

Supplementary Materials: Self-Adaptive Fine-grained Multi-modal Data Augmentation for Semi-supervised Multi-modal Coreference Resolution

Anonymous Authors

A EXPERIMENTAL SETUP

During training, we choose the Adam [3] optimizer with the learning rate of $1e-5$. We train our model by setting the epoch, dropout and batch size to 30, 0.2 and 8 respectively. All our scores are the averaged number over five runs with different random seeds. In Tabel A1, we summarize the hyper-parameter setting.

Table A1: Details of the hyper-parameter settings.

Param.	Value
Leaning rate	$1e-5$
Batch size	8
Epoch size	30
Hidden size	768
Dropout	0.2
Smoothing factor (α_1, α_2)	0.5
filtering intensity β	0.3
Smoothing degree γ	0.4
Separation degree λ	0.6
Loss factor (η_1, η_2)	1.0
CPU	Intel i9
GPU	NVIDIA RTX 3090

B QUALITATIVE ANALYSIS

To gain a deeper understanding of the capabilities of our proposed method, we conduct a comprehensive qualitative analysis on the task of multimodal coreference resolution. Figure A1 presents an example from the CIN dataset [2] where our full model successfully performs multimodal coreference resolution. In contrast, Semi-MCR [1] encounters difficulties in various cases, such as missing predictions for “the glass” and “some objects”, as well as incorrectly identifying “her” as coreferent with “a kid”. Qualitative analysis demonstrates the excellent performance of our method in eliminating textual ambiguities and accurately identifying complete coreference chains, effectively addressing the challenges of multimodal coreference resolution task. These findings highlight the superiority of our SLUDA over Semi-MCR and validate its ability to handle the complex task of coreference resolution in multimodal environments. Furthermore, through our qualitative analysis, we gain deeper insights into the strengths of our approach. We observe that the integration of multimodal information enables our model to leverage both visual and textual cues, resulting in more accurate and comprehensive coreference resolution. By considering the contextual information provided by the multimodal inputs, our

Table A2: Comparison of results and efficiency between using expanded labeled data and not using expanded labeled data.

	CoNLL	MUC	B ³	CEAF _e	Speed(/s)
w/o expanded data	63.59	36.92	78.48	75.39	29.41
with expanded data	66.71	41.43	80.32	78.38	27.03

method effectively resolves textual ambiguities, handles complex coreference relationships, and improves overall performance. Additionally, our analysis reveals the challenges faced by Semi-MCR in handling multimodal coreference resolution. These challenges arise from the limited capacity of Semi-MCR to effectively integrate and exploit visual information, leading to incomplete and inaccurate coreference predictions. This underscores the significance of our proposed method in effectively utilizing multimodal data to enhance coreference resolution.

C ERROR ANALYSIS

In Figure A2, we perform an error analysis on SLUDA to gain valuable insights for future improvements. The analysis reveals a specific issue where our method mistakenly assigns the second occurrence of “one person” and the first occurrence of “one person” to the same coreference chain, as depicted in Figure A2. This misclassification can be attributed to the lack of clear visual features within the blue-boxed region of the image. Insufficient visual clues in this region may result in incorrect pairings during the multimodal coreference resolution task. To address this challenge and improve performance, future work can focus on developing models that are capable of effectively extracting both textual and visual features. One promising direction for future research involves investigating innovative approaches to extract more informative visual features from images. This is particularly crucial in cases where certain regions within the image lack distinct visual clues. By incorporating advanced techniques such as attention mechanisms or region-based analysis, the model can better capture relevant visual information, thereby significantly enhancing the accuracy of multimodal coreference resolution. By exploring these avenues and advancing the extraction of textual and visual features, future models can overcome challenges posed by ambiguous or visually deficient regions, leading to more robust and accurate multimodal coreference resolution.

D EFFICIENCY STUDY

In order to assess the efficiency of our model, we conduct experiments on using labeled data augmentation and not using labeled data augmentation, and the results are shown in Table A2. In terms

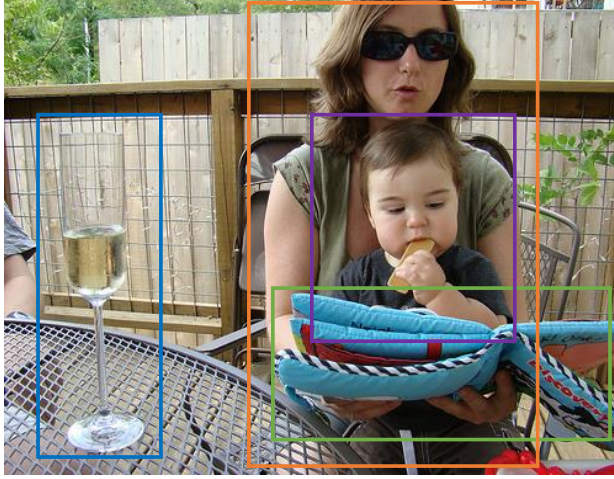


Figure A1: Qualitative results of predictions on the CIN dataset. Like-colored mentions are co-referring.

