Improving Visual Reasoning by Exploiting The Knowledge in Texts

Sahand Sharifzadeh [*] [1] **Sina Moayed Baharlou** [*] [2] **Martin Schmitt** [3]
Hinrich Schütze [3] **Volker Tresp** [1] [4]

Abstract

This paper presents a new framework for training image-based classifiers from a combination of texts and images with very few labels. We consider a classification framework with three modules: a backbone, a relational reasoning component, and a classification component. While the backbone can be trained from unlabeled images by self-supervised learning, we can fine-tune the relational reasoning and the classification components from external sources of knowledge instead of annotated images. By proposing a transformer-based model that creates structured knowledge from textual input, we enable the utilization of the knowledge in texts. We show that, compared to the supervised baselines with 1% of the annotated images, we can achieve ~8x more accuracte results in scene graph classification, ~3x in object classification, and ~1.5x in predicate classification.

1. Introduction

Relational reasoning is one of the essential components of intelligence; humans explore their environment by considering the entire context of a scene rather than studying each item in isolation from the others. Furthermore, they expand their understanding of the world by educating themselves about novel relations through reading or listening. For example, we might have never seen a "cow wearing a dress" but might have read about Hindu traditions of decorating cows. While we already have a robust visual system that can extract basic visual features such as edges and curves from a scene, the description of a "cow wearing a dress" refines our visual understanding of relations on an *object level* and enables us to recognize one when seeing it quickly.

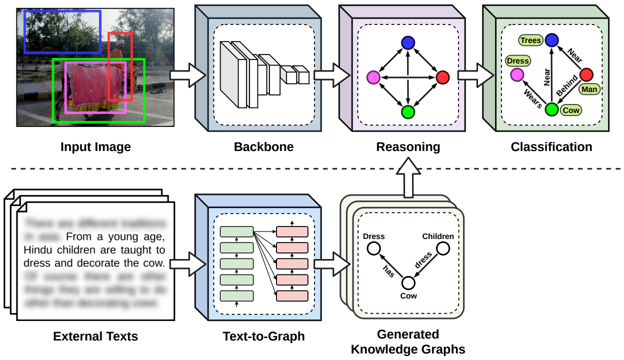

Figure 1. An example of scene graph classification given an image from the Visual Genome dataset. In this architecture, the backbone extracts image-based features from the objects in an image, while the reasoning component passes messages among different objects. In this work, we extract external knowledge from texts and utilize it to fine-tune the reasoning component directly and without requiring additional images.

Relational reasoning is also gaining growing popularity in the Computer Vision community. While most of the approaches still treat each object representation as independent from the others in an image, some recent works propose to contextualize object representations by applying message-passing functions before classification (Xu et al., 2017; Zellers et al., 2018; Yang et al., 2018; Sharifzadeh et al., 2021). These works generally consider a classification pipeline of three-parts: a *backbone*, a *relational reasoning* and a *classification* component (Figure 1). The backbone is typically a convolutional neural network (CNN) that detects objects and extracts image-based representations for each of them. The relational reasoning can be a variant of a recurrent neural network (Hochreiter and Schmidhuber, 1997) or graph convolutional networks (Kipf and Welling, 2016). This component operates on an object level by taking the representations of all the extracted objects from the backbone and propagating them to each other. In the end, the classification component takes the enriched visual representations of each object and classifies them.

Having a relational reasoning component in the classification pipeline opens up exciting new directions awaiting exploration. While the backbone requires images

[*] Equal contribution [1] Department of Informatics, LMU Munich, Germany [2] Sapienza University of Rome, Italy [3] Center for Information and Language Processing (CIS), LMU Munich, Germany [4] Siemens AG, Munich, Germany. Correspondence to: S. Sharifzadeh <sahand.sharifzadeh@gmail.com>, S.M. Baharlou <sina.baharlou@gmail.com>.

Preliminary work. Under review.

as input, the relational reasoning component operates on graphs with their nodes representing the object and their edges representing the relation embeddings. Therefore, instead of requiring images and depending on a backbone, one can fine-tune the relational reasoning component directly from knowledge graphs (KGs) and by employing a canonical form of object and relation representations as input (Sharifzadeh et al., 2021). Knowledge graphs can contain curated facts represented as `(head, predicate, tail)` such as `(Person, Rides, Horse)`. However, relational knowledge is not always stored in the structured form of graphs. In fact, most of the collective human knowledge is only available in the unstructured form of texts and documents. Exploiting this form of knowledge can be extremely beneficial, specially in domains where the knowledge is not stored in the machine-accessible form of KGs.

In this paper, we propose *Texema*, the first image-based classification pipeline equipped with a relational reasoning component that is trained from the large corpora of unstructured knowledge available in texts. Our framework employs a transformer-based model to create structured graphs from textual input and utilizes them to improve the relational reasoning module. We show that not only our text-to-graph component outperforms previous works in that domain by a large margin, but also our results in image-based object and relation classification mark a new milestone in few-shot learning. Specifically, we evaluate our approach on the scene graph classification task on the Visual Genome dataset. Scene graph classification is a fundamental task in scene understanding where the goal is to classify both objects and their relations in an image. We show that by utilizing a combination of self-supervised learning and external knowledge of texts, and compared to the supervised baselines, we can achieve ~3x more accurate results in object classification, ~8x in scene graph classification and ~1.5x in predicate classification using as little as only 500 annotated images.

