Semantic Aligned Multi-modal Transformer for Vision-Language Understanding: A Preliminary Study on Visual QA

Han Ding* **Li Erran Li*** **Zhiting Hu** **Yi Xu** **Dilek Hakkani-Tur** **Zheng Du** **Belinda Zeng**
Alexa AI, Amazon
{handing, lilimam, huzhitin, yxaamzn, hakkanit, zhengdu, zengb}@amazon.com

Abstract

Recent vision-language understanding approaches adopt a multi-modal transformer pretraining and finetuning paradigm. Prior work learns representations of text tokens and visual features with cross-attention mechanisms and captures the alignment solely based on indirect signals. In this work, we propose to enhance the alignment mechanism by incorporating image scene graph structures as the bridge between the two modalities, and learning with new contrastive objectives. In our preliminary study on the challenging compositional visual question answering task, we show the proposed approach achieves improved results, demonstrating potentials to enhance vision-language understanding.

1 Introduction

Vision-language tasks, such as image captioning (Vinyals et al., 2015), visual question answering (Antol et al., 2015), and visual commonsense reasoning (Zellers et al., 2018), serve as rich test-beds for evaluating the reasoning capabilities of visually informed systems. These tasks require joint understanding of visual contents, language semantics, and cross-modal alignments. In particular, beyond simply detecting what objects are present, models have to understand comprehensively the semantic information in an image, such as objects, attributes, relationships, actions, and intentions, and how all of these are referred to in natural language.

Inspired by the success of BERT (Devlin et al., 2019) on a variety of NLP tasks, there has been a surge of building pretrained models for vision-language tasks, such as ViLBERT (Lu et al., 2019), VL-BERT (Su et al., 2020), and UNITER (Chen et al., 2020). Despite the impressive performance on several vision-language tasks, these models suffer from fundamental difficulties in learning effective visually grounded representations, as they rely solely on cross-attention mechanisms to capture the alignment between image and text features, and learn from indirect signals without any explicit supervisions. Recently, Oscar (Li et al., 2020) introduced object tags detected in images as anchor points to ease the learning of semantic alignments between image regions and word sequences. However, individual object tags in isolation ignore the rich visual information, such as attributes and relationships between objects. Without such information as contextual cues, the core challenge of ambiguity in visual grounding remains difficult to solve. As Figure 1 shows, in order to answer the question correctly, the model needs to reason about object relationships. Without the relation "on" between "cup" and "table", the model mistakenly thinks the "cup" is on the "tray".

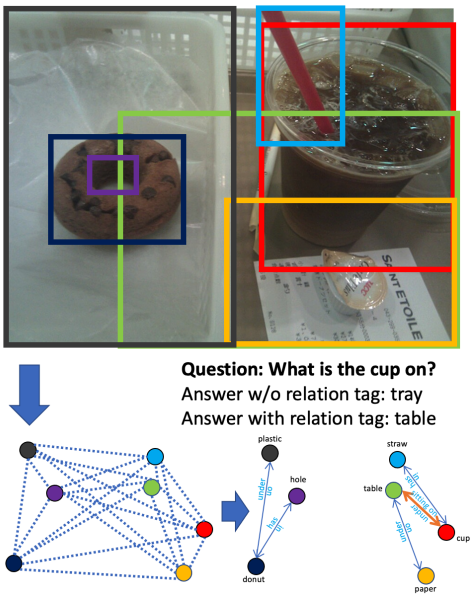

Figure 1: A Visual question-answering example illustrating the effectiveness of using scene graph as the bridge for cross-modal alignment

This work tackles the above challenges by introducing *visual scene graphs* as the bridge to align vision-language semantics. Extracted from the image using modern scene graph generators, a visual

*Equal contribution

Proceedings of the Third Workshop on Multimodal Artificial Intelligence, pages 74–78

scene graph effectively depicts salient objects and their relationships. The visually-grounded intermediate abstraction permits more effective vision language cross attention for disambiguation and finer-grained alignment. Specifically, we propose *Samformer* (Semantic Aligned Multi-modal transFORMER) that learns the alignment between the modalities of text, image, and graphical structure. For each of object-relation labels in the scene graph, the model can easily find the referring text segments in natural language, and then learn to align to the image regions already associated with the scene graph. On the basis of the visually-grounded graph, we apply a contrastive loss and a masked language model loss that explicitly encourage image-text alignment. Furthermore, we propose a per-triplet (object, relation, subject) contrastive loss to align object and relation representations across the two modalities respectively.

