Visually Grounded Commonsense Knowledge Acquisition

Yuan Yao[1], Tianyu Yu[2], Ao Zhang[4], Mengdi Li[5], Ruobing Xie[6], Cornelius Weber[4], Zhiyuan Liu[1*], Hai-Tao Zheng[2,3*], Stefan Wermter[5], Tat-Seng Chua[4], Maosong Sun[1]

[1]Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University, Beijing, China
[2]Shenzhen International Graduate School, Tsinghua University
[3]Peng Cheng Laboratory
[4]School of Computing, National University of Singapore, Singapore
[5]Department of Informatics, University of Hamburg, Hamburg, Germany
[6]WeChat AI, Tencent
yaoyuanthu@163.com

Abstract

Large-scale commonsense knowledge bases empower a broad range of AI applications, where the automatic extraction of commonsense knowledge (CKE) is a fundamental and challenging problem. CKE from text is known for suffering from the inherent sparsity and reporting bias of commonsense in text. Visual perception, on the other hand, contains rich commonsense knowledge about real-world entities, e.g., (*person*, `can_hold`, *bottle*), which can serve as promising sources for acquiring grounded commonsense knowledge. In this work, we present CLEVER, which formulates CKE as a distantLy supErVised multi-instancE leaRning problem, where models learn to summarize commonsense relations from a bag of images about an entity pair without any human annotation on image instances. To address the problem, CLEVER leverages vision-language pre-training models for deep understanding of each image in the bag, and selects informative instances from the bag to summarize commonsense entity relations via a novel contrastive attention mechanism. Comprehensive experimental results in held-out and human evaluation show that CLEVER can extract commonsense knowledge in promising quality, outperforming pre-trained language model-based methods by 3.9 AUC and 6.4 mAUC points. The predicted commonsense scores show strong correlation with human judgment with a 0.78 Spearman coefficient. Moreover, the extracted commonsense can also be grounded into images with reasonable interpretability. The data and codes can be obtained at https://github.com/thunlp/CLEVER.

Introduction

Providing machines with commonsense knowledge is a longstanding goal of artificial intelligence (Davis, Shrobe, and Szolovits 1993). Tremendous efforts have been devoted to building commonsense knowledge bases (KBs) (Liu and Singh 2004; Speer, Chin, and Havasi 2017; Sap et al. 2019), which have facilitated various important applications in both computer vision (Wu et al. 2017; Narasimhan, Lazebnik, and Schwing 2018; Gu et al. 2019; Gardères et al. 2020) and natural language processing (Zhou et al. 2018; Wu et al. 2020; Lv et al. 2020). However, most commonsense KBs are manually curated, which greatly limits their coverage and scale.

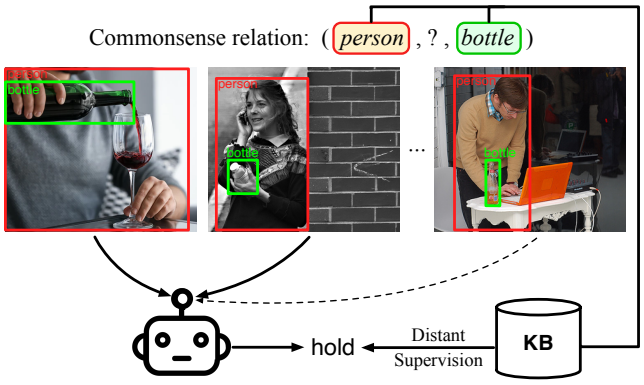


Figure 1: Visually grounded commonsense knowledge acquisition as a distantly supervised multi-instance learning problem. Given an entity pair and associated images, our model first understands entity interactions in each image, and then selects informative ones (solid line) to summarize the commonsense relations.

This paper studies the fundamental and challenging problem of commonsense knowledge extraction (CKE), which aims to extract plausible commonsense interactions between entities, e.g., (*person*, `can_hold`, *bottle*). Previous works have attempted to extract commonsense knowledge from plain text (Li et al. 2016) or pre-trained language models (PLMs) (Petroni et al. 2019; Bosselut et al. 2019). However, there is a growing consensus that obvious commonsense is rarely reported in text (Gordon and Van Durme 2013; Paik et al. 2021), and commonsense in PLMs suffers from low consistency and significant reporting bias (Shwartz and Choi 2020; Zhou et al. 2020; Elazar et al. 2021). There is also widespread doubt whether learning purely from the surface text forms can lead to real understanding of common-

*Corresponding authors: Z.Liu (liuzy@tsinghua.edu.cn), H.Zheng (zheng.haitao@sz.tsinghua.edu.cn)
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

sense meanings (Bender and Koller 2020).

Visual perceptions (e.g., images), on the other hand, contain rich commonsense knowledge about real-world entities that can be consistently grounded. According to our statistics, 83% of the triplets in visual relation learning datasets cannot be found in ConceptNet,[1] indicating a promising direction for CKE from image data. However, most existing image-based CKE methods are either confined to restricted interaction types (e.g., spatial or partonomy relations) (Chen, Shrivastava, and Gupta 2013; Collell, Van Gool, and Moens 2018; Xu, Lin, and Zhu 2018) or require extensive human annotation (Vedantam et al. 2015).