2. Related Works

Scene Graph Classification There is an extensive body of work on visual reasoning in general (Wu et al., 2014; Deng et al., 2014; Hu et al., 2016; 2017; Santoro et al., 2017; Zellers et al., 2019). Here, we mainly review the works that are focused on the scene graph classification and were published following the release of Visual Relation Detection (VRD) (Lu et al., 2016) and the Visual Genome (Krishna et al., 2017) datasets. While the original papers on VRD and VG provided the baselines for scene graph classification by treating objects as independent of each other, later, several works considered contextualizing the entities before classification. Iterative Message Passing (IMP) (Xu et al., 2017), Neural Motifs (Zellers et al., 2018) (NM), Graph R-CNN (Yang et al., 2018), and Schemata (Sharifzadeh et al., 2021) proposed to propagate the image context using basic RNNs, LSTMs, graph convolutions, and graph transformers respectively. On the other hand, authors of VTransE (Zhang et al., 2017) proposed to capture relations by applying TransE (Bordes et al., 2013), a knowledge graph embedding model, on the visual embeddings, Tang et al. (2019) exploited dynamic tree structures to place the object in an image into a visual context. Chen et al. (2019a) proposed a multi-agent policy gradient method that frames objects into cooperative agents and then directly maximizes a graph-level metric as the reward. In tangent to those works, Sharifzadeh et al. (2019) proposed to enrich the input domain in scene graph classification by employing the predicted pseudo depth maps of VG images that were released as an extension called *VG-Depth*.

Commonsense in Scene Understanding Several recent works have proposed to employ external or internal sources of knowledge to improve visual understanding (Wang et al., 2018; Jiang et al., 2018; Singh et al., 2018; Kato et al., 2018). In the scene graph classification domain, some of the works have proposed to correct the SG prediction errors by merely comparing them to the co-occurrence statistics of internal triples as a form of commonsense knowledge (Chen et al., 2019c;b; Zellers et al., 2018). Earlier, Baier et al. (2017; 2018) proposed the first scene graph classification model that employed prior knowledge in the form of Knowledge Graph Embeddings (KGEs) that generalize beyond the given co-occurrence statistics. Zareian et al. (2020a;b) followed this approach by extending it to models that are based on graph convolutional networks. More recently, Sharifzadeh et al. (2021) proposed *Schemata* as a generalized form of a KGE model that is learned directly from the images rather than triples. In general, scene graph classification methods are closely related to the KGE models. Therefore, we refer the interested readers to (Nickel et al., 2016; Ali et al., 2020a;b) for a review and large scale study on the KG models, and to (Tresp et al., 2019; 2020) for an extensive investigation of the connection between perception, KG models, and cognition.

Nevertheless, to the best of our knowledge, the described methods have employed curated knowledge in the form of triples, and none of them have directly exploited the textual knowledge. In this direction, the closest work to ours is by Yu et al. (2017), proposing to distill the external language knowledge using a teacher-student model. However, this work does not include a relational reasoning component and only refines the final predictions. Additionally, as shown in the experiments, the knowledge extraction module that we use in this paper is more than two times more accurate than the Stanford SG Parser (Schuster et al., 2015) used in that work.

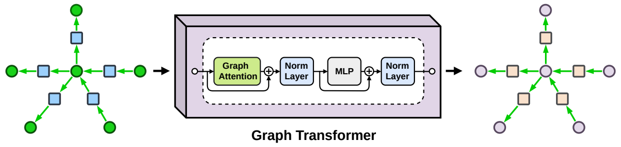

Figure 2. Graph Transformers take the node and edge embeddings of a graph as input and apply message propagations over several attention heads.

Knowledge Extraction from Text Knowledge extraction from text has been studied for a long time (Chinchor, 1991). Previous work ranges from pattern-based approaches (Hearst, 1992) to supervised neural approaches with specialized architectures (Gupta et al., 2019; Yaghoobzadeh et al., 2017). Recently, Schmitt et al. (2020) successfully applied a general sequence-to-sequence architecture to graph↔text conversion. With the recent rise of transfer learning in NLP, an increasing number of approaches is based on large language models, pretrained in a self-supervised manner on massive amounts of texts (Devlin et al., 2019). Inspired from previous work that explores transfer learning for graph-to-text conversion (Ribeiro et al., 2020), we base our text-to-graph model on a pretrained T5 model (Raffel et al., 2019).

3. Methods

In this section, we first describe the backbone, relational reasoning, and classification components. Then, we introduce the Text-to-Graph module and discuss the utilization of the external knowledge through these components.

In what follows, bold lower case letters denote vectors, bold upper case letters denote matrices, and the letters denote scalar quantities or random variables. We use subscripts and superscripts to denote variables and calligraphic upper case letters for sets.