We adopt a set of datasets, including Microsoft COCO Captions dataset (Lin et al., 2014), Visual Genome (Krishna et al., 2016), VQA (Antol et al., 2015), GQA (Hudson and Manning, 2019), Flicker 30k (Young et al., 2014), SBU (Ordonez et al., 2011), and Conceptual Caption (Sharma et al., 2018) to pre-train our model and fine-tune it on visual compositional question answering (GQA) (Hudson and Manning, 2019). Our preliminary analyses show improved performance and demonstrate the potential of the proposed approach on broader visual-language applications.

2 Semantic Aligned Vision and Language Transformer

This section presents the proposed semantic aligned multi-modal transformer (Samformer) for vision-language pre-training. Figure 2 provides an overall architectural view of the method.

Given a pair of an image I and a text sequence $\boldsymbol{w}$ describing the image, the goal of vision-language pre-training is to learn a joint representation of the pair which captures the alignment between the words and image regions and can be adapted to assist downstream tasks. Same as the previous vision-language models (Li et al., 2020; Chen et al., 2020), the proposed Samformer first separately encodes each modality into singular embedding features, and then employs a multi-layer self-attention transformer to align the features and obtain a cross-modal contextualized representation.

Samformer differs critically from previous methods in that we incorporate the visual scene graph extracted from the image to enhance the cross-modal representation learning. The structured, visually-grounded graph encodes rich semantic information (e.g., objects, relationships), which, compared to isolated object tags (Li et al., 2020) and bare image text singular features (Chen et al., 2020; Lu et al., 2019; Su et al., 2020), offers valuable cues to resolve ambiguity and *bridge* together text and visual semantics. We describe in details how visual scene graph is integrated to interplay with the text and image modalities for better alignment (section 2.1), and on this basis how contrastive learning strategies are devised for fine-grained alignment supervisions (section 2.2).

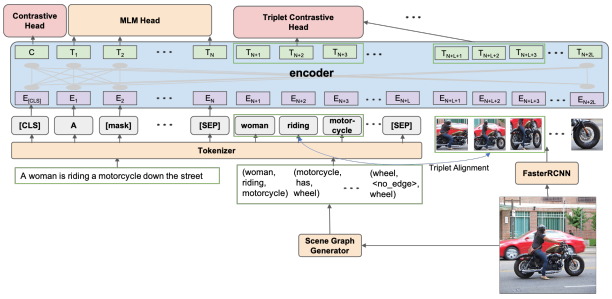

Figure 2: Architecture of the proposed Samformer.

2.1 Cross-modal Alignment with Visual Scene Graph Encoding

Given an image-text pair $(I, \boldsymbol{w})$, we first extract the visual scene graph G from the image with an off-the-shelf scene graph generator (Tang et al., 2020). A scene graph is a directed graph with the nodes representing the objects and the edges depicting their pairwise relationships. We represent the graph as a set of triplets, where a triplet (o_i, r_{ij}, o_j) denotes the relation type r_{ij} between object o_i and object o_j, e.g., ("woman", "riding", "motorcycle") in Figure 2. Crucially, the scene graph is already visually grounded. That is, each of the components in the triplets is associated with the corresponding regions in the image. For example, the object "woman" is associated with the bounding box of woman while the relationship "riding" corresponds to the bounding box that contains both the woman and the motorcycle. With such aligned object/relationship tokens and image regions, the visual scene graph thus serves as a bridge between the original image I and text sequence $\boldsymbol{w}$. That is, the model can easily find the correspondence between the text segments in the sequence $\boldsymbol{w}$ and the triplet tokens in the scene graph, since both are in the text modality. The text segments are then

naturally aligned with the respective image regions associated with the scene graph. More importantly, the triplets containing both object and relationship information provide the model with ample contextual cues to resolve ambiguity. For example, Figure 1 shows the relationship "on" between "cup" and "table" resolves the ambiguity whether the cup is on the table or the tray.

