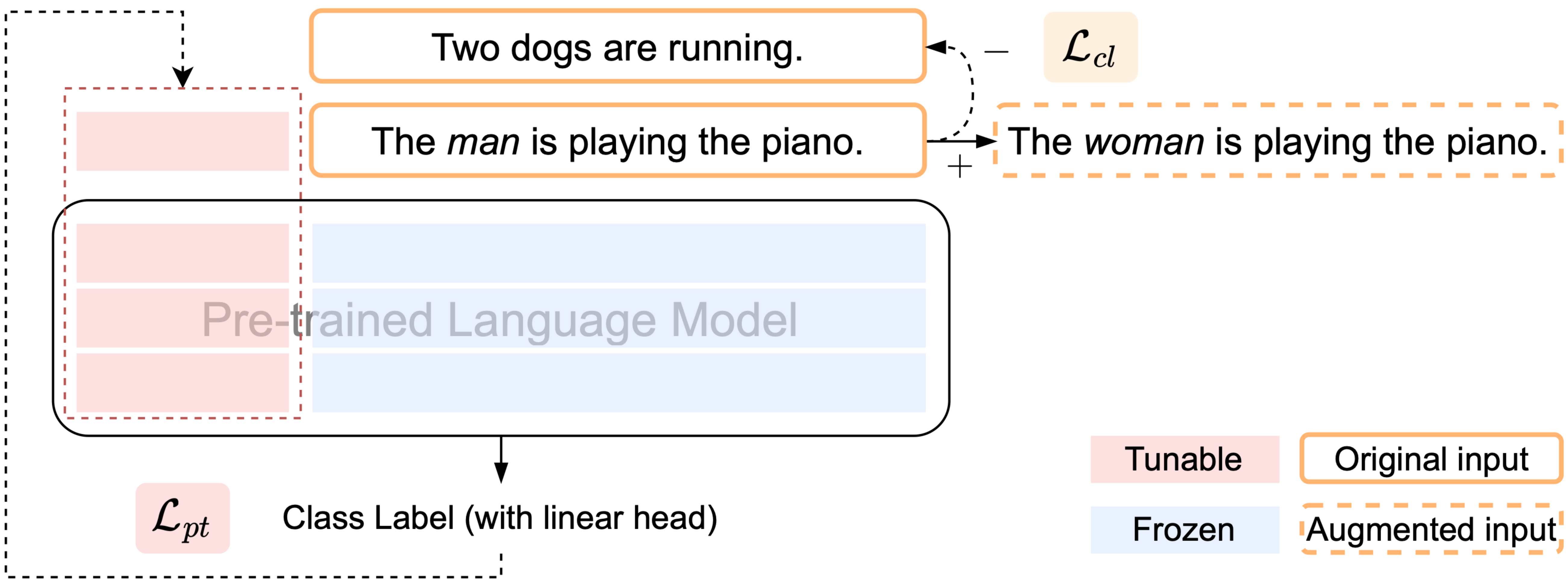
# Co²PT: Mitigating Bias in Pre-trained Language Models through Counterfactual Contrastive Prompt Tuning

Xiangjue Dong[1], Ziwei Zhu[2], Zhuoer Wang[1], Maria Teleki[1], James Caverlee[1]
[1] Texas A&M University, [2] George Mason University




Paper here

## INTRODUCTION

- Relationship between intrinsic and extrinsic benchmarks (which evaluate fairness in downstream applications) **correlates weakly** (Kaneko et al., 2022).
- Models after being debiased, tend to **re-acquire or even amplify biases** during the fine-tuning process on downstream tasks (Zhao et al., 2017; Leino et al., 2019).

We propose **Co²PT**, an efficient and effective *debias-while-prompt tuning* method to mitigate biases via counterfactual contrastive prompt tuning on **downstream tasks**.

## METHODS

- **Deep Prompt Tuning.** We incorporate continuous prompts as prefix tokens in every layer of the PLM, denoted as $\mathcal{L}_{pt}$ .
- **Counterfactual Pairs Construction.** Take the binary-gender debiasing task shown in the above Figure for example, the bias-attribute terms are (***man***, ***woman***), (***he***, ***she***) … The ***man*** is in the input sentence ***The man is playing the piano***. We replace it with ***woman*** while leaving non-attribute words unchanged. Then the counterfactually augmented sentence is ***The woman is playing the piano***.
- **Counterfactual Contrastive Learning.**

$$\mathcal{L}_{cl} = -\log \frac{e^{\mathrm{sim}(\mathbf{p}\oplus\mathbf{h}_i, \mathbf{p}\oplus\mathbf{h}'_i)/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{p}\oplus\mathbf{h}_i, \mathbf{p}\oplus\mathbf{h}'_j)/\tau}}$$

