Contrastive Visual and Language Translational Embeddings for Visual Relationship Detection

Thanh Tran[1], Paulo E. Santos[1] and David Powers[1]

[1]*College of Science and Engineering, Flinders University of South Australia,*
1284 South Rd, Clovelly Park SA 5042, Australia

Abstract

Visual relationship detection aims to understand real-world interactions between object pairs by detecting visual relation triples written in the form of *(subject, predicate, object)*. Previous work has explored the use of contrastive learning to generate joint visual and language embeddings that aid the detection of both seen and unseen visual relation triples. However, these contrastive approaches often learned the mapping functions implicitly and did not fully consider the underlying structure of visual relation triples, limiting the models' use cases and their ability to generalize to unseen compositions. This ongoing work aims to construct joint visual and language embedding models that can capture such hierarchical structure between objects and predicates by explicitly imposing structural loss constraints. In this short paper, we propose VLTransE, a novel embedding model that applies translational loss in conjunction with the visual-language contrastive loss to learn transferable embedding spaces for subjects, objects, and predicates. At test time, the model ranks potential visual relationships by aggregating the visual-language consistency score and the translational score. The preliminary results show that the contrastive model trained with the translational loss constraint can capture hierarchical information which aids the prediction of not only visual predicates but also masked-out objects, while achieving comparable predicate prediction results to the model trained without the translational loss.

Keywords

Visual Relationship Detection, Scene Graph, Translational Embedding, Zero-shot Learning, Contrastive Learning

1. Introduction

Understanding the visual world is essential for many modern machine learning tasks including visual question answering [1], image retrieval [2], and image captioning [3]. Visual relationship detection (VRD) [4] aims to facilitate such understanding by bridging the gap between low-level visual information and high-level symbolic visual relation triples, written in the form of *(subject, predicate, object)*. Given the successful performance of deep neural networks in low-level perception tasks such as object classification and object detection, multiple works [4, 5, 6, 7, 8] for VRD have built neural classification models that directly predict the visual predicate from the

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE 2022), Stanford University, Palo Alto, California, USA, March 21–23, 2022.

thanh.tran1725@gmail.com (T. Tran); paulo.santos@flinders.edu.au (P. E. Santos); david.powers@flinders.edu.au (D. Powers)

https://www.flinders.edu.au/people/paulo.santos (P. E. Santos); https://www.flinders.edu.au/people/david.powers (D. Powers)

input image and text, achieving state-of-the-art results on the VRD benchmarks [4, 9]. However, these methods have two main limitations. First, the models learn directly from the dataset distribution, making them susceptible to dataset biases and limiting their ability to generalize to rare compositions of visual relation triples at test time. For example, the classification model can detect *(person, riding, horse)*, while struggle to detect *(person, riding, cow)* or *(person, riding, dog)*. Second, these models are optimized on a narrowly defined task in the given benchmarks, making it difficult to extend the model's use case beyond the given task and domain.

To address these two issues, this research approaches the problem from a different angle. Instead of tackling the VRD problem as an end-to-end classification task, this work assumes the graphical structure of these visual relation triples, interprets this structure as a knowledge graph [10], and formulates the VRD problem as a knowledge graph completion problem [11]. However, unlike traditional knowledge graphs that are based on factual knowledge bases, the knowledge graphs here are represented by a set of visual entities and their interactions, where the nodes are subjects and objects grounded in the image through bounding boxes, and the edges are the relation predicates that exist between pairs of subjects and objects [12]. In the current literature, such formulation of a knowledge graph is also called a scene graph [12], and the task of knowledge graph completion is called scene graph completion [13].

Central to the scene graph completion is the idea of scene graph embedding (SGE), which aims to build embedding models that transform the entities and relations into low-dimensional vector spaces while preserving the structure of the original knowledge graph [10]. Such embedding approach is beneficial to VRD in two ways. First, because these embedding spaces preserve the graphical structure, unseen relations can be inferred by aggregating the relevant neighbors' features [14]. Second, like any other knowledge graph, a scene graph can be augmented with other domain-specific knowledge graphs [15, 16] or common-sense knowledge graphs [17, 18] during training, allowing the model to make out-of-domain inferences at test time.