In implementation, we first embed tokens in both the text sequence $\boldsymbol{w}$ and scene graph triplets (extracted by SGG (Tang et al., 2020)) with a pre-trained BERT embedder (Devlin et al., 2019). We then extract the visual embedding of each image region and also the union region of each triplet with the Faster R-CNN component (Ren et al., 2015) used in the bottom-up-attention (Anderson et al., 2018). All the embedding vectors are then fed into a transformer network with self-attention mechanisms to infer the alignment, as shown in Figure 2. In particular, to inform the transformer about the known alignment between the scene graph triplet tokens and image regions, we augment each triplet embedding and its corresponding image region embedding with the same position embedding.

2.2 Pre-training

We describe the pre-training method of the model. After pre-training, the model can then be applied to downstream visual-language tasks with efficient finetuning.

2.2.1 Masked Language Modeling (MLM)

This task is very similar to the Masked Language Modeling (MLM) task utilized in BERT (Devlin et al., 2019). The key difference is that visual clues are incorporated to predict the masked words for capturing the dependencies among visual and linguistic contents. During pre-training, each word in the input sentence is randomly masked (at a probability of 15%). For the masked word w_m, its token is replaced with a special token `[MASK]`. The model is trained to predict the masked words, based on the unmasked words $\boldsymbol{w}_{\setminus m}$, the scene graph G, and the visual features $\boldsymbol{v}$ of image regions (Figure 2). During pre-training, the final output feature at the position of the masked word is fed into a classifier over the whole vocabulary, and we minimize the prediction loss:

$$\mathcal{L}_{\mathrm{MLM}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{w},m} \log P_\theta \left(\boldsymbol{w}_m \mid \boldsymbol{w}_{\setminus m}, G, \boldsymbol{v}\right) \quad (1)$$

The MLM task learns to use the relevant tokens in triplet tags which effectively aligns the representation between text $\boldsymbol{w}$ and graph G.

2.2.2 Contrastive Losses for Cross-Modal Alignment

As shown in Subsection 2.1, our model aligns the scene graph of an image with paired text using triplet tags as the bridge. We use two Contrastive loss terms. One is at the sequence level to align G and $\boldsymbol{v}$. The other is to align each triplet tag and its region features. For each training example, we randomly decide whether to use the first term or second. As training progresses, we increase the probability of using the first term. The reason to use the sequence level loss is because many downstream visual-language problems directly finetune the sequence level representation.

Specifically, given an image, we sample a object-relation triplet g from its scene graph G. We then replace the scene graph G by G' randomly sampled from the entire dataset with probability 50%. Denote H the resulting scene graph. We apply a fully-connected (FC) layer as a binary classifier on top of the encoder output of `[CLS]` to predict whether the scene graph is original (y = 1 if $H = G$) or has been replaced (y=0 if $H = G'$). The cross-modal contrastive loss at a global level (CMCG) is defined as:

$$\mathcal{L}_{\mathrm{CMCG}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{w},G} \log P_\theta \left(y \mid \boldsymbol{w}, H, \boldsymbol{v}\right) \quad (2)$$

The second contrastive loss at the triplet tag level is constructed as follows. For each triplet tag g, we randomly determine with probability 50% whether we replace with another tag, g'. We apply a fully-connected (FC) layer as a binary classifier on top of the encoder output of g' and its region features to predict whether the tag is original (z = 1) or has been polluted (z = 0). The cross-modal contrastive loss for each triplet tag (CMCT) is defined as:

$$\mathcal{L}_{\mathrm{CMCT}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{w},g} \log P_\theta \left(z \mid \boldsymbol{w}, G_{\setminus g}, g', \boldsymbol{v}\right) \quad (3)$$

3 Preliminary Experiments

3.1 Experimental Settings

We initialize our model with Oscar (Li et al., 2020) base model weights and pre-train it further on the collected image-text corpus. The scene graph used in our model is extracted using the pretrained model of SGG (Tang et al., 2020).

After pre-training, we conduct our preliminary experiments on GQA (Hudson and Manning, 2019). The task focuses on visual reasoning and compositional question answering in real-world settings

Method	Test-dev	Test-std
Oscar	58.40	59.01
Samformer w/ CMCG	60.46	60.33
Samformer w/ CMCG+CMCT	**60.51**	**60.62**
Improvement	2.11	1.61

Table 1: Comparison of Samformer and Oscar on GQA test sets (fine-tuned on train-bal only).