In this work, we present CLEVER, which formulates CKE as a distantLy supErVised multi-instancE leaRning problem (Dietterich, Lathrop, and Lozano-Pérez 1997), where models learn to summarize general commonsense relations of an entity pair from a bag of images, as shown in Figure 1. The commonsense relation labels are automatically created by aligning relational facts in existing KBs to image bags to provide distantly supervised learning signals. In this way, commonsense learning can easily scale up in general domain without costly manual image annotation.

To extract commonsense facts about a pair of query entities, models need to first understand their semantic interactions in each image of the bag, and then select informative ones (i.e., images that express interactions of interest between query entities) to synthesize the commonsense relations. However, our pilot experiments show that existing multi-instance learning methods cannot serve the task well, due to the complexity of real-world commonsense relations. Therefore, we propose a dedicated framework that models image-level entity interactions via vision-language pre-training (VLP) models, and selects meaningful images to summarize bag-level commonsense relations via a novel contrastive attention mechanism.

Comprehensive experimental results in held-out and human evaluation show that CLEVER can extract commonsense knowledge in promising quality, outperforming PLM-based approaches by 3.9 AUC and 6.4 mAUC points. The predicted commonsense scores show strong correlation with human judgment, achieving 0.78 Spearman's rank correlation coefficient. Moreover, the extracted commonsense can also be grounded into images with reasonable interpretability. Compared with PLM-based methods that produce commonsense purely based on text surface forms in a black-box fashion, the interpretability of CLEVER can be leveraged to provide supporting evidence for commonsense knowledge in KBs, which can be useful for downstream applications.

Our contributions are summarized as fourfold: (1) We propose to formulate CKE as a distantly supervised multi-instance learning problem, which can easily scale up for commonsense relations in a general domain without manual image annotation. (2) We conduct extensive experiments on existing and adapted CKE methods from different data sources, showing their effectiveness and limitations. (3) We present a dedicated CKE framework that integrates VLP models with a novel contrastive attention mechanism to deal with complex commonsense relation learning. (4) We conduct comprehensive experiments which demonstrate the effectiveness of the proposed framework.

[1]We randomly sample 200 distinct relational triplets from Visual Genome (Krishna et al. 2017) and manually verify if the triplet or its variations are included in the ConceptNet.

Related Work

Knowledge Bases. Large-scale knowledge bases (KBs) that store abundant structured human knowledge facilitate various AI applications. Many efforts have been devoted to building KBs of different knowledge types, including linguistic knowledge (Miller 1994), world knowledge (Bollacker et al. 2008) and commonsense knowledge (Liu and Singh 2004; Speer, Chin, and Havasi 2017; Sap et al. 2019). However, existing KBs are mainly constructed with human annotation, which greatly limits their coverage and scale.

Commonsense Knowledge Acquisition. To acquire commonsense knowledge, some works attempt to learn from internal structures of existing triplets (Speer, Havasi, and Lieberman 2008; Malaviya et al. 2020). However, these models usually suffer from the data sparsity of existing KBs. A more promising direction is to extract the commonsense contained in external data, i.e., commonsense knowledge extraction (CKE). Previous efforts in CKE can be divided into three categories according to the knowledge sources, including text-based, PLM-based and image-based models.

(1) *Text-based methods.* Early works attempt to extract commonsense from text (Angeli and Manning 2013; Li et al. 2016). However, CKE from text endures inherent reporting bias (Gordon and Van Durme 2013), i.e., people rarely state the obvious commonsense facts in text, making text not an ideal commonsense knowledge source. (2) *PLM-based methods.* Since PLMs learn certain commonsense knowledge during pre-training, they can be probed or fine-tuned to generate commonsense knowledge (Petroni et al. 2019; Davison, Feldman, and Rush 2019; Bosselut et al. 2019). However, it has been found that the commonsense in PLMs suffers from both low consistency, where small changes in the query templates can lead to substantially different predictions (Zhou et al. 2020; Elazar et al. 2021), and significant bias where the commonsense predictions can greatly differ from human judgments (Shwartz and Choi 2020; Paik et al. 2021). (3) *Image-based methods.* Some works have explored CKE from images that contain rich grounded commonsense knowledge. Chen, Shrivastava, and Gupta (2013) learn partonomy (i.e, `part_of`) and taxonomy (i.e., `is_a`) commonsense from images. Yatskar, Ordonez, and Farhadi (2016); Xu, Lin, and Zhu (2018) extract spatial commonsense (e.g., `located_near`). Chao et al. (2015) learn unary affordance commonsense about entities. Vedantam et al. (2015); Chen et al. (2022) extract more general commonsense interactions based on human annotation. Sadeghi, Kumar Divvala, and Farhadi (2015) mine commonsense based on spatial consistency of entities. Different from previous works, we extract general type commonsense interactions between entities without human annotation or restricted assumptions about commonsense knowledge.

Scene Graph Generation. Understanding visual interac-

tions between objects also lies in the interest of scene graph generation (Krishna et al. 2017; Lu et al. 2016; Xu et al. 2017; Tang et al. 2020; Yao et al. 2021b,c; Zhang et al. 2022). Different from CKE which aims to summarize global commonsense relations between entities from a bag of images, the goal of scene graph generation is to identify the local relation in a specific image. Moreover, scene graph models usually require large amounts of image annotations, whereas the proposed distantly supervised CKE framework does not need annotated images.