In this work, we aim to perform scene graph embedding using the contrastive learning approach [19, 20], which learns representations by pulling together the target vector (or anchor) and a matching (positive) vector, while pushing apart the anchor from non-matching (negative) vectors. We believe that such contrastive approach can help us construct a better scene graph representation that can be transferred to other downstream tasks while giving us more control over the output embedding spaces. Thus, this short paper proposes VLTransE (Figure 1), a visual-language contrastive scene graph embedding model that preserves the local structure of the graph through the use of translational loss constraint [21]. At test time, the model is evaluated on the predicate detection task and tail entity (object) prediction task (Figure 2). The preliminary results in Tables 1 and 2 show that the method performs reasonably well on both tasks, while Table 3 shows that the model trained with translational loss can achieve comparable predicate prediction results to the model trained without translational loss on unseen triples.

2. Related Work

This section presents a review of the work related to compositional grounding of visual concepts on language [9], with an emphasis on visual relationship detection and scene graph representational learning through the use of contrastive learning and translational embeddings.

Visual Relationship Detection aims to capture real-world interactions between subject and object pairs (e.g person-riding-horse), allowing the model to detect not only objects but also relations between objects. However, due to the large number of potential real-world interactions, existing visual relationship datasets including VRD [4] and Visual Genome [9] are often sparse and unbalanced, where common relationships occur more frequently than rarer but plausible ones. While Lu et al. [4] and Yu et al. [5] have shown that leveraging language biases can help the models learn co-occurrences statistical priors, such approaches often limit the model's generalization ability and prevent the model from dealing with the variability of visual appearances. Thus, other works [22, 23] have interpreted VRD as a zero-shot detection task, and uses contrastive learning to construct joint visual and language embedding spaces that can be transferred to detect unseen visual relation triples. Here, [22] emphasizes the importance of analogy transfer, which is a downstream neural network module that leverages compositional embedding parts to compose novel visual relation triples. While our method also constructs distinct subject, object, and predicates embedding spaces using contrastive learning, the entire pipeline is trained end-to-end and the method focuses on the use of translation scoring functions (Figure 2) to rank the predicates instead of having a separate downstream network.

Contrastive Learning focuses on minimizing the distance between the target embedding (anchor) vector and the matching (positive) embedding vector, while maximizing the distance between the anchor vector and the non-matching (negative) embedding vectors. Recent work on contrastive learning have shown that discriminative or contrastive approaches can (i) produce transferable embeddings for visual objects through the use of data augmentation [20], and (ii) learn joint visual and language embedding space that can be used to perform zero-shot detection [24]. Given the sparseness and long-tailed property of scene graph datasets, application (i) of contrastive approach can help the model learn better visual appearance embeddings of *(subject, object)* pairs under limited resource settings. Moreover, in application (ii), contrastive learning gives a clearer separation of the visual embeddings and language embeddings compared to the traditional black-box neural fusion approaches [25, 26], giving us more control over both the symbolic triples input and the final output embedding spaces.

Scene Graph Embedding and Translational Embedding. The above task of constructing joint transferable visual and language embedding spaces for $(subject, predicate, object)$ can also be interpreted as a scene graph embedding task. Here, a scene graph is a graph-based formulation that explicitly models objects, attributes of objects, and relationships between objects [12]. Because a scene graph can be interpreted as a knowledge graph, common knowledge graph embedding techniques [27] can also be used for scene graph embeddings. Thus, inspired by translational embeddings [21], H. Zhang et al. [6] builds a model called VTransE that predicts the visual predicate by assuming the translational properties of the $(subject, predicate, object)$ triples, where EMB(subject) + EMB(predicate) $\approx$ EMB(object). Similar to VTransE, the model proposed presented in the present paper also enforces the translational loss constraints to preserve the local graph structure. However, instead of training an end-to-end softmax predictor, the method here uses contrastive learning with negative sampling to learn the three separate visual-language embedding spaces.