- **Learning Objectives.**

$$\mathcal{L} = \mathcal{L}_{pt} + \alpha\mathcal{L}_{cl}$$

## DATASETS

| | Train | Validation | Bias-Test |
|---|---|---|---|
| STS-B | 5,749 | 1,500 | 16,980 |
| SNLI | 550,152 | 10,000 | 1,936,512 |
| Bias-in-Bios | 255,710 | 39,369 | 98,344 |

## RESULTS

- We observe a **significant reduction** in the bias score. These findings indicate a substantial improvement in the ability to mitigate bias.

| Model | Diff.↓ | $\tau$:0.1↓ | $\tau$:0.3↓ | Pear. / Spear. |
|---|---|---|---|---|
| BERT | 0.282 | 0.867 | 0.417 | 0.883 / 0.879 |
| BERT+CDA | 0.131 | 0.511 | 0.080 | 0.885 / 0.881 |
| ZariCDA* | 0.112 | 0.445 | 0.048 | 0.892 / 0.889 |
| ZariDO* | 0.347 | 0.922 | 0.585 | 0.880 / 0.878 |
| ADELE† | 0.121 | - | - | 0.889 / - |
| Context-Debias | 0.332 | 0.916 | 0.539 | 0.879 / 0.876 |
| Auto-Debias | 0.312 | 0.902 | 0.502 | 0.884 / 0.880 |
| MABEL | 0.066 | 0.204 | 0.013 | 0.889 / 0.885 |
| PT | 0.321 | 0.749 | 0.369 | 0.889 / 0.885 |
| Co²PT (ours) | **0.058** | **0.167** | **0.005** | 0.884 / 0.880 |

- These results clearly demonstrate the **effectiveness of integrating** Co²PT into established debiased models for downstream tasks.

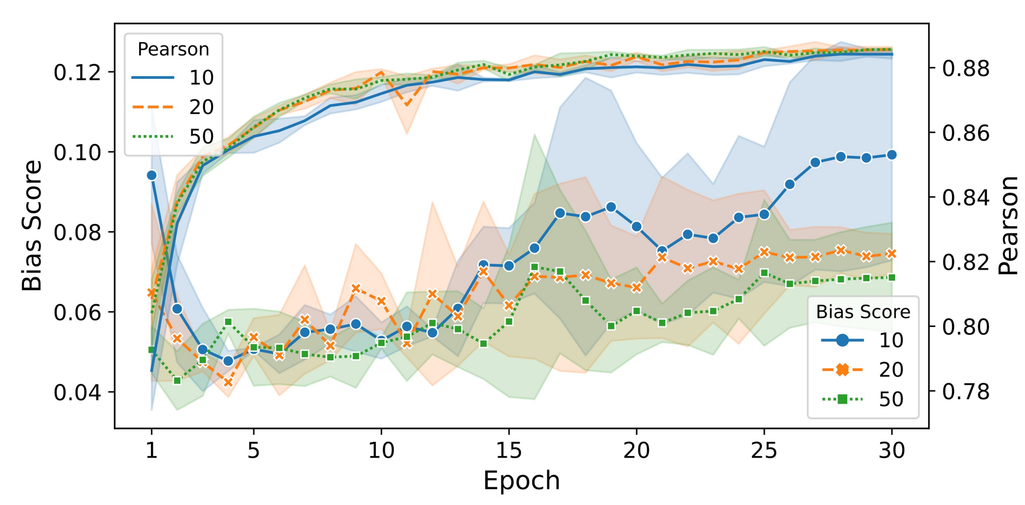
| Model | Diff.↓ | $\tau$:0.1↓ | $\tau$:0.3↓ | Pear. / Spear. |
|---|---|---|---|---|
| Context-Debias | 0.332 | 0.916 | 0.539 | 0.879 / 0.876 |
| + Co²PT | **0.088** | **0.361** | **0.010** | 0.885 / 0.881 |
| Auto-Debias | 0.312 | 0.902 | 0.502 | 0.884 / 0.880 |
| + Co²PT | **0.068** | **0.231** | **0.005** | 0.883 / 0.878 |
| MABEL | **0.066** | **0.204** | 0.013 | 0.889 / 0.885 |
| + Co²PT | 0.068 | 0.228 | **0.005** | 0.892 / 0.889 |

- We perform an extensive **ablation study** to show how different components affect Co²PT.