World Knowledge Acquisition. The extraction of factual world knowledge, e.g., (*Bob Dylan*, `composer`, *Blowin' in the Wind*), is an important tool to supplement world knowledge bases. Most works in world knowledge acquisition focus on text as the knowledge source (Nguyen and Grishman 2015; Soares et al. 2019; Wu et al. 2019; Dong et al. 2020; Chen et al. 2021; Yao et al. 2019, 2021a; Zhang et al. 2021a), with some attempts in multimodal world knowledge acquisition (Wen et al. 2021). To alleviate human annotation, Mintz et al. (2009) propose distant supervision that aligns KBs to text to create noisy relation labels. Following works focus on dealing with the noise in distant supervision under the multi-instance learning formulation (Riedel, Yao, and McCallum 2010; Zeng et al. 2015; Liu et al. 2018). The most widely adopted method is the selective attention model (Lin et al. 2016) which selects high-quality instances in the bag based on the attention mechanism. In comparison, we aim to extract commonsense knowledge from bag of images. We find in our experiment that existing multi-instance learning models cannot serve the complex commonsense learning well, and therefore we propose a dedicated approach for the task.

Pilot Experiment and Analysis

To investigate the effectiveness and limitation of existing CKE methods, we first perform an empirical study of representative methods from different information sources, including text-based, PLM-based and image-based models.

Problem Definition. CKE aims to extract commonsense relational triplet (s, r, o), which depicts plausible interactions $r \in \mathcal{R}$ between entities (s, o). For example, (*person*, `can_hold`, *bottle*) reflects the commonsense knowledge that a person can hold a bottle. A special `NA` relation is also included, indicating no relation between the entity pair.

Benchmark Construction. We construct the CKE benchmark based on Visual Genome (Krishna et al. 2017), which contains relational triplets about entities from real-world image data. Specifically, we select distinct triplets with the top 100 entity types and relation types. For automatic held-out evaluation (Mintz et al. 2009), we split the triplets into disjoint training, validation and test sets. Each entity pair is associated with Visual Genome images that contain the entities. The training/validation/test data contains 13,780/1,166/3,496 commonsense facts, 6,443/678/1,964 entity pairs, and 55,911/5,224/13,722 images respectively.

Existing CKE Models. We select representative CKE models for empirical study. (1) *Text-based models.* We adopt RTP (Schuster et al. 2015), a widely used triplet parser, which extracts commonsense triplets from captions based on dependency trees. We extract triplets from Conceptual Caption (Sharma et al. 2018) containing 3M captions, and obtain the confidence of the global triplets according to their frequency in the caption data. (2) *PLM-based models.* We adopt LAMA (Petroni et al. 2019) that probes knowledge in BERT by filling the prompting template containing the query entity pair and the masked relation (e.g., "*person* `[MASK]` *bottle*").[2] Following Lin et al. (2020), we further fine-tune the model based on the same prompts using the triplets in the training set to better learn the commonsense knowledge. Following Peng et al. (2020), we also adopt a vanilla fine-tuned BERT model which predicts relations based on the entity names using `[CLS]` token.

Multi-instance Learning for Image-based CKE. Intuitively, images are raw visual perceptions of rich real-world entity interactions, which can serve as a scalable and promising information source for CKE. However, most existing image-based CKE methods are either restricted in relation types, or require manual image annotation.

For general and scalable commonsense KB construction, it is desirable to extract general type commonsense knowledge from large-scale images without human annotation. To this end, we propose to formulate CKE as a *multi-instance learning* problem (Dietterich, Lathrop, and Lozano-Pérez 1997), where commonsense relation r between entities (s, o) is summarized from a bag of images $\mathcal{B}_{(s,o)} = \{v_i\}_{i=1}^{N}$ containing the entity pair. Inspired by Mintz et al. (2009), we align existing commonsense KBs to image bags to provide distantly supervised learning signals. Specifically, the image bag $\mathcal{B}_{(s,o)}$ is labeled with the relation r between (s, o) in the KB, assuming that at least a subset of images in the bag expresses the triplet (s, r, o), and there might be some images in the bag that do not express the triplet. To extract the commonsense triplet, models need to first understand the entity interactions in each image of the bag, and then select the meaningful ones to synthesize the commonsense relations.

We note some works exploring problems in similar formulation in world knowledge extraction from text. To investigate the effectiveness of existing multi-instance learning methods for image-based CKE, we adapt representative approaches that select and summarize bag of instances using average pooling (Lin et al. 2016), at-least-one strategy (Zeng et al. 2015), or attention mechanism (Lin et al. 2016).

Specifically, given a triplet (s, r, o), we first select a bag of images containing the query entity pair. In practice, the number of candidate images can be large (e.g., $\sim$1,000), while only a small portion reflects entity interactions. To compose a proper size of image bags, inspired by Zellers et al. (2018), we select images with top spatial overlaps (i.e., intersection over union in pixels) of query entities, which are more likely to exhibit interactions. The query entity pair in each image of the bag is encoded into feature representations $\{\boldsymbol{v}_i\}_{i=1}^{N}$ using an adapted Neural Motif (Zellers et al. 2018) model, a widely used CNN-based entity pair encoder.

To obtain the bag representation $\boldsymbol{B}_{(s,o)}$, (1) *average pool-*

[2] We also experimented with masking the entities, and find that masking relations achieves better performance.