In the preliminary research reported in this paper, we aim to perform zero-shot visual re-

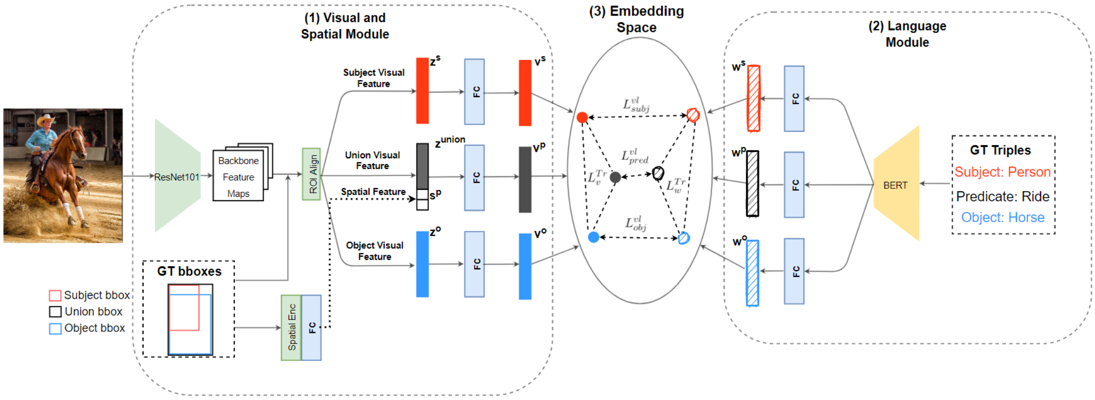

Figure 1: Overview of the proposed model. Red, black, and blue colors represent subject, predicate, and object respectively. The blue FC rectangles are independent fully connected layers with RELU activation function. The final output of the model consists of six embeddings with three visual embeddings (v^s, v^p, and v^o) and three language embeddings (w^s, w^p, and w^o). These embeddings are then trained on two set of losses: (L^{vl}_{subj}, L^{vl}_{pred}, L^{vl}_{obj}) are the visual-language consistency losses, while L^{Tr}_v, L^{Tr}_w are the translational losses for visual and language embeddings triples.

lationship detection through the use of scene graph embedding, where we construct three separate visual and language embedding spaces for subject, predicate, and object using contrastive loss. While there are multiple contrastive loss functions [28, 29, 30], the visual-language contrastive loss in this work uses triplet margin loss, where one anchor vector of one modality is contrasted against one positive and one negative vector of the other modality. To preserve the local structure of the scene graph during embedding, the method also enforces the translational loss constraint separately in the language triplet embeddings and the visual triplet embeddings. While translational loss only preserves the first-order proximity or local structure of the scene graph, we believe that this method can be extended to other scene graph embedding techniques in the future.

3. The VLTransE Architecture

This section describes proposed architecture and outlines the details of the current implementation. The general architecture consists of three modules: (1) the **Visual and Spatial Module** that generates visual embeddings based on the extracted features from the images and bounding boxes' coordinates (Figure 1, left), (2) the **Language Module** that learns contextualized token embeddings which changes according to the context of the input triples (Figure 1, right), (3) the **Loss Functions** that enforce translational losses to preserve the first-order graph structure and visual-language contrastive losses to ensure the consistency between the (visual, language) embeddings pairs (Figure 1, center).

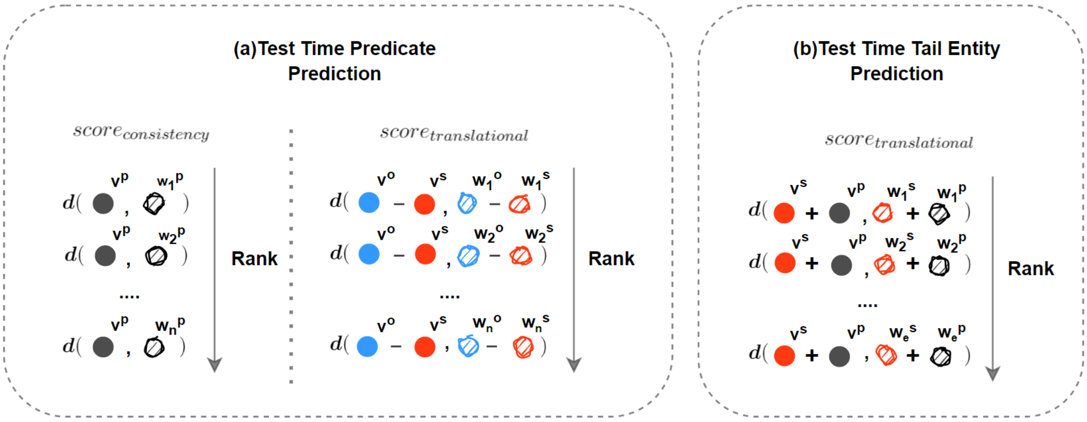
(a)Test Time Predicate Prediction
$score_{consistency}$
$score_{translational}$
$d($ v^p , w_1^p $)$
$d($ v^p , w_2^p $)$
....
$d($ v^p , w_n^p $)$
Rank
$d($ v^o − v^s , w_1^o − w_1^s $)$
$d($ v^o − v^s , w_2^o − w_2^s $)$
....
$d($ v^o − v^s , w_n^o − w_n^s $)$
Rank
(b)Test Time Tail Entity Prediction
$score_{translational}$
$d($ v^s + v^p , w_1^s + w_1^p $)$
$d($ v^s + v^p , w_2^s + w_2^p $)$
....
$d($ v^s + v^p , w_e^s + w_e^p $)$
Rank

Figure 2: Test Time Scoring Functions. Red, black, and blue colors represent subject, predicate, and object respectively. $\boldsymbol{d}$(x,y) computes the cosine distance between x and y, and the distances are ranked in an ascending order. in (a), $\boldsymbol{n}$ is the number of predicate classes in the dataset. In (b), $\boldsymbol{e}$ is the number of object classes in the dataset.

3.1. Visual and Spatial Module

Visual Module. 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 for downstream embeddings. Given the success of CNN-based architecture [31] in learning image representations from large scale pre-training, the visual feature extraction module in this work uses Faster R-CNN [32] pre-trained on the COCO 2017 dataset. In the current implementation, Faster-RCNN consists of a shared ResNet-101 backbone network [33], a region proposal network (RPN), and a region of interest (ROI) detector. Thus, given the ground truth subject, object, and union bounding boxes, the visual features are extracted from the shared backbone and a region of interest pooling operation, yielding $\boldsymbol{z^s}$, $\boldsymbol{z^o}$, and $\boldsymbol{z^{union}}$ respectively.

Spatial Module. The model also extracts spatial information from the subject, object, and union bounding boxes to incorporate spatial and position priors. Similar to J.Zhang et al. [7], given the three boxes b^s, b^{union}, b^o in $[x, y, w, h]$ format, where (x, y) is the starting coordinate and (w, h) is the width and height of the box, the *spatial encoder* generates a 22-dimensional feature vector:

$$\Delta(b^1, b^2) =< \frac{x^1 - x^2}{w^2}, \frac{y^1 - y^2}{h^2}, log(\frac{w^1}{w^2}), log(\frac{h^1}{h^2}) > \tag{1}$$

$$\boldsymbol{c}(b) =< \frac{x}{w_{img}}, \frac{y}{h_{img}}, \frac{x+w}{w_{img}}, \frac{y+h}{h_{img}}, \frac{wh}{w_{img}h_{img}}) > \tag{2}$$

$$< \Delta(b^s, b^o), \Delta(b^s, b^{union}), \Delta(b^{union}, b^o), \boldsymbol{c}(b^s), \boldsymbol{c}(b^o) > \tag{3}$$

The spatial feature vector in Equation (3) then goes through two fully connected layers to get the final 64-dimensional spatial embedding, $\boldsymbol{s^p}$.