3.1. Backbone

For the backbone, we use a convolutional neural network (ResNet-50) that has been pre-trained from unlabeled images of ImageNet (Deng et al., 2009) and Visual Genome (Krishna et al., 2017) by BYOL (Grill et al., 2020), a self-supervised representation learning technique. Given an image $\mathbf{I}$ with several objects in bounding boxes $\mathcal{B} = \{\mathbf{b}_i\}_{i=1}^{n}$, $\mathbf{b}_i = [b_i^x, b_i^y, b_i^w, b_i^h]$, we apply the ResNet-50 to extract pooled object features $\mathcal{X}^o = \{\mathbf{x}_i^o\}_{i=1}^{n}$, $\mathbf{x}_i^o \in \mathbb{R}^d$. Here $[b_i^x, b_i^y]$ are the coordinates of $\mathbf{b}_i$ and $[b_i^w, b_i^h]$ are its width and height, and d are the vector dimensions. Following (Zellers et al., 2018), we define $\mathcal{X}^p = \{\mathbf{x}_i^p\}_{i=1}^{m}$, $\mathbf{x}_i^p \in \mathbb{R}^d$ as the relational features between each pair of objects. Each $\mathbf{x}_i^p$ is initialized by applying a two layered fully connected network on the relational position vector $\mathbf{t}$ between a head i and a tail j where $\mathbf{t} = [t_x, t_y, t_w, t_h]$, $t_x = (b_i^x - b_j^x)/b_i^w{}_j, t_y = (b_i^y - b_j^y)/b_j^h, t_w = \log(b_i^w/b_j^w), t_h = \log(b_i^h/b_j^h)$. The implementation and pre-training details of the layers are provided in Experiments.

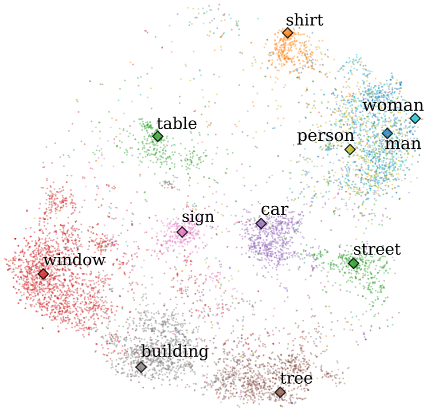

Figure 3. The t-SNE of image-based representations and schemata of 11 sample classes from the training set of VG. Each colored dot represents examples from a unique class. The labeled shapes represent the schemata of each class.

Given $\mathcal{X}^o$ and $\mathcal{X}^p$ we form a structured presentation of the objects and predicates in $\mathbf{I}$ called a **Scene Representation Graph** (SRG) (Sharifzadeh et al., 2021). SRG is a fully connected graph with each node representing either an object or a predicate, such that each object node is a direct neighbor to predicate nodes and each predicate node is a direct neighbor with its head and tail object nodes (Figure 2).

3.2. Relational Reasoning

In the relational reasoning component we take the initial SRG representations and update them by message passing with four layers of a Graph Transformer (Koncel-Kedziorski et al., 2019; Sharifzadeh et al., 2021) as shown in Figure 2. The contextualized representations in the SRG have the same number of dimensions as $\mathcal{X}$s and are represented as $\mathcal{Z}^o = \{\mathbf{z}_i^o\}_{i=1}^{n}$, $\mathbf{z}_i^o \in \mathbb{R}^d$ and $\mathcal{Z}^p = \{\mathbf{z}_i^p\}_{i=1}^{m}$, $\mathbf{z}_i^p \in \mathbb{R}^d$.

3.3. Classification

Following Sharifzadeh et al. (2021), we define classification as the attention between $\mathcal{Z}$s and the set of *schemata* as class-based vectors $\mathcal{S}^o = \{\mathbf{s}_i^o\}_{i=1}^{|\mathcal{C}^o|}$, $\mathbf{s}_i^o \in \mathbb{R}^d$ and $\mathcal{S}^p = \{\mathbf{s}_i^p\}_{i=1}^{|\mathcal{C}^p|}$, $\mathbf{s}_i^p \in \mathbb{R}^d$ such that $\mathcal{C}^o$ and $\mathcal{C}^p$ are the set of all object and predicate classes. Therefore, the attention co-efficients, i.e.,

Input	man standing with child on ski slope
Reference Graph (RG)	(child, on, ski slope) (man, standing with, child) (man, on, ski slope)
$R^{\text{text}\rightarrow\text{graph}}$	🟥 (man, standing, child)
Stanford Scene Graph Parser	🟥 (standing, with, child) , 🟥 (standing, on, slope)
CopyNet (1%)	🟩 (man, standing with, child)
Ours (1%)	🟩 (man, standing with, child)
CopyNet (10%)	🟩 (man, standing with, child) 🟥 (child, on, slope)
Ours (10%)	🟩 (man, standing with, child) 🟩 (child, on, ski slope)

Table 1. Example fact extractions and evaluation wrt. reference graph (RG). Green: correct (∈ RG). Red: incorrect (∉ RG).

the classification outputs, are computed as

$$\alpha_{ic} = \text{softmax}(a(\mathbf{z}_i, \mathbf{s}_c)), \tag{1}$$

and the attention values as

$$\boldsymbol{\delta}_i = \sum_{c \in \mathcal{C}} \alpha_{ic} \mathbf{s}_c, \tag{2}$$

such that $a(.)$ is a dot-product attention function. We omitted the superscripts of o and p for brevity. The schema of each class is originally initialized by a random vector. However, during the training we apply a categorical cross entropy, between the one-hot encoded ground truth labels and the attention coefficients α_{ic}. As a result, the schema of each class, will adapt such that it can capture the gist of visual relation information from that class given the image-based embeddings (Figure 3). We use this observation to utilize the knowledge in texts.

