Does Structural Attention Improve Compositional Representations in Vision-Language Models?

Rohan Pandey **Rulin Shao** **Paul Pu Liang** **Louis-Philippe Morency**
Language Technologies Institute
Carnegie Mellon University
{rspandey, rulins, pliang, morency}@cs.cmu.edu

Abstract

Although scaling self-supervised approaches has gained widespread success in Vision-Language pre-training, a number of works providing structural knowledge of visually-grounded semantics have recently shown incremental performance gains. Past work hypothesizes that providing structural knowledge to models in the form of scene graphs, syntax parses, etc. will result in better *Structure Alignment* and thus maintain representational compositionality, a core feature of human cognition. We compare one such *Structural Training* model to a *Structural Attention* model which has only implicitly learned inter-modal structure alignment through a self-supervised attention regularizer. We report that the latter model results in a 52% improvement over its baseline on the Winoground evaluation dataset, establishing a new vision-language compositionality state-of-the-art (Group=16.00). We begin exploring why this self-supervised approach succeeds where a more strongly supervised approach fails, specifically analyzing what the auxiliary loss implicitly conveys about structural knowledge.

1 Introduction

Much of the expressive power of human language derives from the principle of compositionality—the ability to combine meanings of constituents according to structured rules. Compositionality applies not only to language but to vision as well, with the ideal representation of an image capturing the relations of its subcomponents. However, recent work shows that Vision-Language Models (VLMs) fail to construct compositional representations and generally ignore syntactic & structural information [14, 12, 10].

Winoground [14] is a simple vision-language compositionality task that tests a VLM's ability to match syntactic permutations of a caption to their corresponding images; Thrush et al. [14] find that all recent state-of-the-art VLMs perform below chance levels on the Winoground evaluation dataset. Contemporaneously, Milewski et al. [12] probe for structural knowledge in VLMs, finding that they encode significantly less linguistic syntax than LMs and virtually no visual structure, treating images as a 'bag of objects'.

While there has been much work focusing on vision-language alignment of objects and words [8, 6, 4, 11], there has been less highly influential work on aligning structures and relations across vision & language. We tentatively group these Structure Alignment approaches into three main categories: *Structural Training* (providing a model with multimodal structural input or pre-training to predict structures) [23, 3, 15, 7, 18, 9, 24], *Structural Architecture* (e.g. graph neural nets that are dynamically constructed based on a multimodal structure) [4, 1, 16, 17, 26, 5], and *Structural Attention* (indirectly using structure to guide the attention mechanism) [13, 20, 21, 19].

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

We focused on comparing Structural Training and Structural Attention approaches since we were unable to locate publicly available Structural Architecture models for image-text matching. As a representative for Structural Training, we choose ROSITA [3] and for Structural Attention, we choose IAIS [13]. After discovering that IAIS not only outperforms ROSITA, but achieves a new state-of-the-art on Winoground, we explore how IAIS works with structurally complex data.

Why does an attention regularizer (IAIS loss) calculated from attention values with no explicit knowledge of multimodal structure result in more structural understanding than a model trained on scene graphs & textual relations (ROSITA)? This phenomenon is broadly reminiscent of the emergence of semantic analogies in word embedding spaces solely through co-occurrence training.

Our main contributions are twofold:

1. IAIS improves on its baseline (UNITER [2]) by 52% to beat the current Winoground state-of-the-art (VinVL [25]) across all 3 score types.
2. Correlation between IAIS loss and Winoground performance of a model may serve as a measure of whether a model represents intra-modal relations well.

2 Methods

We replicated two Structural Alignment models: ROSITA [3] and IAIS [13] based on their publicly available Flickr30k [22] fine-tuned checkpoints and tested both of them on the Winoground evaluation dataset. ROSITA takes a structured knowledge pre-training approach, where scene graphs and text relations are jointly masked & predicted. IAIS builds on UNITER by introducing IAIS loss to implicitly encourage inter-modal alignment between attention relations. In simple terms, it works by identifying the linguistic tokens that correspond to visual objects, and pulling together the attention values between token pairs and their corresponding object pairs.

To better understand the performance of these models, we calculate their IAIS loss on Winoground examples. Since IAIS loss approximates ISDa [13], it should inversely correlate with the model's ability to inter-modally align the intra-modal relations of an example. Winoground examples require high inter-modal alignment between these relations, and therefore we expect:

1. Models should exhibit higher IAIS loss on Winoground than on Flickr30k or COCO, since the former is more relationally complex.
2. A 'good' model should exhibit greater difference in IAIS loss between Winoground pairs and anti-pairs, since the model only 'understands' the correct alignment.