These extracted visual and spatial feature vectors are then passed through three separate neural networks to generate the subject, predicate, and object embeddings. For the subject and object embeddings, the visual feature vectors go through three fully connected layers with RELU activation function to get 256-dimensional embedding vectors, v^s and v^o. Similarly, the union feature vector, z^{union}, is first concatenated with the spatial embedding, s^p, before going through three fully connected layers with RELU activation function to get the 256-dimensional predicate embedding vector, v^p.

3.2. Language Module

For the language module, the model also learns three separate neural networks for subject, predicate, and object that map the pre-trained language features toward the final joint visual and language spaces. Here, the architecture uses BERT [34] instead of word2vec [35] as our pre-trained language encoder to leverage the contextualized information from the entire triplet. We believe that contextualized encoders like BERT are beneficial for visual relationship detection because the same predicate can have different meanings under different (subject, object) contexts. Thus, after extracting contextualized feature embeddings from BERT, and passing them through the three separate neural networks, the final output are three 256-dimensional embedding vectors: w^s, w^p, and w^o.

3.3. Loss Functions

The model uses triplet margin loss as the primary metric loss function, although this can be replaced with other contrastive loss functions [28, 30]. Here, cosine similarity is used as the distance metric $\boldsymbol{d}$ for all triplet margin loss functions.

Visual and Language Consistency Loss. Using triplet margin loss, the following loss function aims to bring the three positive visual embeddings (v^s, v^p, v^o) closer to the three positive language embeddings (w^s, w^p, w^o), while pushing apart negative pairs. To reduce the number of equations, the loss function in Equation (6) or L^{vl} is applied separately for the three subject, predicate, and object heads. Therefore, given the set $\boldsymbol{V} = \{\boldsymbol{v}, \boldsymbol{w}\}$ of positive visual and language embedding pairs, the set $\boldsymbol{V}^{v-} = \{\boldsymbol{v}^-, \boldsymbol{w}\}$ of negative visual with positive language pairs, and $\boldsymbol{V}^{w-} = \{\boldsymbol{v}, \boldsymbol{w}^-\}$ of positive visual with negative language pairs, the triplet losses are:

$$L_v^{vl} = \sum_{(v,w)\in\boldsymbol{V}} \frac{1}{|\boldsymbol{V}^{\boldsymbol{v}-}|} \sum_{(v^-,w)\in\boldsymbol{V}^{\boldsymbol{v}-}} [m + \boldsymbol{d}(v,w) - \boldsymbol{d}(v^-,w)]_+ \tag{4}$$

$$L_w^{vl} = \sum_{(v,w)\in\boldsymbol{V}} \frac{1}{|\boldsymbol{V}^{\boldsymbol{w}-}|} \sum_{(v,w^-)\in\boldsymbol{V}^{\boldsymbol{w}-}} [m + \boldsymbol{d}(v,w) - \boldsymbol{d}(v,w^-)]_+ \tag{5}$$

$$L^{vl} = L_v^{vl} + L_w^{vl} \tag{6}$$

where $[x]_+ = max(0,x)$ denotes only the positive part of the input, m denotes a margin of 0.2, and $\boldsymbol{d}$ is cosine similarity distance metric. L^{vl} is applied correspondingly for objects, subjects, and predicates pairs.

Translational Loss. To enforce the structural priors of visual relation triples, the model also enforces the translational loss on the visual embeddings and language embeddings. Thus given a set $\boldsymbol{S}$ of valid triples (s, p, o), and $\boldsymbol{S}^-$ of randomly selected negative triples (s', p', o'), the translational losses are defined as:

$$L_v^{Tr} = \sum_{(s,p,o)\in\boldsymbol{S}} \frac{1}{|\boldsymbol{S}^-|} \sum_{(s',p,o')\in\boldsymbol{S}^-} [m + \boldsymbol{d}(v^s + v^p, v^o) - \boldsymbol{d}(v^{s'} + v^p, v^{o'})]_+ \tag{7}$$

$$L_w^{Tr} = \sum_{(s,p,o)\in\boldsymbol{S}} \frac{1}{|\boldsymbol{S}^-|} \sum_{(s',p,o')\in\boldsymbol{S}^-} [m + \boldsymbol{d}(w^s + w^p, w^o) - \boldsymbol{d}(w^{s'} + w^p, w^{o'})]_+ \tag{8}$$

Here, the predicate embeddings act as the translational vector between the subject and object embeddings. Thus, the final combined loss function is defined as:

$$L = L_{subj}^{vl} + L_{obj}^{vl} + L_{pred}^{vl} + L_v^{Tr} + L_w^{Tr} \tag{9}$$

Test-Time Inference. To perform test-time inference on the generated visual and language embeddings, the evaluation algorithm computes the cosine similarity distances between the visual embeddings and language embeddings, and ranks them to select the top predictions. Depending on the evaluation task, different embedding parts of *(subject, predicate, object)* can be used (Figure 2). In this paper, we evaluated the model on two tasks: (i) the predicate prediction task, and (ii) the tail entity prediction task.

For the predicate prediction task (Figure 2a), both the ground truth bounding boxes and labels for subject and object are given. Thus, given the ground truth bounding boxes and an image, the three visual embeddings (v^s, v^p, v^o) for subject, predicate and object are first generated by the *visual and spatial module* (Figure 1, 1). For the language modality, due to the usage of BERT contextualized encoder, the evaluation algorithm first enumerates all possible $(subject, predicate_i, object)$ triples where $i \in (0, \boldsymbol{n})$ and $\boldsymbol{n}$ is the number of predicates. These triples are then passed through the *language module* (Figure 1, 2) to generate $\boldsymbol{n}$ language embeddings triples, $(w^s, w^p, w^o)_{i\in(0,\boldsymbol{n})}$. Thus, given the visual (v^s, v^p, v^o) embedding triple and the language $(w^s, w^p, w^o)_{i\in(0,\boldsymbol{n})}$ embedding triples, the visual-language consistency score for the predicate is computed as:

$$score_{consistency} = \boldsymbol{d}(v^p, w^p) \tag{10}$$

and the translational score is computed from:

$$score_{translational} = \boldsymbol{d}(v^o - v^s, w^o - w^s) \tag{11}$$

These two scores are then multiplied to get the final ranking score

$$score_{combine} = score_{consistency} * score_{translational} \tag{12}$$

For the tail entity prediction task (Figure 2b), only the ground truth bounding box for subject and ground truth labels for subject and predicate is provided. Thus, without the ground truth object bounding box, the union box is set to be the subject bounding box. Therefore, given the image and the subject bounding box, the *visual and spatial module* (Figure 1, 1) generates the visual embeddings (v^s, v^p) for the subject and predicate. Similarly, from the subject and predicate ground truth labels, the evaluation algorithm first enumerates all possible $(subject, predicate, object_i)$ triples where $i \in \boldsymbol{e}$ and $\boldsymbol{e}$ is the number of object classes. These triples are then passed through the *language module* (Figure 1, 2) to generate $(w^s, w^p, w^o)_{i \in (0, \boldsymbol{e})}$ embedding triples. Given the visual (v^s, v^p) embeddings and the language $(w^s, w^p, w^o)_{i \in (0, \boldsymbol{e})}$ embedding triples, the translational score function is computed as:

$$score_{translational} = \boldsymbol{d}(v^s + v^p, w^s + w^p) \tag{13}$$

4. Preliminary Results

This section evaluates the performance of VLTransE on the VRD dataset, which contains 4000 images for training and 1000 images for testing. In total, the VRD dataset contains 100 object classes, 70 predicate classes, and 37,993 relationships.

Table 1
Predicate Prediction Results on VRD test set

	seen and unseen triples		unseen triples only	
scoring function	Recall top@1	Recall top@5	Recall top@1	Recall top@5
$score_{consistency}$	15.21	36.71	3.93	16.60
$score_{translation}$	6.83	26.01	4.62	16.25
$score_{combined}$	18.64	43.75	6.33	22.58

Table 2
Tail Entity Prediction Results on VRD test set

	seen and unseen triples		unseen triples only	
scoring function	Recall top@1	Recall top@5	Recall top@1	Recall top@5
$score_{translation}$	10.72	33.80	3.51	14.37