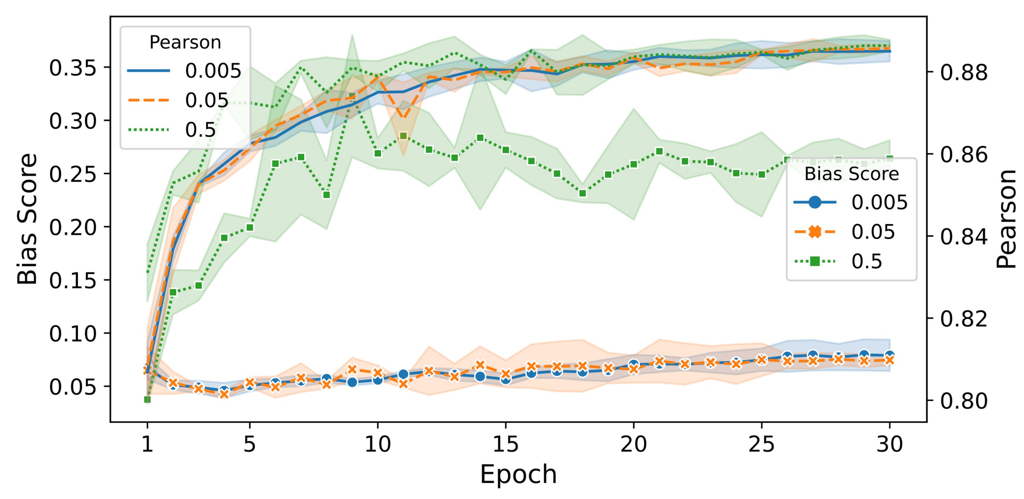
| Model | Diff.↓ | $\tau$:0.1↓ | $\tau$:0.3↓ | Pear. / Spear. |
|---|---|---|---|---|
| PT | 0.321 | 0.749 | 0.369 | 0.889 / 0.885 |
| PT+CDA | 0.291 | 0.747 | 0.351 | 0.890 / 0.886 |
| PT+SCL | 0.161 | 0.548 | 0.133 | 0.883 / 0.878 |
| Co²PT+SCL$_n$ | 0.117 | 0.467 | 0.056 | 0.884 / 0.878 |
| PT+NLI+CL | 0.080 | 0.280 | 0.022 | 0.881 / 0.876 |
| PT+NLI+CL$_p$ | 0.207 | 0.687 | 0.222 | 0.884 / 0.881 |
| PT+CDA+CL$_p$ | 0.271 | 0.725 | 0.338 | 0.886 / 0.883 |
| Co²PT (PT+CDA+CL) | **0.058** | **0.167** | **0.005** | 0.884 / 0.880 |

## IMPACT of HYPERPARAMETERS

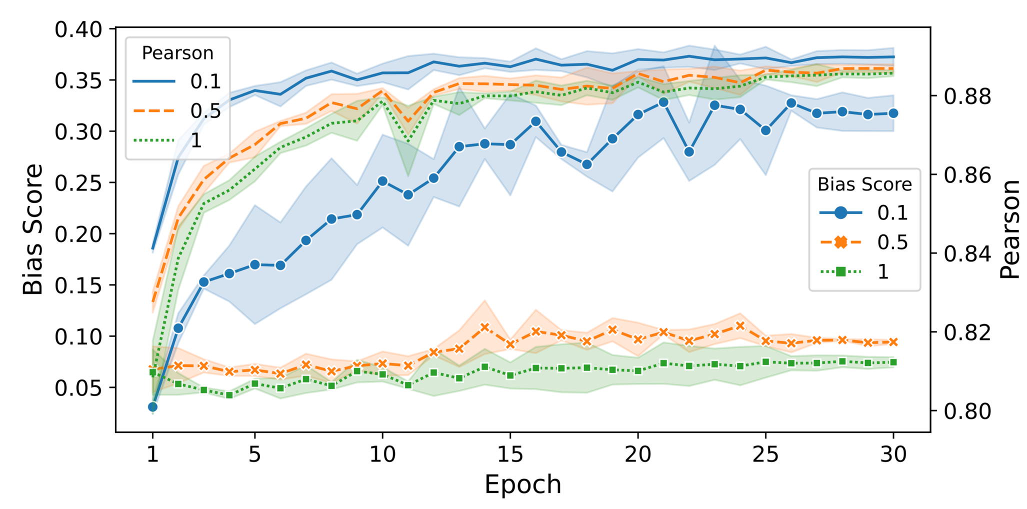
- A larger prompt length enables the model to achieve **better model performance** on downstream tasks more rapidly while still maintaining a lower bias score.




- A higher temperature value corresponds to less weight of the cosine similarity calculation, resulting in decreased effectiveness in bias mitigation.




- A lower coefficient value assigns less weight to the contrastive module, leading to decreased bias mitigation effects.




TEXAS A&M UNIVERSITY
Department of Computer Science & Engineering

Contact me! Actively looking for internship: