Contrastive Visual and Language Learning for Visual Relationship Detection

Thanh Tran, Maëlic Neau, Paulo E. Santos, David Powers

Contrastive Visual and Language Learning for Visual Relationship Detection

Thanh Tran and **Maëlic Neau** and **Paulo E. Santos** and **David Powers**
{tran0553, neau001, paulo.santos, david.powers}@flinders.edu.au
College of Science and Engineering
Flinders University
1284 South Rd, Clovelly Park,
SA5042 Australia

Abstract

Visual Relationship Detection (VRD) aims to understand real-world objects' interactions by grounding visual concepts to compositional visual relation triples, written in the form of *(subject, predicate, object)*. Previous work explored the use of contrastive learning to implicitly predict *predicates* (representing relations) from the relevant image regions. However, these models often directly leverage in-distribution spatial and language co-occurrences biases during training, preventing the models from generalizing to out-of-distribution compositions. In this work, we examined whether contrastive vision and language models, pre-trained on large-scale external image and text datasets, can assist the detection of compositional visual relations. To this end, we propose a contrastive fine-tuning approach for the VRD task. The results obtained from this investigation show that larger models yield better performance when compared with their smaller counterparts, while models pre-trained on larger datasets do not necessarily present the best performance.

1 Introduction

Understanding the visual world is essential for many modern computer vision tasks, including visual question answering (VQA) (Agrawal et al., 2016), image captioning (Hossain et al., 2019), and human-robot interaction (Goodrich and Schultz, 2007). Given the complexity of real world scenes, low-level object classification and recognition tasks are insufficient to solve these problems. Instead, higher-level visual understanding and reasoning tasks are often required. Visual Relationship Detection (VRD) aims to facilitate such understanding by bridging the gap between low-level visual information and high-level symbolic visual relation. Previous work (Lu et al., 2016) on VRD emphasized the use of spatial priors in the form of overlapping bounding boxes and language co-occurrence bias during training to achieve higher benchmark results. However, such explicit incorporation of priors during training often leads to biases that might not transfer to new visual relationship (Zhu et al., 2022).

To address these issues, current research leverage external linguistic commonsense knowledge from structured knowledge bases (Zareian et al., 2020) or unstructured natural language corpora (Ye and Kovashka, 2021) by distilling the information down into low-dimensional distributed vector embeddings. These embeddings indirectly benefit the visual relationship detection task as they preserve the relevant topological structure from the knowledge base or the linguistic contextual information from raw language corpora. In this work, we focus on the later and examine whether contrastive embedding models learned from the abundant amount of unstructured image and text pairs from the web (see Table 1) can contribute to the VRD task.

To this end, we examined the promise of robustness in the large pre-trained image and text deep neural encoders (Radford et al., 2021), focusing on learning joint visual and language embeddings for VRD. The present paper investigates the application of contrastive learning in this context. Contrastive learning aims to learn representations by *pulling together* the matching (or positive) vector pairs, while *pushing apart* non-matching (or negative) vector pairs. We believe that such a contrastive approach, when paired with learning (Oliver et al., 2018), will improve the joint representations and indirectly benefit the VRD task. Thus, this work proposes a fine-tuning framework that extends existing vision and text encoders to classify predicates from the given ground truth object regions and object labels. The results obtained show that the amount of data used in the pre-training process do not necessarily impact the final performance of the VRD predicate classification (VRD-PredCls)

Dataset	No. unique image and pairs
LAION400M	413 millions
LAION2B-en	2.32 billions

Table 1: Number of unique image-text pairs for two different datasets used in pre-training.

task when evaluated on models of the same size.

2 Related Work

This section presents a review of the work related to compositional grounding of visual concepts on language (Krishna et al., 2017), with an emphasis on visual relationship detection through the use of contrastive learning.

Visual Relationship Detection aims to construct a symbolic representation from an unstructured scene by representing it as a set of visual relationship triples in the form of *(subject, predicate, object)* (e.g *(person, riding, horse)*). Here, the *subject* and *object* are labels grounded to the salient image regions through bounding boxes, and the *predicate* is defined as a real-world interaction between these object pairs. However, due to the large number of potential real-world interactions, existing visual relation datasets including VRD (Lu et al., 2016) and Visual Genome (Krishna et al., 2017) are often sparse and unbalanced, where common relations occur more frequently than rarer but plausible ones. To tackle this problem, existing work seek to incorporate *linguistic knowledge* as additional training features to the learning model. For example, Lu et al. (2016) have shown that leveraging pre-trained word embeddings can help the models learn linguistic statistical priors; similarly, Yu et al. (2017) show that distilling knowledge from external Wikipedia datasets can improve the model's performance. Other work have also directly targeted the bias nature of the benchmark by incorporating spatial information (Peyre et al., 2019), and *(subject, object)* co-occurrence priors as learning features for the classifiers (Chen et al., 2019), improving the model's performance significantly. Recent approaches have also tackled the effect of bias by using model-agnostic counterfactual prediction during inference such as Total Direct Effect (TDE) (Tang et al., 2020) or by experimenting with different sampling strategies (Desai et al., 2021). In contrast, the present work does not use any predefined co-occurrence statistics, or spatial information, as learning features at the same time that it does not learn a neural discriminative classifier that directly predicts the predicate label from the input features. Instead, we apply existing contrastive encoder-encoder architecture to construct visual and text embeddings that can be ranked using a cosine similarity scoring metric.

Datasets: Visual Genome (Krishna et al., 2017) is the most used dataset for VRD. It is composed of 108K images that have been labeled with 150 object and 50 predicate classes. The labeling strategy of Visual Genome is simple: first, volunteers were asked to provide captions for regions in images; then, these captions were transformed into *(subject, predicate, object)* triples. This strategy allowed to build a large-scale dataset with an average of 21 relations per image but at the cost of introducing various biases such as asymmetric relations (e.g. *(hair, on, man)* versus *(man, has, hair)*) or confusion between predicates (e.g. *(man, wearing, shirt)* and (man, wears, shirt)). Visual Genome and other datasets used for Visual Relationships Detection such as VRD (Lu et al., 2016), suffer from a long-tail distribution of the predicates, as recurrently stated in previous work (Zellers et al., 2018; Tang et al., 2020; Chen et al., 2019). Nonetheless, Visual Genome has become the most used benchmark for VRD, mainly because of the large scale and diversity of its annotated data.

Contrastive Learning aims to minimize a defined distance metric between the matching or positive embedding vector pairs while maximizing the distance between non-matching or negative embedding vector pairs. Recent work on contrastive learning have shown that such discriminative learning approaches can (i) learn to ignore invariant features and spurious correlations through data augmentation and automatic negative sampling technique (Chen et al., 2020a), and (ii) learn joint visual and language embeddings that can be used to perform zero-shot triple detection on a wide variety of tasks (Peyre et al., 2019; Tran et al., 2022). In this work, we focus on use case (ii) which aims to transfer pre-trained image and text embeddings to the visual relationship detection task. To this end, we built on top of an existing body of work on zero-shot transfer and multi-modal representational learning, with an emphasis on CLIP (Radford et al., 2021), a contrastive learning model that encodes image and text using abundant (image, caption) pairs from the internet. We believe that such contrastive encoder-

encoder model gives a clearer separation of the visual embeddings and language embeddings compared to the traditional black-box neural fusion approaches (Su et al., 2019; Chen et al., 2020b), giving us more control over both the triples input and the final output embedding spaces.

Multi-modal Representational Pre-training. Visual relations consist of both textual tokens from the *(subject, predicate, object)* triples and the visual information from the objects' salient features and bounding boxes. As a result, it is essential to leverage pre-trained embeddings that capture not only the uni-modal information from image or text but also the interactions between both modalities. Still, most existing approaches use uni-modal architecture such as Faster-RCNN (Ren et al., 2015) or BERT (Devlin et al., 2018) that were pre-trained on uni-modal tasks such as object detection, object recognition, masked language modeling, or next sentence prediction, etc. In this work, we focus on applying encoder models that are pre-trained on both language and vision tasks, where each modality can contribute positively to the other, facilitating a more complete set of concepts. More specifically, we apply vision Transformers and language Transformers that were pre-trained using the discriminative contrastive learning framework (Radford et al., 2021). For fair comparisons with previous work and baseline experiments, we also evaluated our approach against CNN vision encoders and language Transformer encoders pre-trained on the contrastive learning framework.

3 Contrastive Image and Text Matching

This section describes the proposed architecture and outlines the details of the current implementation. The general architecture consists of three modules: (1) the **Visual Module** generates visual embeddings based on the extracted features from the *(subject, object)* union image regions, (2) the **Language Module** generates text embeddings of the concatenated *(subject, predicate, object)* string, and (3) the **Contrastive Loss Module** that consists of the visual-language contrastive losses ensuring the consistency between the matching (visual, language) embedding pairs, while pushing apart the non-matching (visual, language) embedding pairs. Here, we use the cosine similarity metric as the distance metric and the cross entropy loss as our main loss objective.

model	image encoder	no. params
CLIP	ResNet50	25.6M
CLIP	ResNet101	44.5M
OpenCLIP	VIT-B-16	86M
OpenCLIP	VIT-B-32	86M
OpenCLIP	VIT-L-14	307M
OpenCLIP	VIT-H-14	632M

Table 2: Number of parameters used in CLIP and OpenCLIP pre-trained image encoder.

3.1 Visual Module

Visual Encoder: One of the main sub-tasks of visual relationship detection is to detect subjects and objects from a given image and extract their visual features. Given the success of CNN-based architecture and Transformer-based architecture in learning image representations from large-scale datasets, we applied two different types of pre-trained backbone as our encoders: (i) the ResNet50 CNN-based visual backbone, and (ii) the ViT Transformer-based image backbone. We used the OpenCLIP pre-trained image encoder and language encoders (section 3.2) on either the LAION 400m or LAION 2b-en datasets to evaluate the impact of scaling up the size of the model and dataset on the final performance of the VRD task. OpenCLIP is the open-source version of CLIP with released pre-trained models. A comparison of CLIP and OpenCLIP image encoder is displayed in Table 2.

Image Preprocessing: Given the ground truth bounding boxes from an input image, we enumerated all n possible union bounding boxes and extracted $I_{i\in(0,n)}$ embedding vectors using one of the image encoders. It is worth noting that most of the encoder parameters were frozen during the fine-tuning process due to the limited resource setting.

3.2 Language Module

Language Encoder: Similar to the Vision Transformer (ViT) used in the Visual Encoder, the language encoder used in this work is a 12-layer 512-wide Transformer architecture (Vaswani et al., 2017) with 8 attention heads that can leverage the contextualized information from the entire input sentence. We believe that such transformer-based contextualized encoders are beneficial for the visual relationship detection task because the same predicate can have different meanings under distinct *(subject, object)* pairs.

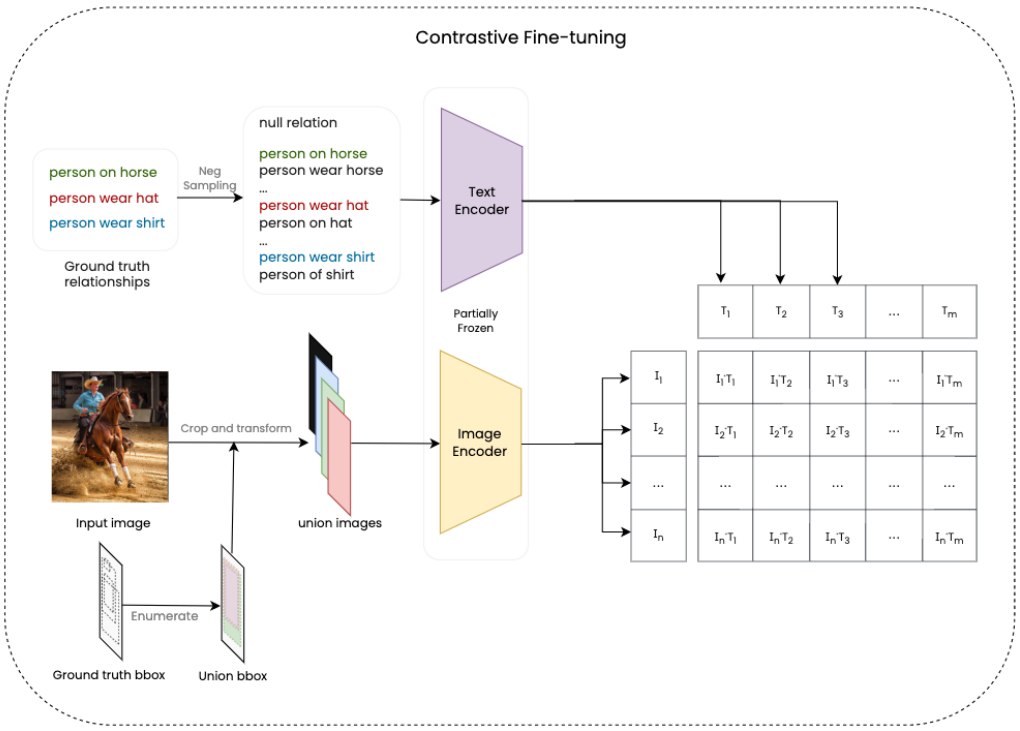

Figure 1: Overview of the proposed contrastive fine-tuning approach. Visualization is partially adapted from Radford et al. (2021).

Language Preprocessing: For each triple in the set of k ground truth triples, we first enumerated $p\,(subject, predicate_{i\in(0,p)}, object)$ triples where p is the number of predicates. We then concatenated each of these triples into '$subject\ predicate\ object$' string format, resulting in $m = k * p$ sentences. We also added a 'null relation' string as a matched pair with union image regions or (subject, object) pairs that have no visual relationship.

3.3 Contrastive Loss Module

Inspired by (Radford et al., 2021), we also used a dual cross entropy loss function for our contrastive visual and language consistency loss. Here, cosine similarity was used as a distance metric d for the loss function.

Negative Sampling: We performed negative sampling during the image and language enumeration and preprocessing step, where embeddings of each modality were paired against multiple negative embeddings of the other modality. This process resulted in a non-symmetric table as shown in Figure 1, where the row represents the n image embedding vectors and the column represents the m text embedding vectors. Using this table mask of positive and negative examples, we constructed the labels y for both image and text.

Visual and Language Consistency Loss: Similarly to the CLIP Loss, we computed a cross entropy loss along the table's rows and a cross entropy loss along the table's columns. Thus, given the set S_v containing all possible image embeddings I and the set S_l containing all possible text embeddings T, the cross entropy loss equation for all visual embeddings is:

$$L^{v-CE} = \frac{1}{|S_v|} \sum_{I \in S_v} \sum_{T \in S_l} -1_{y_I = y_T} \cdot \log\left(\frac{e^{d(I,T)}}{\sum_{T \in S_l} e^{d(I,T)}}\right)$$

Similarly, the cross entropy loss equation for all language embeddings is:

$$L^{l-CE} = \frac{1}{|S_l|} \sum_{T \in S_l} \sum_{I \in S_v} -1_{y_I = y_T} \cdot \log\left(\frac{e^{d(I,T)}}{\sum_{I \in S_v} e^{d(I,T)}}\right)$$

where,

$$1_{y_I = y_T} = \begin{cases} 1, & \textit{if (I,T) is positive} \\ 0, & \textit{otherwise} \end{cases}$$

The final visual and language consistency loss can be defined as:

$$L^{vl-CE} = L^{v-CE} + L^{l-CE}$$

[1]The relationship detection head of these models was retrained using the settings provided in (Han et al., 2021).

[2]We were unable to train the model and replicate the results

Model	Visual Encoder	Finetuned MLP	Finetuned Attention
CLIP-ResNet50-laion400m	ResNet50	53.2	-
CLIP-ResNet101-laion400m	ResNet101	47.7	-
OpenCLIP-VIT-B/16-laion400m	VIT-B/16	42.2	59.1
OpenCLIP-VIT-B/32-laion400m	VIT-B/32	**57.0**	59.8
OpenCLIP-VIT-B/32-laion2b-en	VIT-B/32	50.35	60.0
OpenCLIP-VIT-L/14-laion400m	VIT-L/14	55.6	61.8
OpenCLIP-VIT-L/14-laion2b-en	VIT-L/14	54.8	62.0
OpenCLIP-VIT-H/14-laion2b-en	VIT-H/14	51.4	**63.1**

Table 3: Predicate Prediction Results using the Recall metrics for different contrastive models on the Visual Genome dataset. Here, we examined two different strategies for fine-tuning these models: (i) *Finetuned MLP* and (ii) *Finetuned Attention*. In the approach (i), MLPs were attached on top of the frozen visual and language encoders, whereas in the approach (ii), the last two encoder layers of the Transformer were unfrozen.

Model	Visual Encoder	PredCls
IMP[1] (Xu et al., 2017)	ResNet50	57.6
MSDN[1] (Li et al., 2017)	ResNet50	59.6
G-RCNN[1] (Yang et al., 2018)	ResNet50	59.94
RelDN[2] (Zhang et al., 2019)	ResNet50	60.9
Neural Motif[1] (Zellers et al., 2018)	ResNet50	63.0
GPS-Net [2] (Lin et al., 2020)	ResNet50	66.9
ITS+RTS [2] (Tian et al., 2021)	ResNet101	**67.3**
OpenCLIP-VIT-H/14-laion2b-en	VIT-H/14	63.1

Table 4: Comparing Predicate Prediction Results on Visual Genome dataset with other work using the recall metrics on the PredCls task. The replicated results are slightly different from those in Han et al. (2021)

Test-Time inference: At test time, given a set of ground truth *(subject, object)* class pairs and a set of objects' bounding box regions, the evaluation algorithm first enumerates all possible *(subject, predicate, object)* triples and union image regions. These relation triples are then preprocessed into '$subject\ predicate\ object$' sentences as described in Section 3.2, yielding p textual embeddings for each pair, where p is the number of predicates. Similarly, the union bounding boxes and the image were preprocessed and encoded according to the method described in Section 3.1, yielding n possible embeddings.

For each image embedding $I_{i\in(1,n)}$, the evaluation code measured the similarity scores between the given image embedding and all possible corresponding textual embeddings in $T_{j\in(1,p)}$. The evaluation algorithm then ranks $n * p$ scores and selects the top result to compare against the ground truth using the Recall metrics based on Han et al. (2021).

in these references. Instead, we used the results provided by either Han et al. (2021) or the original paper referenced in the table.

4 Results

This section presents an evaluation of the performance of different fine-tuned OpenCLIP (Wortsman et al., 2022) models on the Visual Genome dataset, from which 60,784 images were used for training and 26,466 images for testing. Results marked with [1] have been computed by retraining the models using the PyTorch implementation from (Han et al., 2021). Since the commonly used Recall metric gives a biased view toward common relationships, we also present a visualization of how models of different sizes perform on individual predicate classes in Figure 2.

4.1 Evaluation

All evaluation results were computed using the recall metric on the top 1 ranked item for the predicate classification task (PredCLS). The Recall metric was used here to facilitate a comparison with previous work. Table 4 shows that the largest fine-tuned model achieved competitive results with respect to state-of-the-art approaches. It should also be noted that the majority of the other approaches use debiasing or balancing techniques to counteract

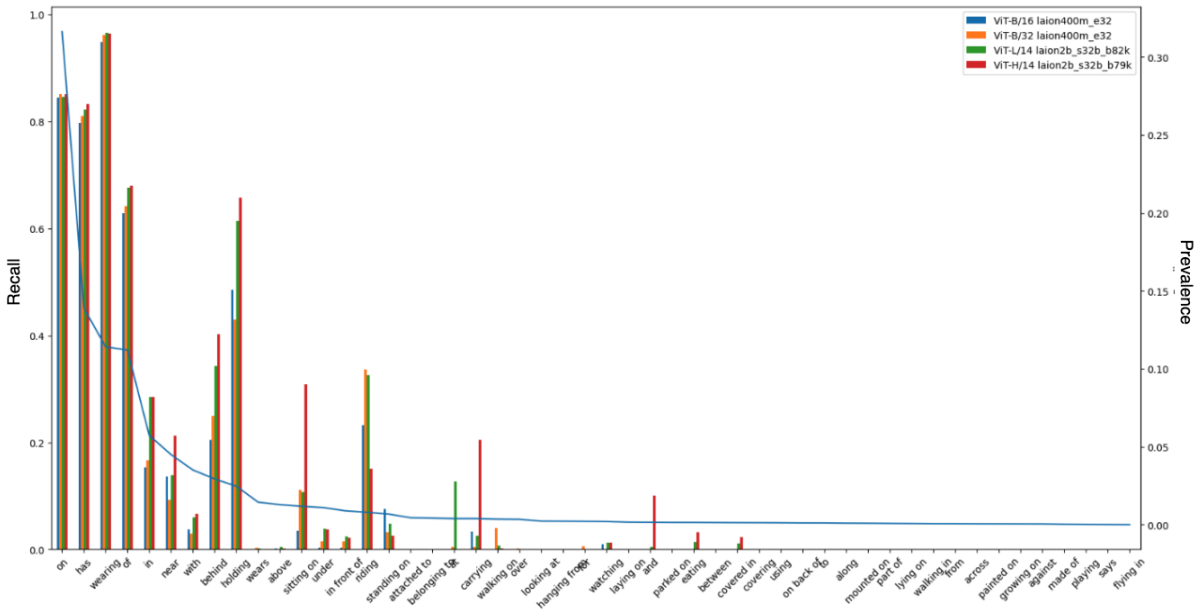

Figure 2: Predicate Prediction Results on Visual Genome dataset by individual relationship class. The blue line represents the prevalence of the class labels.

the long-tail nature of the training dataset. Table 3 indicates that fine-tuning these Transformer-based models without unfreezing the attention layers did not lead to better performance as the model size was increased. On the contrary, when fine tuned with unfrozen attention layers, we observed a consistent improvement in the model's performance as the model size increases. Here, *ViT-B/32* is smaller than *ViT-L/14*, and *ViT-L/14* is smaller than *ViT-H/14*. From Table 3, no significant improvement in performance was observed between models pre-trained on *laion400m* dataset and those pre-trained on the larger *laion2b-en* dataset. For Transformer-based models, an input image can be split into either 16 patches (*ViT-B/16*) or 32 patches (*ViT-B/32*). By comparing the *ViT-B/16* encoder with the*ViT-B/32* encoder trained on the *laion400m*, we observe that having more patches can improve the performance in both fine-tuning approaches.

4.2 Ablation Studies

To better analyze the performance of the different model sizes on rare predicates classification, we compared the distribution of each predicate category with the recall performance obtained by the 4 OpenCLIP models. Figure 2 shows that the largest model VIT-H/14 trained on laion2b-en performed better on rare predicates classes such as $carrying$, $sitting\,on$ or $eating$ while maintaining competitive performance on more common predicates such as on, has or $wearing$. Still, these models were biased towards common relationships and underperformed when evaluated on rare relationships. It is also unknown how much of the performance was due to a direct causal effect rather than chance using solely the Recall metric. Thus, future work should incorporate other evaluation metrics to have a more holistic measure of the model's performance.

5 Conclusion

Visual Relationship Detection is the cornerstone of many modern machine learning tasks that require a comprehensive understanding of the visual scene. In this work, we investigated whether large neural encoders pre-trained on a large set of data could capture the natural web image-text pair distribution (Radford et al., 2021) and assist the detection of visual relations. To this end, we proposed a contrastive fine-tuning technique that leverages this capability to perform Visual Relationship Detection. While the results in Table 4 show that the fine-tuned model achieved competitive results, the model still suffered from biases that stemmed from the long-tailed data distribution.

Moreover, based on results of the tests reported in this work, the success of contrastive learning techniques was highly dependent on the quality of the positive and negative sampling examples, and there was a limit to what automated sampling tech-

niques can do without human intervention. Without high-quality samples, these deep neural networks still learned spurious correlations and patterns, making it difficult for the models to generalize beyond the given domain. Thus, future work may further explore better negative sampling and data augmentation techniques that incorporate external taxonomies or knowledge bases to avoid biases and spurious correlations stemming from the skewed training data distribution.

Finally, the test-time inference algorithm in this work becomes more expensive as the number of predicates p increases since the evaluation technique used measures the distance between the given image embedding and all textual embeddings $(subject, predicate_{i \in (1,p)}, object)$. However, this effect can be mitigated by pre-computing all sentence embeddings from the dataset and storing them in memory or in the file system. Still, such approach may suffer from the I/O bottleneck where the embeddings have to be loaded into memory prior to computation. Future work may also explore alternative strategies to leverage the generated embeddings in predicate prediction tasks.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. VQA: Visual Question Answering. *arXiv:1505.00468 [cs]*.

Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. 2019. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR.

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

Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. 2021. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15404–15413.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Michael A. Goodrich and Alan C. Schultz. 2007. Human-robot interaction: A survey. *Found. Trends Hum.-Comput. Interact.*, 1(3):203–275.

Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. 2021. Image scene graph generation (sgg) benchmark.

MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys*, 51(6):1–36.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision*, pages 1261–1270.

Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. 2020. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual Relationship Detection with Language Priors. *arXiv:1608.00187 [cs]*.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.

Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2019. Detecting Unseen Visual Relations Using Analogies. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1981–1990, Seoul, Korea (South). IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725.

Hongshuo Tian, Ning Xu, An-An Liu, Chenggang Yan, Zhendong Mao, Quan Zhang, and Yongdong Zhang. 2021. Mask and predict: Multi-step reasoning for scene graph generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4128–4136.

Thanh Tran, Paulo E. Santos, and David Powers. 2022. Contrastive Visual and Language Translational Embeddings for Visual Relationship Detection: AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence, AAAI-MAKE 2022. *CEUR Workshop Proceedings*, 3121.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971.

Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685.

Keren Ye and Adriana Kovashka. 2021. Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8289–8299.

Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. 2017. Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1068–1076, Venice. IEEE.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. 2020. Bridging knowledge graphs to generate scene graphs. In *European conference on computer vision*, pages 606–623. Springer.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840.

Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. 2019. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543.

Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, and Mohammed Bennamoun. 2022. Scene Graph Generation: A Comprehensive Survey. ArXiv:2201.00443 [cs].