3.4. Text-to-Graph

We employ a pretrained sequence-to-sequence T5$_{\text{small}}$ model (Raffel et al., 2019) based on transformers (Vaswani et al., 2017) and fine-tune it for the task of knowledge extraction. T5 consists of an encoder with several layers of self-attention (like BERT, Devlin et al., 2019) and a decoder with autoregressive self-attention (like GPT-3, Brown et al., 2020). Adapting the multi-task setting from T5's pretraining, we use the task prefix "make graph: " to mark our text-to-graph task. Following Schmitt et al. (2020), we serialize the graphs by writing out their facts separated by end-of-fact symbols (`EOF`), and separate the elements of each fact (head, predicate, and tail) with `SEP` symbols. Table 1 shows an example text (Input) and its corresponding reference graph (RG).

	Precision		Recall		F1	
	1%	10%	1%	10%	1%	10%
$R^{\text{text}\rightarrow\text{graph}}$	$1.92_{\pm0.00}$	$1.86_{\pm0.01}$	$1.87_{\pm0.00}$	$1.81_{\pm0.01}$	$1.89_{\pm0.00}$	$1.84_{\pm0.01}$
SSGP	$14.86_{\pm0.01}$	$14.52_{\pm0.02}$	$18.47_{\pm0.01}$	$18.05_{\pm0.02}$	$16.47_{\pm0.01}$	$16.09_{\pm0.02}$
CopyNet	$29.20_{\pm0.13}$	$30.77_{\pm0.49}$	$27.19_{\pm0.28}$	$29.79_{\pm0.29}$	$28.16_{\pm0.21}$	$30.27_{\pm0.34}$
Ours	$\mathbf{33.37}_{\pm0.11}$	$\mathbf{33.81}_{\pm0.08}$	$\mathbf{31.06}_{\pm0.18}$	$\mathbf{32.45}_{\pm0.33}$	$\mathbf{32.17}_{\pm0.13}$	$\mathbf{33.12}_{\pm0.16}$

Table 2. The mean and standard deviation of the Precision, Recall and F1 scores from the predicted facts given the textual descriptions. The results are computed on the respected Visual Genome splits using our model and our implementation of the related works.

3.5. Utilizing the External Knowledge

As discussed earlier, while the attention coefficients α_{ic} are the classification outputs, the attention values $\boldsymbol{\delta}_i$ capture a weighted canonical representation of the class that the given object or predicate node belongs to. This brings us to the point that it enables us to fine-tune our relational reasoning component even when no images are available. After training our Text-to-Graph model, we apply it to the external texts and extract a set of knowledge graphs such as $KG = \{\mathcal{V}, \mathcal{E}\}$ with n' nodes and m' edges such that $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^{n'}$ assigns each node to a one-hot vector where the index of the corresponding object class for that node is set to one. Same applies to each edge and its corresponding predicate class in $\mathcal{E} = \{\mathbf{e}_i\}_{i=1}^{m'}$. We then create a graph of schema representations by setting α_{ic}s in Equation 2 to the one-hot vectors in $\mathcal{V}$ and $\mathcal{E}$. We consider a training paradigm similar to denoising autoencoders, where we randomly drop some of the nodes and edges from the schema graph and feed it to the relational reasoning component, and fine-tune it towards predicting the complete graph.

At test time, the recurrent process of (a) classification, (b) injection of schemata to the relational reasoning component, and (c) contextualization, is referred to as one step of *assimilation* (Piaget, 1923; Sharifzadeh et al., 2021). In this work, we assimilate the results only two times.

4. Evaluation

We first measure the performance of *Texema* in predicting the graphs from textual descriptions and compare them to the state-of-the-art models. We then evaluate its performance when classifying objects and relations in images. In particular, we show that the extracted knowledge from the texts can compensate for the lack of annotated images and handcrafted knowledge graphs.

Dataset We use the splits of Visual Genome (VG) dataset (Krishna et al., 2017) proposed by Sharifzadeh et al.

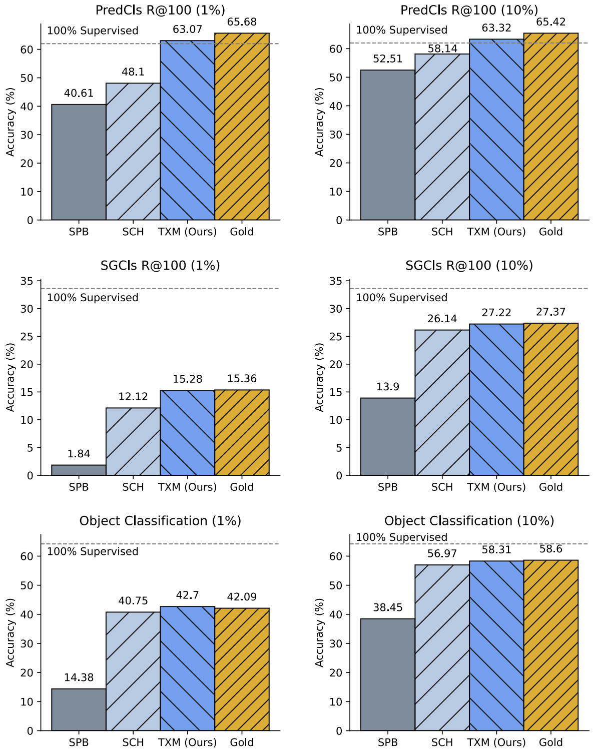

Figure 4. This figure represents the predicate classification (top), scene graph classification (middle) and object classification (bottom) accuracies on 1% and 10% splits of the VG. The results are computed with the supervised baseline, self-supervised baseline, after fine-tuning the self-supervised baseline with the extracted knowledge from texts, and after fine-tuning the self-supervised baseline with the gold standard knowledge provided by the crowdsourcers. The dashed line represents the accuracy of a supervised model on 100% of the data.

(2021). The training splits consist of a *parallel set* and a *disjoint set*. The parallel set has 1% and 10% randomly sampled data from the sanitized training set of VG by Xu et al. (2017) while the disjoint set contains the rest (99% and 90% of the VG). Both sets consist of images and their annotations, i.e., bounding boxes, scene graphs, and scene descriptions in the form of texts. However, while the images in the parallel set are treated as annotated, the images in the disjoint set are considered as unlabeled and separate from their annotations such that: the images are used for self-supervised learning (without labels), the scene graphs annotations are used as the golden standard external knowledge (annotated by crowd-sourcing and cleaned by the expert), and the scene descriptions are considered as the external sources available as texts. Note that the scene graphs and scene descriptions from the VG are collected separately and crowd-sourcing. Therefore, even though the graphs and descriptions describe the same image region, they are disjointed and might contain complementary knowledge.

4.1. Graphs from Texts

The goal of this experiement is to study the effectiveness of the Text-to-Graph model in predicting graphs from texts.

Experiment We fine-tune our pretrained Text-to-Graph model on the scene descriptions and triples in the parallel sets and apply it on the disjoint sets to extract knowledge graphs. Table 1 shows example predictions of our model and three baselines in a qualitative way and Table 2 presents the precision, recall and F1 measures of these models.

We consider the following baselines: (1) $\mathrm{R}^{\text{text}\rightarrow\text{graph}}$ is a simple rule-based system introduced by Schmitt et al. (2020) for general knowledge graph generation from text. (2) The rules of the Stanford Scene Graph Parser (SSGP) (Schuster et al., 2015), also used by (Yu et al., 2017), that are more adapted to the scene graph domain. It was not specifically designed to match the scene graphs from Visual Genome, but it was still engineered to cover typical idiosyncrasies of textual image descriptions and corresponding scene graphs. (3) CopyNet (Gu et al., 2016) is an LSTM sequence-to-sequence model with a dedicated copy mechanism, which allows copying text elements directly into the graph output sequence. It was used for unsupervised text-to-graph generation by Schmitt et al. (2020), but we train it on the supervised data of our parallel sets. We use a vocabulary of around 70k tokens extracted from the VG-graph-text benchmark (Schmitt et al., 2020) and, otherwise, also adopt the hyperparameters from (Schmitt et al., 2020).

4.2. Graphs from Images

These experiments' goal is to study whether the external knowledge of texts can improve scene graph classification and how.

Experiments Here, we conduct the following studies:

We train our models (a) with a backbone trained by supervised learning, while the relational reasoning component has only been trained on the parallel set of data (1% or 10%) (SPB) (b) with a backbone trained by self-supervised learning, while the relational reasoning component has only been trained on the parallel set of data (1% or 10%) (SCH by Sharifzadeh et al. (2021)) (c) with a backbone trained by self-supervised learning, while the relational reasoning component has been trained on the parallel set of data (1% or 10%) and fine-tuned by exploiting the knowledge extracted from the texts (TXM), (d) with a backbone trained by self-supervised learning, while the relational reasoning component has been trained on the parallel set of data (1% or 10%) and fine-tuned from the gold standard external knowledge, and finally (e) with supervised learning on the 100% of the data (dashed line).