Table 3
Comparing the model trained with and without translational loss

	seen and unseen triples		unseen triples only	
model	Recall top@1	Recall top@5	Recall top@1	Recall top@5
without translational loss	24.18	44.20	7.10	22.07
with translational loss	18.64	43.75	6.33	22.58

Evaluation. All evaluation results are computed using the recall metric on the top n ranked items. For the predicate prediction task, Table 1 shows that by multiplying the visual-language consistency score ($score_{consistency}$) and the translational score ($score_{translation}$) instead of using just the visual-language consistency score, the performances of the model when detecting all predicates and unseen predicates improved by 22.6% and 61% respectively on Recall top@1 metric, and by 19.2% and 36% on the Recall top@5 metric. In table 3, it shows the model trained with the additional translational loss performs poorer than the model trained without the translational loss constraint when evaluated on the entire test set. However, the results between the two models become comparable when evaluated solely on unseen compositions of visual relation triples.

With the additional translational structural loss, the embedding space can now be extended to tasks other than visual relationship detection. Here, we evaluated the model on the tail entity prediction task, where the goal is to infer potential objects given only the subject and predicate ground truth label. The results shown in Table 2 indicate that the model can perform reasonably well given that no additional visual information is provided.

5. Discussion and Future Work

Visual Relationship Detection is the cornerstone of many modern machine learning tasks that require a comprehensive understanding of the visual scene. Current contrastive distance metric approaches in learning joint visual-language embeddings for VRD often rely on neural networks learning the necessary transformations implicitly without any structural constraints. To this end, we propose VLTransE, a contrastive visual-language embedding model that preserves the first-order structure of the graph through the use of the translational constraint. While the results shown in Table 3 indicate that additional constraints can interfere with the model learning process and reduce the model's performance on the given VRD benchmark, Table 1 and 2 show the versatility of the embeddings, where the same embedding space can be used for tasks other than visual relationship detection. While the initial results of the model's first iteration is reasonable, further experiments are needed to see the failure corner cases and verify the impact of language biases.

There are certain limitations with the proposed approach that we want to explore in future research. First, the translational loss can only preserve the first order proximity of the scene graph, where only intermediate neighbors features are used, limiting the expressiveness of the embedding spaces. Thus, we might consider extending the method to other graph embedding techniques that consider not only the local structure, but also the global structure. Second, the method shown here uses random triplet selection for the contrastive losses, which could prevent the model from converging to the optimal solution. Therefore, future work may consider other contrastive loss functions and negative sampling techniques. Finally, while the method induces the graphical structure in the final output embedding spaces, it remains open on how to effectively visualize these embeddings or transfer them to create a more explicit representation.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: Visual Question Answering, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, Santiago, Chile, 2015, pp. 2425–2433. doi:10.1109/ICCV.2015.279.

[2] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, C. D. Manning, Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval, in: Proceedings of the Fourth Workshop on Vision and Language, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 70–80. doi:10.18653/v1/W15-2812.

[3] A. Karpathy, L. Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptions, arXiv:1412.2306 [cs] (2015). arXiv:1412.2306.

[4] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual Relationship Detection with Language Priors, arXiv:1608.00187 [cs] (2016). arXiv:1608.00187.

[5] R. Yu, A. Li, V. I. Morariu, L. S. Davis, Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, 2017, pp. 1068–1076. doi:10.1109/ICCV.2017.121.

[6] H. Zhang, Z. Kyaw, S.-F. Chang, T.-S. Chua, Visual Translation Embedding Network for Visual Relation Detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, 2017, pp. 3107–3115. doi:10.1109/CVPR.2017.331.

[7] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, B. Catanzaro, Graphical Contrastive Losses for Scene Graph Parsing, arXiv:1903.02728 [cs] (2019). arXiv:1903.02728.

[8] Y.-C. Su, S. Changpinyo, X. Chen, S. Thoppay, C.-J. Hsieh, L. Shapira, R. Soricut, H. Adam, M. Brown, M.-H. Yang, B. Gong, 2.5D Visual Relationship Detection, arXiv:2104.12727 [cs] (2021). arXiv:2104.12727.