Intuitively, this latter proposition derives from the fact that IAIS loss should be higher for image-text pairs that have unmatched relations, and that a good model should be better at recognizing that a Winoground anti-pair is unmatched. While a good model should hopefully generalize this ability across Winoground, IAIS loss measures attention alignment for individual examples, meaning that we can expect this correlation to hold even if a model is only 'good' at an individual example. Formally,

$$Score_{WG}(\mathcal{M}, X_i) \sim L_{IAIS}(\mathcal{M}, X_{i,ap}) - L_{IAIS}(\mathcal{M}, X_{i,p}) \tag{1}$$

where $X_i = \{I_0, I_1, C_0, C_1\} \in \mathcal{D}_{Winoground}$ contains 2 images and 2 captions. $X_{i,p} = \{(I_0, C_0), (I_1, C_1)\}$ denotes the true paired samples in X_i, and $X_{i,ap} = \{(I_0, C_1), (I_1, C_0)\}$ denotes the anti-pairs in X_i. Eq. 1 states that the difference between pair and anti-pair IAIS loss on Winoground examples which a model $\mathcal{M}$ successfully matches should be higher than on those which the model unsuccessfully matches.

3 Results

In table 1 we report new scores for IAIS (built on UNITER$_{large}$) and ROSITA. Observe that IAIS fine-tuned on Flickr30k results in a new state-of-the-art Group score–a 10% improvement on the prior VinVL and an 52% improvement on UNITER$_{large}$, and comparable improvements for Text and Image scores. Although ROSITA underperforms VinVL, it would still achieve third place in Group score among the 19 models originally tested by Milewski et al. [12].

Model	Text	Image	Group
MTurk Human	89.50	88.50	85.50
Random Chance	25.00	25.00	16.67
IAIS$_{Flickr30k}$	**42.50**	**19.75**	**16.00**
IAIS$_{COCO}$	41.75	**19.75**	15.50
VinVL (OSCAR+)	37.75	17.75	14.50
ROSITA$_{Flickr30k}$	35.25	15.25	12.25
UNITER$_{large}$	38.00	14.00	10.50
CLIP (ViT-B/32)	30.75	10.50	8.00

Table 1: Winoground scores for IAIS and ROSITA compared with the previous state-of-the-art (VinVL) and IAIS' baseline (UNITER)

We should acknowledge that these scores do indeed continue to remain below the computed 'Random Chance' levels. However, improvements even below this threshold are meaningful because the models that we're testing also have near state-of-the-art performance on standard image-text matching tasks; the same may not be said of a random selector. Thus, even though IAIS remains below 'chance' levels, its considerable improvement on prior models across all 3 score types means that it acquires these improvements by genuinely learning better structural alignment, not simply by luck.

Next, let us address the two analysis hypotheses in Sec. 2 about IAIS loss applied to models on Winoground. We do indeed find that the average IAIS loss of UNITER models is higher on Winoground than what is reported for Flickr30k + COCO. Specifically, UNITER-base has a mean IAIS loss of 1.31 on Winoground, compared to 0.59 on the original dataset. This suggests that UNITER's attention fails to capture the complex relations of Winoground as easily as it does on traditional vision-language datasets. Intuitively, this serves as confirmation that IAIS loss correlates with a human notion of compositional complexity.

Since all 3 traditional Winoground scores are boolean for individual examples, we define a 'soft' notion of Score$_{WG}$, which is computed as a continuous analog of Group score. Concretely,

$$Score_{WG} = s(C_0, I_0) + s(C_1, I_1) - (s(C_0, I_1) + s(C_1, I_0)) \tag{2}$$

As shown in Fig. 1, the difference of IAIS loss between pairs and anti-pairs has a linear correlation with Score$_{WG}$. The correlation for UNITER$_{large}$ is strong and positive ($r = .82$), while the correlation is negative and weaker for UNITER$_{base}$ ($r = -.61$, though still certainly significant, $p = .00$). The fact that these correlations have different directions is curious. Intuitively, the positive correlation of UNITER$_{large}$ says that for Winoground examples which IAIS clearly distinguishes correct alignment, the model will successfully match them.

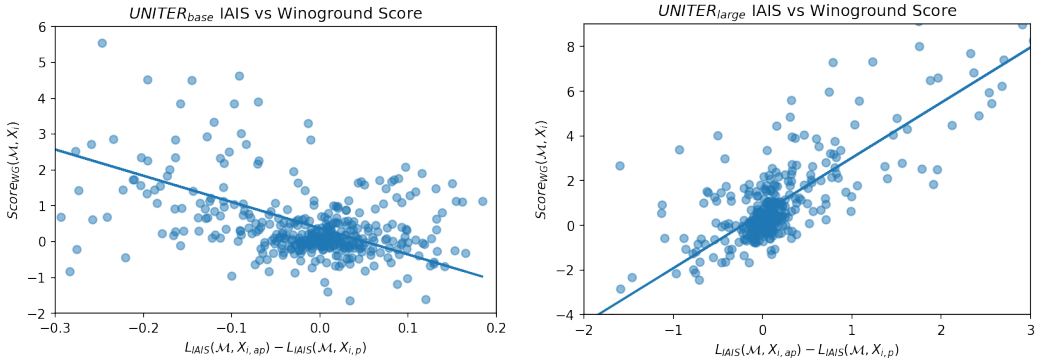

Figure 1: UNITER's difference in IAIS loss against its soft Winoground score

On the other hand, the negative correlation of UNITER$_{base}$ says that if IAIS fails to correctly identify attention alignment on a Winoground example, the model has a higher chance of matching them. Our working hypothesis to explain this strange relation is that as a model gets better, even if it isn't explicitly trained on IAIS loss, its behavior will become more predictable by IAIS since its attentions

are better aligned due to performance increase. In a bad vision-language model, the abstractions necessary to encourage relation alignment in final layer attention aren't even present yet, preventing IAIS from predicting results.