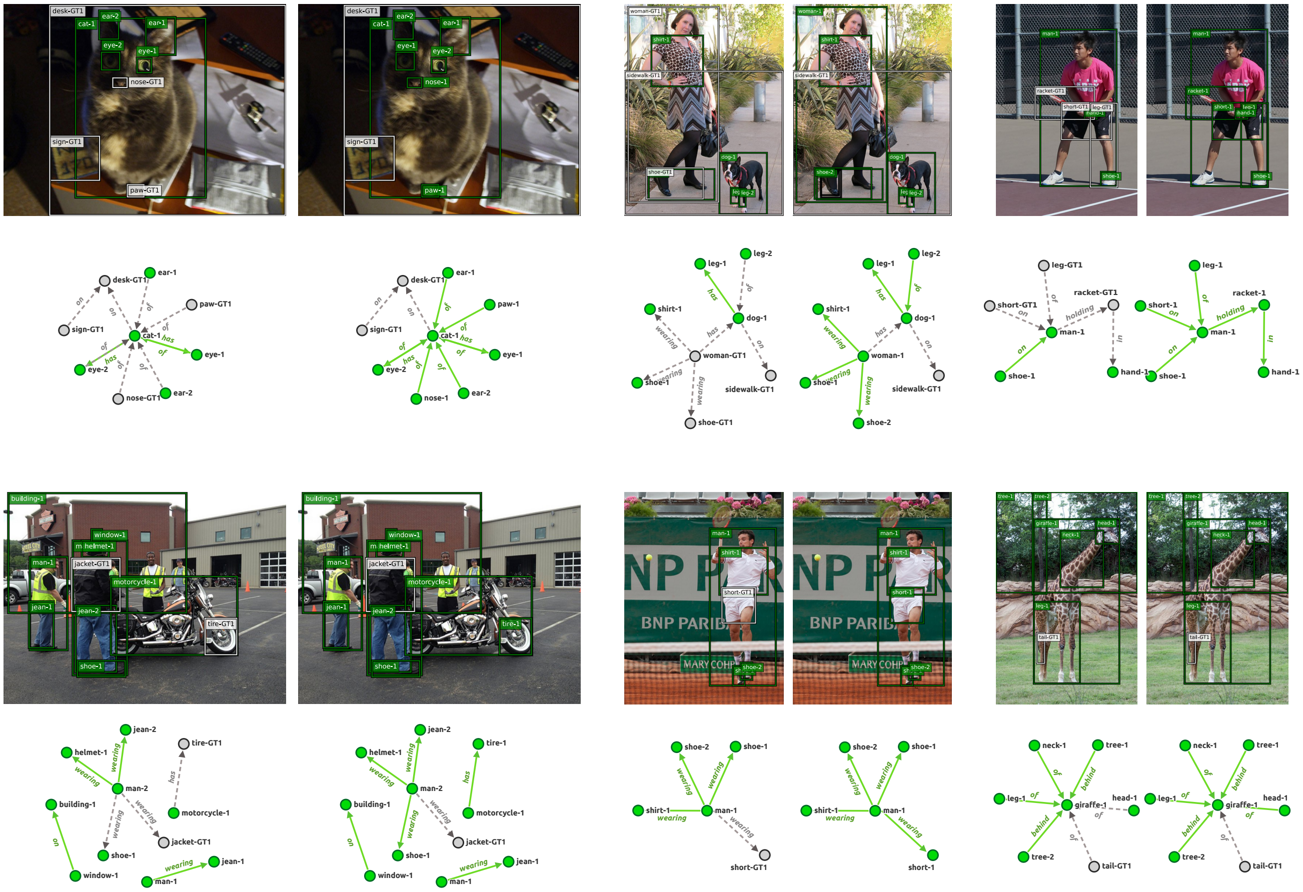

Figure 5. Qualitative examples of improved scene graph classification results (Recall@100) before and after (from left to right) fine-tuning the model using the knowledge in texts. Green and gray colors indicate true positives and false negatives concluded by the model.

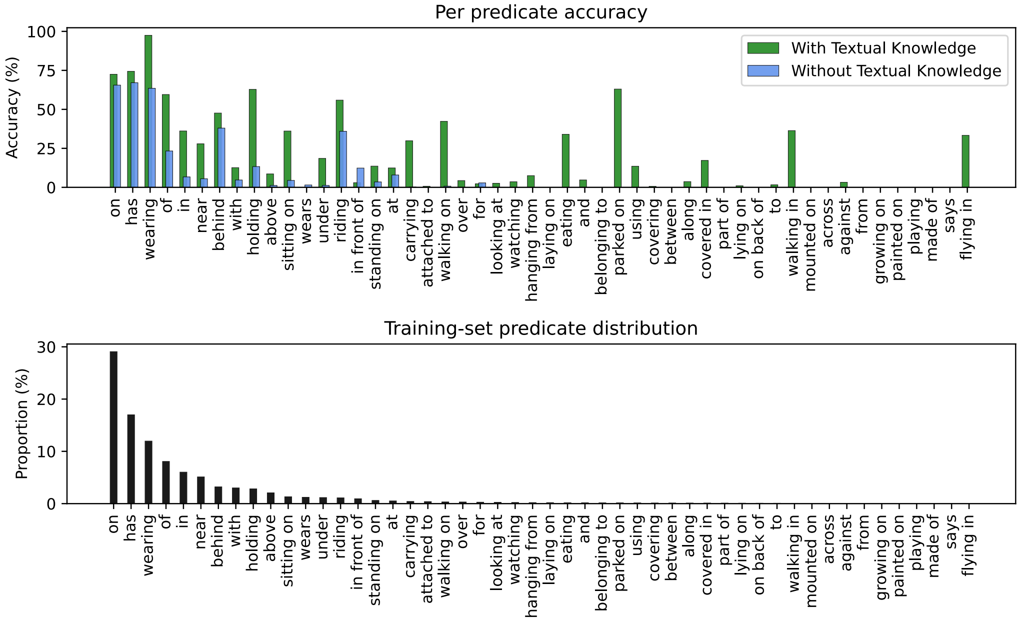

Figure 6. This plot shows the per predicate classification improvement (Recall@ 100) before and after fine-tuning our model with the external textual knowledge.

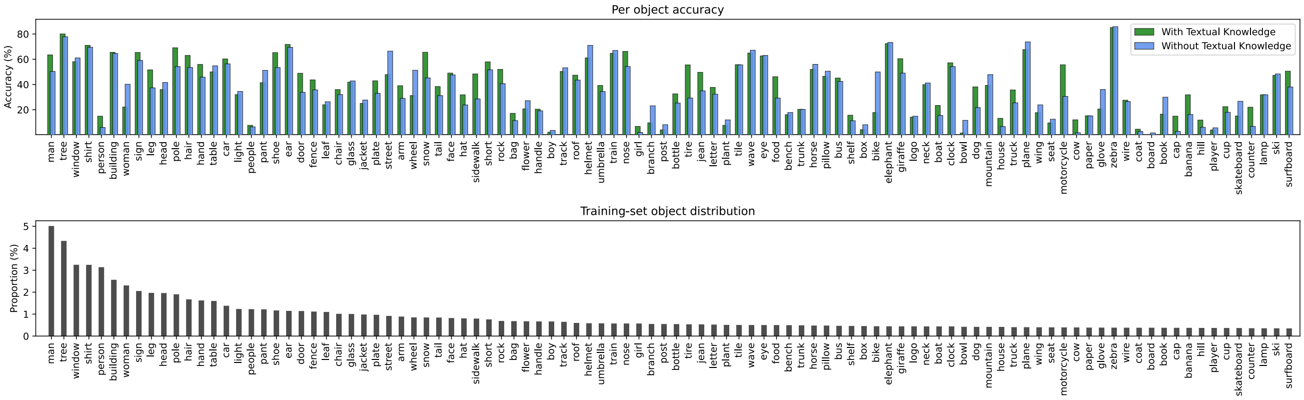

Figure 7. This plot shows the per object classification improvement before and after fine-tuning our model with the external textual knowledge.

Figure 4 presents the results on the test set for object classification, predicate classification (PredCls - predicting predicate labels given a ground truth set of object boxes and object labels) as well as scene graph classification (SGCls - predicting object and predicate labels, given the set of object boxes). Since the focus of our study is not on improving the object detector backbone, we do not report the scene graph *detection* results here. For a metric, we use the Recall@K (**R@K**). R@K computes the mean prediction accuracy in each image given the top K predictions. For the complete set of results under *constrained* and *unconstrained* setups (Yu et al., 2017), and also with the Macro Recall (Sharifzadeh et al., 2019; Chen et al., 2019c) (**mR@K**), refer to the supplementary materials.

As show in Figure 4, the external knowledge from the texts can not only improve the classification results under all settings, but their combination with self-supervised model can already outperform the supervised models in PredCls. Also, they are almost as effective as when using the gold standard crowd-sourced triples, and in some settings even richer than those (object classification with 1%). Notice that compared to the self-supervised baseline, we gained up to ∼5% *relative improvement* in object classification, more than ∼26% in scene graph classification, and ∼31% in predicate prediction accuracy.

Additionally, Table 7 and Table 6 display the improvements per object and per predicate class after fine-tuning the model with external knowledge of texts (From SCH to TXM) together with the table representing the frequency of their appearance in the dataset. The goal of these tables is to understand better where the improvements are happening; the results indicate that most improvements occur in under-represented classes. This means that we have achieved a generalization performance that is beyond the simple reflection of the dataset's statistical bias. For example, interestingly we have improved the classification of objects such as a `Motorcycle` and `Surfboard` and predicates such as `Flying in` and `Parked On`.

Moreover, Figure 5 provides qualitative examples of the predicted scene graphs before and after fine-tuning the relational reasoning with the external knowledge from texts. For example, in the top-left image, after fine-tuning the model with external knowledge, not only we can correctly classify the `Paw` and `Nose`, but we can also detect that all the other parts also belong to the `cat`. We can observe a similar effect in all other examples.

Note that while a major stream of research works on the Visual Genome has been focused on utilizing 100% of the data and using a pre-trained VGG-16 (Simonyan and Zisserman, 2014) backbone, in this work, similar to (Sharifzadeh et al., 2021), we are focused on the few-shot learning setting and using the self-supervised backbone by BYOL (Grill et al., 2020) (based on ResNet-50 (He et al., 2016)). In other words, the goal of this work is not to compare with the major stream of research that uses 100% of the data but rather to evaluate the effectiveness of textual knowledge in scarsity of annotated data.

4.3. Implementation Details

We use ResNet-50 (He et al., 2016) for the backbone and follow Sharifzadeh et al. (2021) for trainings. We train the supervised backbones on the corresponding split of visual genome training set with the Adam optimizer (Kingma and Ba, 2014) and a learning rate of 10^{-5} for 20 epochs with a batch size of 6. We train the self-supervised backbones with the BYOL (Grill et al., 2020) approach. We fine-tune the pre-trained *self-supervised* weights over ImageNet (Deng et al., 2009), on the entire training set of Visual Genome images in a self-supervised manner with no labels, for 3 epochs with a batch-size of 6, SGD optimizer with a learning rate of 6×10^{-5}, momentum of 0.9 and weight decay of

4×10^{-4}. Similar to BYOL, we use a MLP hidden size of 512 and a projection layer size of 128 neurons. Then for each corresponding split, we fine-tune the weights in a supervised manner with the Adam optimizer and a learning rate of 10^{-5} for 4 epochs with a batch size of 6.

After extracting the image-based embeddings from the penultimate fully connected (fc) layer of the backbones, we feed them to a fc-layer with 512 neurons and a Leaky ReLU with a slope of 0.2, together with a dropout rate of 0.8. This gives us initial object node embeddings. We apply a fc-layer with 512 neurons and Leaky ReLU with a slope 0.2 and dropout rate of 0.1, to the extracted spatial vector $\mathbf{t}$ to initialize predicate embeddings. For the relational reasoning component, we take four graph transformer layers of 5 heads each with 2048 and 512 neurons in each fully connected layer. We initial the layers with using Glorot weights (Glorot and Bengio, 2010). We train our supervised models with the Adam optimizer and a learning rate of 10^{-5} and a batch size of 22, with 5 epochs for the 1% and 11 epochs for 10%. We train our self-supervised model with the Adam optimizer and a learning rate of 10^{-5} and a batch size of 22, with 6 epochs for the 1% and 11 epochs for 10%. Finally, we train our self-supervised model including the external knowledge with the Adam optimizer and a learning rate of 10^{-5} with a batch size of 16, with 6 epochs for the 1% and 11 epochs for 10%. We train these models with 2 assimilations.