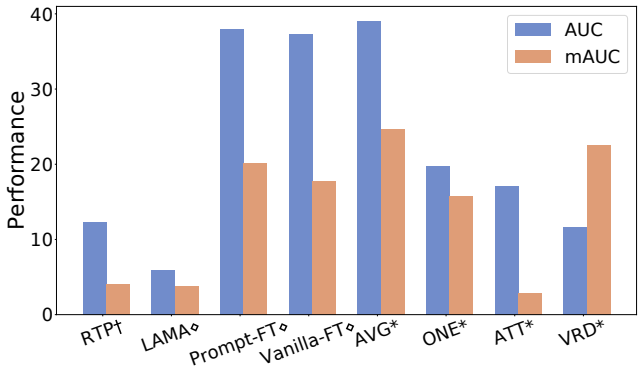


Figure 2: Results of CKE models from different information sources. †: text-based, ◇: PLM-based, *: image-based.

ing (AVG) (Lin et al. 2016) computes the mean of instance representations: $\boldsymbol{B}_{(s,o)} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{v}_i$; (2) *at-least-one strategy* (ONE) (Zeng et al. 2015) selects the most likely instance: $\boldsymbol{B}_{(s,o)} = \boldsymbol{v}_j$, where v_j achieves the highest score on the golden relation r^* of the given training triplet; (3) *attention mechanism* (ATT) (Lin et al. 2016) computes the weighted sum of instance representations: $\boldsymbol{B}_{(s,o)} = \sum_{i=1}^{N}\alpha_i\boldsymbol{v}_i$, where the attention weight is computed based on the golden relation query: $\alpha_i = \text{Softmax}_i(\boldsymbol{v}_i^\top \boldsymbol{r}^*)$. The bag representation $\boldsymbol{B}_{(s,o)}$ is optimized towards the golden label r^* via a softmax classifier. During inference, since the relation label is unknown, ONE and ATT enumerate relation queries for the corresponding relation prediction score.

In addition to multi-instance learning based approaches, we also adapt visual relation detection models for image-based CKE. To simulate a scalable scenario, we randomly select a moderate number (i.e., 100) of image-level annotations for each relation from Visual Genome, and train a Neural Motif (Zellers et al. 2018) model to predict the relation between an entity pair in specific images. During inference, the relation score of a bag is obtained by max pooling over relation scores of all images in the bag.

Results. Following previous works in knowledge acquisition (Zeng et al. 2015; Lin et al. 2016), to provide a rigorous evaluation, we draw the precision-recall curve of held-out triplet predictions, and report the area under curve (AUC). Besides the traditional micro result, we also report mAUC, the area under the macro curve (i.e., the average curve of different relations) to evaluate the performance on long-tail relations. From Figure 2 we have the following observations:

(1) Text-based method (RTP) and knowledge probing from PLMs (LAMA) struggle on CKE. The reason is the inherent lack of commonsense knowledge in text, and the models are not fine-tuned for the task. Further fine-tuning PLMs (Prompt-FT and Vanilla-FT) on the task can boost the performance to achieve a strong result.

(2) Visual perceptions from images can provide rich information for commonsense knowledge acquisition. Based on a relatively proper summarization approach (AVG), multi-instance learning-based models on images achieve the best results over all existing CKE models.

(3) Multi-instance learning formulation is necessary for scalable image-based CKE in open-domain. Adapted image-level visual relation detection models (VRD) do not perform well on CKE, despite more image-level relation annotations used (e.g., 100 image-level annotations per relation).

(4) Simple adaptation of existing multi-instance learning approaches cannot serve CKE well. The overall performance is still not satisfactory for all models. Notably, despite their competitive performance in world knowledge acquisition from text, ONE and ATT perform poorly on CKE. The reason is that compared with the relation schemes of world knowledge, commonsense relations exhibit higher complexity, where fine-grained relations with overlapping semantics (e.g., `stand_on` and `walk_on`), and hyponym-hypernym conflicts (e.g., `stand_on` and `on`) frequently occur. Compared with AVG, the *golden-query-only* problem of ONE and ATT hinders them from distinguishing complex commonsense relations. We refer readers to the methodology section for a more detailed discussion on the problem.

Methodology

The pilot experiment results show that dedicated approaches need to be developed to address the unique challenges of commonsense knowledge acquisition. Essentially, due to the complexity of commonsense relations, multi-instance learning based CKE presents challenges on two levels: (1) on the image level, models need to first understand complex entity interactions in each image, (2) on the bag level, models are required to select informative instances to summarize the fine-grained commonsense relations between the entities. We present a dedicated model for CKE from images, as shown in Figure 3, which (1) achieves deep understanding of the image-level interactions between entities through powerful vision-language pre-training (VLP) models, and (2) selects meaningful images to summarize bag-level commonsense relations via a contrastive attention mechanism.

Vision-language Pre-training Models for Image-level Entity Interaction Understanding. Recently VLP models have pushed forward the state-of-art of many multimodal tasks in a foundation role (Bommasani et al. 2021), such as visual question answering and visual grounding. However, few works have explored leveraging VLP methods to model complex visual relations for entity pairs. We show that pre-trained Transformers can serve as powerful foundation models to resolve complex image-level entity interactions.

Given a query entity pair (s, o) and the associated image bag $\mathcal{B}_{(s,o)} = \{v_i\}_{i=1}^{N}$, each query entity pair instance in the bag is encoded into deep representations $\boldsymbol{v}_i$ via detector-based VLP models. In this work, we adopt VinVL (Zhang et al. 2021b), a state-of-the-art VLP model as the encoder. Specifically, the query and context entities in each image are first encoded by object detectors to obtain a series of visual features $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\}$. The visual features and token embeddings of entity tags $\{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}$ are then fed into pre-trained Transformers to obtain deep multimodal hidden representations $\{\boldsymbol{h}_u^1, \boldsymbol{h}_u^2, \ldots, \boldsymbol{h}_u^n, \boldsymbol{h}_t^1, \boldsymbol{h}_t^2, \ldots, \boldsymbol{h}_t^n\}$. The image-level entity pair representation is obtained by the concatenation of visual and text hidden representations: $\boldsymbol{v}_i =$

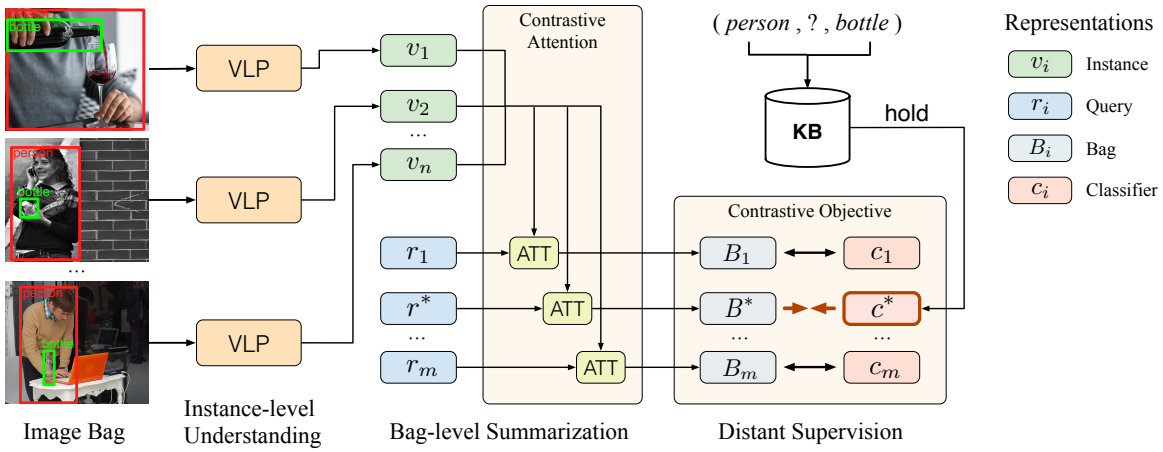


Figure 3: The CLEVER framework for visually grounded commonsense knowledge acquisition. Given a bag of images about an entity pair, our model leverages VLP models for image-level entity interaction understanding, and selects informative images to summarize bag-level commonsense relations via a contrastive attention mechanism.

$[\boldsymbol{h}_u^s; \boldsymbol{h}_u^o; \boldsymbol{h}_t^s; \boldsymbol{h}_t^o]$.

Despite the simplicity, the approach exhibits three important advantages in image-level entity interaction modeling: (1) The messages of entities (including query and context entities) are fused through multiple self-attention layers in Transformers to help model complex entity interactions. (2) Visual and textual information of entities are fused into deep multimodal representations. (3) Pre-trained deep vision-language representations are utilized to facilitate commonsense understanding.

Contrastive Attention Mechanism for Bag-level Commonsense Summarization. From the pilot experimental results, we observe that the complexity of commonsense relations (e.g., overlapping semantics and hyponym-hypernym conflicts) makes the relation boundaries hard to distinguish by existing multi-instance learning methods. In particular, despite its success in world knowledge acquisition from text, attention mechanism (ATT) (Lin et al. 2016) performs poorly on CKE. Here we identify that *golden-query-only* is the key limitation of ATT in CKE, and show that by making the attention mechanism *contrastive* over golden relation and other negative relations, the boundaries of complex commonsense relations can be effectively distinguished to achieve significantly better CKE performance.

We begin by discussing the *golden-query-only* problem in ATT. During ATT training, the bag representation $\boldsymbol{B}_{(s,o)}$ is static for the prediction of different relations, and computed only based on the golden relation query. However, during inference, since the golden relation is unknown, all possible relations need to be enumerated to query the bag to predict the corresponding relation score. The golden-query-only problem leads to a lack of effective supervision for the bag representations (and relation scores) of other negative relations, resulting in indistinguishable negative bag representations from the golden ones.

To address the problem, we present a novel *contrastive attention mechanism* that imposes contrastive supervision for golden and negative bag representations and relation scores. Specifically, for the prediction of *each relation* $r_i \in \mathcal{R}$, a relation-aware bag representation $\boldsymbol{B}_{(s,r_i,o)}$ is obtained by a weighted sum of instance representations, where the attention weights are computed using the corresponding relation query $\boldsymbol{r}_i$ as follows:

$$\boldsymbol{B}_{(s,r_i,o)} = \sum_{j=1}^{N} \alpha_j^{r_i} \boldsymbol{v}_j, \tag{1}$$

$$\alpha_j^{r_i} = \text{Softmax}_j(\boldsymbol{v}_j^\top \boldsymbol{r}_i). \tag{2}$$

The bag representations are optimized via a contrastive InfoNCE loss (Oord, Li, and Vinyals 2018) as follows:

$$\mathcal{L} = -\log \frac{\exp(\boldsymbol{c}^{*\top} \boldsymbol{B}_{(s,r^*,o)})}{\sum_i \exp(\boldsymbol{c}_i^\top \boldsymbol{B}_{(s,r_i,o)})}, \tag{3}$$

where $\boldsymbol{c}_i$ is the classifier embedding of r_i. In this way, the contrastive attention imposes clear boundaries between the bag representations of golden and negative relations to deal with the summarization of complex commonsense relations. The contrastive attention can also be viewed as a kind of cross-attention (Vaswani et al. 2017) between relation queries and image instances, which can potentially benefit from multi-layer stacking. We leave it for future work.