Method	Test-dev-open	Test-std-open
Oscar	42.27	43.32
Samformer w/ CMCG	**46.36**	**46.77**
Samformer w/ CMCG+CMCT	45.88	46.26
Improvement	4.09	3.45

Table 2: Comparison of Samformer and Oscar on GQA open questions (fine-tuned on train-bal only).

which involve diverse reasoning skills including spatial reasoning, relational reasoning, logic and comparisons. The task is formulated as a classification problem that chooses an answer from a shared set of 1,852 candidate answers. We select the particular task in our preliminary study because the task would benefit from effective alignment of the text-vision modalities on objects, relationships and attributes. In particular, GQA needs rich scene graph information from images to answer challenging compositional questions.

Since we build our model upon Oscar, we use it as the baseline for comparison. We choose Oscar base which has 12 layers and each layer has 12 attention heads. Both Oscar and ours were fine-tuned on the GQA train-balance dataset. For Oscar, we reproduced it with the official published pretrain model on the smaller balance training set.

3.2 Results

In this section, we study the performance on the downstream GQA task. As shown in Table 1, our Samformer by incorporating scene graphs improves the accuracy by 2.11% on GQA test-dev and 1.61% on test-std. The improvement is stronger if we focus on the challenging open questions (non-binary) in GQA, as shown in Table 2. Specifically, our method achieves 4.09% and 3.45% improvement, respectively, suggesting that including scene graph triplets help with understanding complex scenes and questions. The per-triplet contrastive loss, CMCT further improves the gains.

For fine-grained analysis of our method, we evaluate the performance grouping by the semantic type on the validation set. Among 5 semantic types, our method achieves 3.84% improvement on category type and 2.33% relation type. Although category questions are not directly asking about relation, the question itself sometimes related to a relation, for instance "Who is walking?". A triplet tag such as "man walking on street" would help the model better answer question like this.

To understand the full potential of the proposed approach that makes use of scene graphs, we evaluate the performance of Samformer when *ground-truth* scene graph is available. Figure 3 shows the results of evaluation accuracy as training proceeds. By including ground-truth scene graph relation tags, we can see significantly improved results compared to the baseline model that does not use relation tags at all. Using predicted relation tags also helps, though the improvement margin is more narrow since the predicted tags can be noisy.

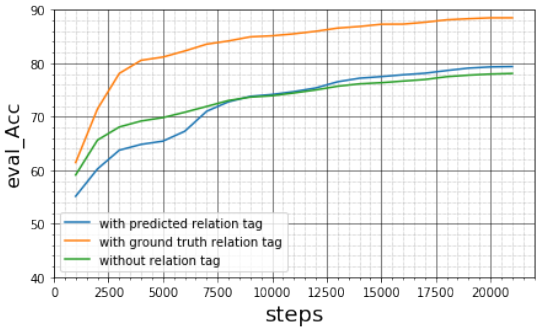

Figure 3: GQA evaluation accuracy curve without relation tags, with predicted relation tags, and with ground-truth relation tags.

4 Conclusion and Future Work

In this work, we propose Samformer, a novel semantic aligned multi-modal transformer model for vision-language pre-training. We explicitly align the visual scene graphs and text using triplet tags as anchors as well as a contrastive loss between each triplet tags and its paired visual features. We show improved preliminary results on GQA.

As shown in the empirical study, the performance is to some extent capped by the rather limited relations and object categories that can be extracted from off-the-shelf pre-trained scene graph models and object detectors. For future work, we plan to jointly train with scene graph models to more effectively learn from limited labeled data and weak supervision signals from paired text.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision (ICCV)*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt representation learning. In *European conference on computer vision (ECCV),*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal of Computer Vision (IJCV)*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics aligned pre-training for Vision-Language tasks. In *European conference on computer vision (ECCV)*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic visiolinguistic representations for Vision-and-Language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic Visual-Linguistic representations. In *International Conference on Learning Representations (ICLR)*.

Kaihua Tang, Yulei Niu, Jianqiang Huang, and Jiaxin Shi. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.