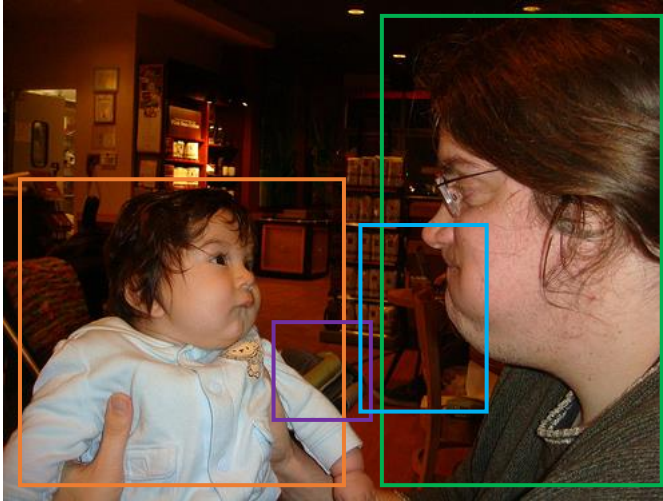


Figure A2: An example of error analysis on SLUDA method. Like-colored mentions are co-referring.

of model performance, we can find that using labeled data augmentation is much better than not using data augmentation on the four commonly used coreference resolution evaluation indicators, with ConLL F1 increasing by 3.12% and MUC F1 increasing by 4.51%. Furthermore, the inference speeds between the two approaches are relatively consistent, indicating that the utilization of labeled data augmentation does not have a significant impact on the model's inference speed. These comprehensive findings underline the advantages of incorporating labeled data augmentation in enhancing the accuracy of multi-modal coreference resolution without compromising the efficiency of the model's inference process. By leveraging additional labeled data, the model gains a better understanding of multi-modal coreference, leading to more accurate results while maintaining a comparable inference speed. In summary, the experiments clearly demonstrate the effectiveness of labeled data augmentation in improving model performance for

Ground Truth

in this image there is a drink in a glass and the glass is placed on the table. there is a person sitting on the chair and she is holding some objects in her hands. there is a kid sitting on her lap and she is holding a food item in her hand.

Semi-MCR

in this image there is a drink in a glass and the glass is placed on the table. there is a person sitting on the chair and she is holding some objects in her hands. there is a kid sitting on her lap and she is holding a food item in her hand.

Ours

in this image there is a drink in a glass and the glass is placed on the table. there is a person sitting on the chair and she is holding some objects in her hands. there is a kid sitting on her lap and she is holding a food item in her hand.

Ground Truth

in this image in the foreground there is one person who is wearing spectacles, and the person is holding a baby. and in the background it looks like there are chairs, shoe and it looks like there is one person sitting and there is a bag it seems and we could see floor.

Ours

in this image in the foreground there is one person who is wearing spectacles, and the person is holding a baby. and in the background it looks like there are chairs, shoe and it looks like there is one person sitting and there is a bag it seems and we could see floor.

multi-modal coreference resolution tasks. By leveraging additional labeled data, the model achieves superior results across evaluation metrics, while the inference speed remains consistent. This highlights the efficiency and benefits of incorporating labeled data augmentation in multi-modal coreference resolution models.

E THE EFFECT OF VAES

To assess the effectiveness of the multi-modal VAE module, we conduct experiments comparing the overall results of using VAE versus not using VAE on the CIN dataset, as shown in Table A3. Upon analyzing the results, we find that utilizing VAE significantly outperform the non-VAE approach. Specifically, there is a 1.06% improvement in MUC F1 and a 0.81% improvement in ConLL F1. These results indicate the effectiveness of the VAE module in capturing the inherent relationships and structures between each modality's features. By modeling text and image features separately as their

Table A3: Results of the comparison of the self-adaptive selection strategy with different fixed proportion selection strategies.

Training Set	MUC			B ³			CEAF _e			CoNLL
	R	P	F1	R	P	F1	R	P	F1	F1
w/o VAEs	39.64	42.99	40.37	72.41	89.18	79.83	65.47	96.21	77.49	65.90
with VAEs	39.83	44.70	41.43	72.42	90.27	80.32	66.51	96.63	78.38	66.71

respective latent representations, the VAE module reduces noise and uncertainty during the model’s prediction process. This enables the model to better understand the distinctive characteristics of each modality and model them in the latent representation space. This modeling approach contributes to enhancing the predictive performance and stability of the model. By learning the latent representations, the model can better capture the correlations between different modalities, resulting in more accurate predictions.

REFERENCES

[1] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2023. Semi-supervised multimodal coreference resolution in image narrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 11067–11081.

[2] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2023. Who are you referring to? Coreference resolution in image narrations. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*. IEEE, 15201–15212.

[3] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, Yoshua Bengio and Yann LeCun (Eds.).