Integrating Multi-source Information for CKE. Intuitively, multiple heterogeneous data sources can provide complementary information for commonsense learning. We show that this complementarity can be leveraged by a simple ensemble of models from each information source, where the aggregated triplet score is a weighted sum of the prediction score from each source.

Experiments

In this section, we empirically assess the effectiveness of the proposed model. We refer readers to the appendix for implementation details.

Source	Method	AUC	F1	P@2%	mAUC	mF1	mP@2%
-	Random	1.76	3.51	1.71	2.04	5.13	1.94
Text	RTP (Schuster et al. 2015)	12.30	23.67	16.65	4.10	8.62	7.34
PLM	LAMA (Petroni et al. 2019)	5.97	14.11	12.80	3.84	3.59	5.59
	Vanilla-FT (Peng et al. 2020)	37.28	47.06	44.21	17.75	30.98	17.34
	Prompt-FT (Lin et al. 2020)	37.99	44.43	41.69	20.15	35.37	19.81
Image	AVG (Lin et al. 2016)	39.04	47.49	44.34	24.73	41.07	20.83
	ONE (Zeng et al. 2015)	19.69	31.10	25.20	15.70	30.40	12.82
	ATT (Lin et al. 2016)	17.13	28.37	25.07	2.91	6.09	2.20
	CLEVER (Ours)	41.92	48.96	45.84	26.57	**43.62**	22.02
All	Ensemble (Ours)	**45.68**	**49.93**	**47.09**	**27.38**	43.13	**22.80**

Table 1: Experimental results of CKE methods from different information sources. The best results are highlighted in bold, and best single model results are underlined.

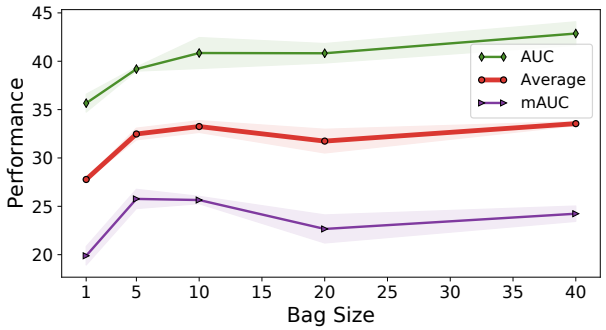


Figure 4: Experimental results of our model with different bag sizes. We report AUC, mAUC and their average.

Method	AUC	F1	mAUC	mF1
CLEVER	**41.92**	**48.96**	**26.57**	**43.62**
VLP → CNN	39.86	48.48	24.99	41.51
CST-ATT → AVG	39.95	47.73	25.56	41.51
CST-ATT → ONE	16.16	26.47	5.17	13.00
CST-ATT → ATT	16.07	25.59	2.14	4.87

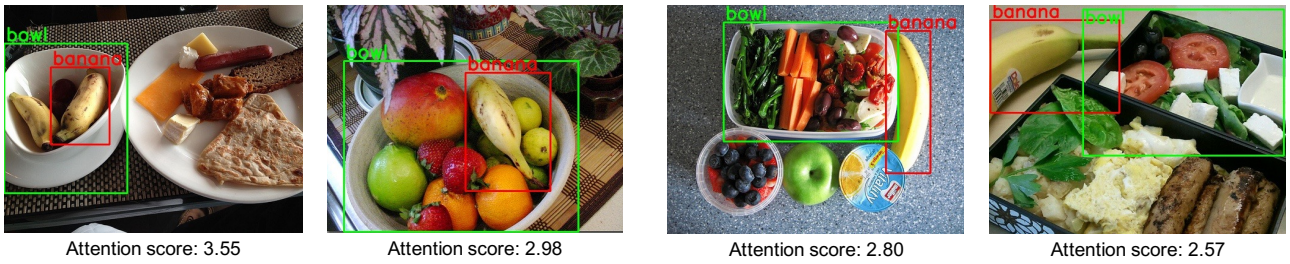
Table 2: Ablation study on the instance encoder and the commonsense summarization method.

Experimental Settings. (1) *Benchmark and baselines.* We perform experiments on the CKE benchmark constructed from Visual Genome as described in the pilot experiment section, and compare to strong baselines from different information sources. We also include a random baseline that randomly predicts relations for entity pairs. For multi-source information integration, we ensemble CLEVER, RTP and Vanilla-FT. (2) *Evaluation metrics.* To provide multi-dimensional evaluation, we also report the maximum F1 on curve, and precision@K% (P@K%) triplet prediction.

Main Results. From the experimental results in Table 1, we have the following observations: (1) CLEVER consistently achieves the best results among all baseline models in both micro and macro metrics. Specifically, CLEVER improves the performance of image-based models, and significantly outperforms previous best PLM-based results by 3.9 AUC and 6.4 mAUC points. The results show that CLEVER can extract commonsense knowledge from visual perceptions with promising quality. (2) Ensemble multi-source information further improves the performance over single-source models. This indicates that CKE can benefit from exploiting complementary information in different sources.

Human Evaluation. In addition to the held-out evaluation, we also perform a human evaluation on top predictions. We select models that achieve the best micro performance on each source, including RTP, Vanilla-FT and CLEVER. Specifically, for each model, we sample from the top 10% triplet predictions in a 1:50 ratio, resulting in 1,200 triplets for human evaluation. Each triplet is labeled by three independent annotators to decide the commonsense score: implausible (0), plausible but rare (1), common (2). We report the locally averaged triplet commonsense score given by human annotators in Figure 6. We can observe that triplets extracted by CLEVER are assigned with significantly higher commonsense scores in most cases. In addition, the commonsense scores of CLEVER achieve a strong 0.78 Spearman's rank correlation coefficient with human score, which shows that commonsense scores from our model can be well aligned to human judgments. The reason is that the contrastive attention mechanism can implicitly leverage the redundancy of instances to reflect the commonsense degree, where multiple informative instances in a bag can contribute to higher commonsense scores.

Interpretability. In addition to the competitive performance, a crucial advantage of CLEVER is that the extracted commonsense knowledge can be grounded into visual perceptions through contrastive attention scores over image instances. As shown in Figure 5, informative images are assigned with larger attention scores for commonsense learning. Compared with PLM-based approaches that produce commonsense knowledge purely based on correlations be-



(a) Informative Instances (b) Uninformative instances

Figure 5: Unnormalized attention scores of the extracted commonsense triplet (*banana*, `in`, *bowl*) over several images in a bag.

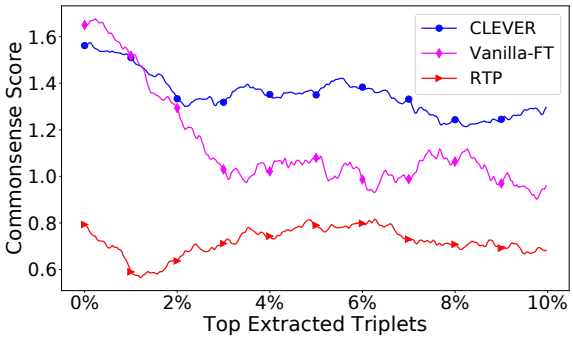


Figure 6: Human evaluation results on top extracted triplets.

tween text tokens in a black-box fashion, CLEVER enables trustworthy commonsense knowledge acquisition with better interpretability in the extraction process. From an application perspective, the selected informative images can also serve as supporting evidence for the extracted triplets in KBs for better knowledge utilization in downstream applications.

Ablation Study. We perform an ablation study by replacing the VLP encoder with the CNN-based encoder, and replacing the contrastive attention mechanism with existing multi-instance learning methods respectively. From the results in Table 2, we can see that both components contribute to the final results. The results show that image-level entity interaction understanding and bag-level summarization are both important for good CKE performance.

Effect of Bag Size. Intuitively, multiple images in a bag can provide diverse and complementary information about an entity pair for robust commonsense learning. To investigate the effect of bag size, we perform experiments on CLEVER with different bag sizes. From the results in Figure 4, we observe that: (1) A certain number of images is necessary to learn the commonsense interactions. The performance drops significantly when very small bag sizes are used. (2) The performance improvement is not significant when the bag size grows larger than 20. We hypothesize the reason is that although a larger bag provides richer commonsense information, it also challenges the model with more noisy instances. Therefore, more advanced methods need to be developed to better exploit the rich information in larger image bags, which we leave for future work.

Method	AUC	F1	P@2%	mAUC	mF1	mP@2%
Random	41.3	45.7	43.0	23.7	38.0	21.7
CLIP	**44.2**	47.3	44.5	24.0	38.5	**22.6**
Overlap	41.9	**49.0**	**45.8**	**26.6**	**43.6**	22.0

Table 3: Image sampling strategies for bag construction.

Type	Examples
I	(*woman*, `hold`, *umbrella*), (*horse*, `pull`, *person*), (*skateboard*, `under`, *man*), (*flower*, `near`, *fence*), (*girl*, `wear`, *glove*), (*truck*, `has`, *handle*)
II	(*snow*, `cover`, *tire*), (*cow*, `with`, *nose*), (*flower*, `in`, *mountain*), (*wire*, `in`, *building*), (*logo*, `printed_on`, *train*), (*boy*, `hold`, *pillow*)
III	(*clock*, `has`, *flower*), (*boat*, `behind`, *car*), (*sheep*, `behind`, *bench*), (*tail*, `on`, *book*)

Table 4: Extracted commonsense triplet examples in different types. I: Reasonable triplets unseen during training, II: novel facts for both Visual Genome and ConceptNet (i.e., newly discovered), III: Uncommonly observed facts.