5. Conclusion

In this work, we proposed the first relational image-based classification pipeline that can be fine-tuned directly from the large corpora of unstructured knowledge available in texts. We proposed to genarate structured graphs from textual input using a transformer-based model. We then fine-tuned the relational reasoning component of our classification pipeline by employing the canonical representations (schemata) of each entity in the generated graphs. We showed that our text-to-graph component outperforms other graph generator methods. Also, after employing the generated knowledge within our classification pipeline we gained a significant improvement in all settings. In most cases, the accuracy was similar to when using the golden standard graphs that are manually annotated by crowd-sourcing.

References

M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, M. Galkin, S. Sharifzadeh, A. Fischer, V. Tresp, and J. Lehmann. Bringing light into the dark: A large-scale evaluation of knowledge graph embedding models under a unified framework. *arXiv preprint arXiv:2006.13365*, 2020a.

M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, S. Sharifzadeh, V. Tresp, and J. Lehmann. Pykeen 1.0: A python library for training and evaluating knowledge graph emebddings. *arXiv preprint arXiv:2007.14175*, 2020b.

S. Baier, Y. Ma, and V. Tresp. Improving visual relationship detection using semantic modeling of scene descriptions. In *International Semantic Web Conference*, pages 53–68. Springer, 2017.

S. Baier, Y. Ma, and V. Tresp. Improving information extraction from images with learned semantic models. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 5214–5218. AAAI Press, 2018.

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *Computing Research Repository*, arXiv:2005.14165, 2020. URL `https://arxiv.org/abs/2005.14165`.

L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, and S.-F. Chang. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4613–4623, 2019a.

T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 522–531, 2019b.

T. Chen, W. Yu, R. Chen, and L. Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019c.

N. Chinchor. MUC-3 linguistic phenomena test experiment. In *Third Message Uunderstanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991*, 1991. URL `https://www.aclweb.org/anthology/M91-1004`.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *European conference on computer vision*, pages 48–64. Springer, 2014.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://www.aclweb.org/anthology/N19-1423`.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

J. Gu, Z. Lu, H. Li, and V. O. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1154. URL `https://www.aclweb.org/anthology/P16-1154`.

P. Gupta, S. Rajaram, H. Schütze, and T. A. Runkler. Neural relation extraction within and across sentence boundaries. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6513–6520, 2019. doi: 10.1609/aaai.v33i01.33016513. URL `https://doi.org/10.1609/aaai.v33i01.33016513`.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*, 1992. URL `https://www.aclweb.org/anthology/C92-2082`.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori. Learning structured inference neural networks with label relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2960–2968, 2016.

H. Hu, Z. Deng, G.-T. Zhou, F. Sha, and G. Mori. Labelbank: Revisiting global perspectives for semantic segmentation. *arXiv preprint arXiv:1703.09891*, 2017.

C. Jiang, H. Xu, X. Liang, and L. Lin. Hybrid knowledge routed modules for large-scale object detection. *arXiv preprint arXiv:1810.12681*, 2018.

K. Kato, Y. Li, and A. Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

R. Koncel-Kedziorski, D. Bekal, Y. Luan, M. Lapata, and H. Hajishirzi. Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*, 2019.

R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

J. Piaget. *Langage et pensée chez l'enfant*. Delachaux et Niestlé, 1923.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

L. F. R. Ribeiro, M. Schmitt, H. Schütze, and I. Gurevych. Investigating pretrained language models for graph-to-text generation. *Computing Research Repository*, arXiv:2007.08426, 2020. URL `https://arxiv.org/abs/2007.08426`.

A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*, pages 4967–4976, 2017.

M. Schmitt, S. Sharifzadeh, V. Tresp, and H. Schütze. An unsupervised joint system for text generation from knowledge graphs and semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7117–7130, 2020.

S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and 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*, pages 70–80, 2015.

S. Sharifzadeh, S. Moayed Baharlou, M. Berrendorf, R. Koner, and V. Tresp. Improving visual relation detection using depth maps. *arXiv preprint arXiv:1905.00966*, 2019.

S. Sharifzadeh, S. M. Baharlou, and V. Tresp. Classification by attention: Scene graph classification with prior knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

K. K. Singh, S. Divvala, A. Farhadi, and Y. J. Lee. Dock: Detecting objects by transferring common-sense knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 492–508, 2018.

K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019.

V. Tresp, S. Sharifzadeh, and D. Konopatzki. A model for perception and memory. 2019.

V. Tresp, S. Sharifzadeh, D. Konopatzki, and Y. Ma. The tensor brain: Semantic decoding for perception and memory. *arXiv preprint arXiv:2001.11027*, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018.

C. Wu, I. Lenz, and A. Saxena. Hierarchical semantic labeling for task-relevant rgb-d perception. In *Robotics: Science and systems*, 2014.

D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017.

Y. Yaghoobzadeh, H. Adel, and H. Schütze. Noise mitigation for neural entity typing and relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1183–1194, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/E17-1111`.

J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–685, 2018.

R. Yu, A. Li, V. I. Morariu, and L. S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

A. Zareian, S. Karaman, and S.-F. Chang. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision*, pages 606–623. Springer, 2020a.

A. Zareian, H. You, Z. Wang, and S.-F. Chang. Learning visual commonsense for robust scene graph generation. *arXiv preprint arXiv:2006.09623*, 2020b.

R. Zellers, M. Yatskar, S. Thomson, and Y. Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.

R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.

H. Zhang, Z. Kyaw, S. Chang, and T. Chua. Visual translation embedding network for visual relation detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3107–3115. IEEE Computer Society, 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.331.

URL https://doi.org/10.1109/CVPR.2017.331.