Ren et al. [13] notice that annealing IAIS loss such that it increases in weight towards the end of training improves performance of the resulting model. Further work may use the correlation metric we demonstrate to determine an optimal loss annealing schedule. In addition, it remains to be confirmed whether this correlation develops smoothly with model performance, scale, or something else.

At a high level, IAIS loss can only inter-modally compute relation alignment if relations are intra-modally represented already. The fact that UNITER$_{base}$ has a negative correlation suggests that it doesn't represent those intra-modal relations adequately yet, and thus isn't 'ready' for training on IAIS. Even though UNITER$_{base}$ achieves comparable performance to UNITER$_{large}$ on Flickr30k when fine-tuned on IAIS loss, we hypothesize for future work that the base model wouldn't generalize as well to Winoground as the large model, since IAIS doesn't have high enough quality relation representations to act upon.

4 Conclusion

In this work, we reported a new state-of-the-art on the Winoground task as achieved by a self-supervised Structural Attention approach, IAIS. We explored how IAIS loss acts upon the attention mechanism of a few models and how this may encourage structural alignment in the Winoground task. We briefly discussed why a more strongly supervised Structural Training model, ROSITA, doesn't achieve similar levels of performance despite training on data that intuitively seems relevant to Winoground.

More broadly, recent work in Vision-Language Modeling seems to have come to a fork, with those with access to computational resources focusing on performance gains through scaling self-supervised approaches, and those without focusing on models that leverage explicit inductive biases such as multimodal structure. This work confirms that much remains to be done in the middle with approaches like IAIS that are inspired by a structural inductive bias but highly scalable due to their self-supervised nature and lack of need for complex, structural data.

The longer term goal of this line of work is to explore how compositional vision-language representations may be learned, in which the structure of the world may be aligned with the structure of words as effectively as in the human mind. In this context, Winoground serves as an excellent benchmark for human notions of vision-language compositional ability while metrics like IAIS loss, ISDa, and Milewski score [12] serve as tools to probe model-internal structure representations. We seek to leverage representation probes like these to better understand what enables strong performance on compositionality benchmarks, contributing to the development of models of grounded meaning that better approximate human ability.

References

[1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[3] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021.

[4] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. Aligning linguistic words and visual semantic units for image captioning. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 765–773, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350943. URL `https://doi.org/10.1145/3343031.3350943`.

[5] Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. Vlgrammar: Grounded grammar induction of vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1665–1674, October 2021.

[6] Corentin Kervadec, Grigory Antipov, Moez Baccouche, and Christian Wolf. Weak supervision helps emergence of word-object alignment and improves vision-language tasks.

[7] Zaid Khan, Vijay Kumar BG, Xiang Yu, Samuel Schulter, Manmohan Chandraker, and Yun Fu. Single-stream multi-level alignment for vision-language pretraining. *arXiv preprint arXiv:2203.14395*, 2022.

[8] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[9] Zejun Li, Zhihao Fan, Huaixiao Tou, and Zhongyu Wei. Mvp: Multi-stage vision-language pre-training via multi-level semantic alignment. *arXiv preprint arXiv:2201.12596*, 2022.

[10] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

[11] Yongfei Liu, Chenfei Wu, Shao-yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*, 2021.

[12] Victor Milewski, Miryam de Lhoneux, and Marie Francine Moens. Finding structural knowledge in multimodal-bert. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5658–5671, 2022.

[13] Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. Learning relation alignment for calibrated cross-modal retrieval. 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 514–524, 2021.

[14] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[15] Bo Wan, Wenjuan Han, Zilong Zheng, and Tinne Tuytelaars. Unsupervised vision-language grammar induction with shared structure modeling. In *International Conference on Learning Representations*, 2021.

[16] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. *arXiv preprint arXiv:2205.11501*, 2022.

[17] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning. 2022.

[18] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. *Advances in Neural Information Processing Systems*, 34:4514–4528, 2021.

[20] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2197–2207, 2021.

[21] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9847–9857, June 2021.

[22] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[23] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.

[24] Junchao Zhang and Yuxin Peng. Hierarchical vision-language alignment for video captioning. In *International Conference on Multimedia Modeling*, pages 42–54. Springer, 2019.

[25] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[26] Xi Zhang, Feifei Zhang, and Changsheng Xu. Explicit cross-modal representation learning for visual commonsense reasoning. *IEEE Transactions on Multimedia*, 24:2986–2997, 2022. doi: 10.1109/TMM.2021.3091882.