[9] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations, International Journal of Computer Vision 123 (2017) 32–73. doi:10.1007/s11263-016-0981-7.

[10] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A Survey on Knowledge Graphs: Representation, Acquisition and Applications, IEEE Transactions on Neural Networks and Learning Systems (2021) 1–21. doi:10.1109/TNNLS.2021.3070843. arXiv:2002.00388.

[11] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, Z. Duan, Knowledge Graph Completion: A Review, IEEE Access 8 (2020) 192435–192456. doi:10.1109/ACCESS.2020.3030076.

[12] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Image retrieval using scene graphs, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, 2015, pp. 3668–3678. doi:10.1109/CVPR.2015.7298990.

[13] H. Wan, Y. Luo, B. Peng, W.-S. Zheng, Representation Learning for Scene Graph Completion via Jointly Structural and Visual Embedding, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, 2018, pp. 949–956. doi:10.24963/ijcai.2018/132.

[14] P. Maheshwari, R. Chaudhry, V. Vinay, Scene Graph Embeddings Using Relative Similarity Supervision, arXiv:2104.02381 [cs] (2021). arXiv:2104.02381.

[15] [1412.0691] RoboBrain: Large-Scale Knowledge Engine for Robots, https://arxiv.org/abs/1412.0691, ????

[16] C. Henson, S. Schmid, T. Tran, A. Karatzoglou, Using a Knowledge Graph of Scenes to Enable Search of Autonomous Driving Data (????) 2.

[17] R. Speer, J. Chin, C. Havasi, ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, arXiv:1612.03975 [cs] (2018). arXiv:1612.03975.

[18] M. Sap, R. LeBras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning, arXiv:1811.00146 [cs] (2019). arXiv:1811.00146.

[19] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, W.-Y. Ma, Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 6602–6611. doi:10.1109/CVPR.2019.00677.

[20] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, in: Proceedings of the 37th International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.

[21] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, in: Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013.

[22] J. Peyre, J. Sivic, I. Laptev, C. Schmid, Detecting Unseen Visual Relations Using Analogies, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Seoul, Korea (South), 2019, pp. 1981–1990. doi:10.1109/ICCV.2019.00207.

[23] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, M. Elhoseiny, Large-Scale Visual Relationship Understanding, arXiv:1804.10660 [cs] (2019). arXiv:1804.10660.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, arXiv:2103.00020 [cs] (2021). arXiv:2103.00020.

[25] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, UNITER: UNiversal Image-TExt Representation Learning, arXiv:1909.11740 [cs] (2020). arXiv:1909.11740.

[26] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: Pre-training of Generic Visual-Linguistic Representations, arXiv:1908.08530 [cs] (2020). arXiv:1908.08530.

[27] F. Bianchi, G. Rossiello, L. Costabello, M. Palmonari, P. Minervini, Knowledge Graph Embeddings and Explainable AI, arXiv:2004.14843 [cs] (2020). doi:10.3233/SSW200011. arXiv:2004.14843.

[28] K. Sohn, Improved Deep Metric Learning with Multi-class N-pair Loss Objective, in: Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016.

[29] B. Yu, T. Liu, M. Gong, C. Ding, D. Tao, Correcting the Triplet Selection Bias for triplet loss, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, volume 11210, 2018, p. 17.

[30] A. van den Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, arXiv:1807.03748 [cs, stat] (2019). arXiv:1807.03748.

[31] Y. LeCun, P. Haffner, L. Bottou, Y. Bengio, Object Recognition with Gradient-Based Learning, in: D. A. Forsyth, J. L. Mundy, V. di Gesú, R. Cipolla (Eds.), Shape, Contour and Grouping in Computer Vision, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 1999, pp. 319–345. doi:10.1007/3-540-46805-6_19.

[32] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in: Advances in Neural Information Processing Systems, volume 28, Curran Associates, Inc., 2015.

[33] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, arXiv:1512.03385 [cs] (2015). arXiv:1512.03385.

[34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805 [cs] (2019). arXiv:1810.04805.

[35] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781 [cs] (2013). arXiv:1301.3781.