Effect of Instance Sampling Strategy for Bag Construction. Given the typically large number of open images containing an entity pair, it is desirable to select instances that are likely to express commonsense interactions at low costs to construct the bag. Besides the spatial overlap strategy, we experiment with another two sampling strategies: (1) *Random sampling.* Random candidate images are selected to compose the bag. (2) *CLIP-based sampling.* A text query is constructed for the entity pair as: "*s* has some relation with *o*". Then we encode the text query and image candidates using CLIP (Radford et al. 2021), and select the images with top similarity scores. We can see from Table 3 that: (1) Entity interaction priors from CLIP and spatial overlap help select informative images for bag construction. (2) CLIP does not show significant advantage over spatial overlap. The reason is that spatial overlap incorporates more inductive bias for entity pair interactions, while CLIP is optimized to handle general sentences. Therefore, we choose spatial overlap for bag construction due to its simplicity and efficiency.

Case Study. We provide examples of the extracted triplets

from CLEVER in Table 4. We can see that our model can extract reasonable commonsense knowledge unseen during training, and most importantly, novel facts to supplement commonsense KBs. We note that our model can sometimes produce uncommonly observed facts from accidental scene images. We refer readers to the appendix for the supporting images selected by our model for examples in type III.

Conclusion

In this work, we propose a novel formulation for commonsense knowledge acquisition as an image-based distantly supervised multi-instance learning problem. We present a dedicated framework that achieves deep image-level understanding via vision-language pre-training models, and bag-level summarization via a contrastive attention mechanism. Comprehensive experiments show the effectiveness of our framework. In the future, we will explore more advanced multi-instance learning approach, and acquire visual commonsense knowledge in more complex forms and types.

Acknowledgements

This work is funded by the Natural Science Foundation of China (NSFC 62061136001), the German Research Foundation (DFG TRR-169) in Project Crossmodal Learning, National Natural Science Foundation of China (Grant No.62276154), AMiner.Shenzhen SciBrain Fund, Shenzhen Science and Technology Innovation Commission (Research Center for Computer Network (Shenzhen) Ministry of Education), Beijing Academy of Artificial Intelligence (BAAI), the Natural Science Foundation of Guangdong Province (Grant No. 2021A1515012640), Basic Research Fund of Shenzhen City (Grant No. JCYJ20210324120012033 and JSGG20210802154402007), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (Grant No. HW2021008).

For author contributions, Yuan Yao designed the framework and experiments, and wrote the paper. Tianyu Yu conducted the experiments. Ao Zhang, Mengdi Li, Ruobing Xie, Cornelius Weber, Zhiyuan Liu, Hai-Tao Zheng, Stefan Wermter, Tat-Seng Chua and Maosong Sun provided valuable suggestions.

References

Angeli, G.; and Manning, C. D. 2013. Philosophers are mortal: Inferring the truth of unseen facts. In *Proceedings of CoNLL*, 133–142.

Bender, E. M.; and Koller, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of ACL*, 5185–5198.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD*, 1247–1250.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of ACL*, 4762–4779.

Chao, Y.-W.; Wang, Z.; Mihalcea, R.; and Deng, J. 2015. Mining semantic affordances of visual object categories. In *Proceedings of ICCV*, 4259–4267.

Chen, T.; Shi, H.; Tang, S.; Chen, Z.; Wu, F.; and Zhuang, Y. 2021. CIL: Contrastive Instance Learning Framework for Distantly Supervised Relation Extraction. In *Proceedings of ACL*, 6191–6200.

Chen, X.; Shrivastava, A.; and Gupta, A. 2013. NEIL: Extracting visual knowledge from web data. In *Proceedings of ICCV*, 1409–1416.

Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022. Hybrid Transformer with Multi-Level Fusion for Multimodal Knowledge Graph Completion. In *Proceedings of ACM SIGIR*, 904–915.

Collell, G.; Van Gool, L.; and Moens, M.-F. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Proceedings of AAAI*, volume 32.

Davis, R.; Shrobe, H.; and Szolovits, P. 1993. What is a knowledge representation? *AI magazine*, 14(1): 17–17.

Davison, J.; Feldman, J.; and Rush, A. 2019. Commonsense Knowledge Mining from Pretrained Models. In *Proceedings of EMNLP-IJCNLP*, 1173–1178.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2): 31–71.

Dong, B.; Yao, Y.; Xie, R.; Gao, T.; Han, X.; Liu, Z.; Lin, F.; Lin, L.; and Sun, M. 2020. Meta-information guided meta-learning for few-shot relation classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1594–1605.

Elazar, Y.; Kassner, N.; Ravfogel, S.; Ravichander, A.; Hovy, E.; Schütze, H.; and Goldberg, Y. 2021. Measuring and improving consistency in pretrained language models. *TACL*, 9: 1012–1031.

Gardères, F.; Ziaeefard, M.; Abeloos, B.; and Lecue, F. 2020. ConceptBert: Concept-Aware Representation for Visual Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 489–498.

Gordon, J.; and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 25–30.

Gu, J.; Zhao, H.; Lin, Z.; Li, S.; Cai, J.; and Ling, M. 2019. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of CVPR*, 1969–1978.

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

Li, X.; Taheri, A.; Tu, L.; and Gimpel, K. 2016. Commonsense knowledge base completion. In *Proceedings of ACL*, 1445–1455.

Lin, B. Y.; Lee, S.; Khanna, R.; and Ren, X. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of EMNLP*, 6862–6868.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, 2124–2133.

Liu, H.; and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4): 211–226.

Liu, T.; Zhang, X.; Zhou, W.; and Jia, W. 2018. Neural Relation Extraction via Inner-Sentence Noise Reduction and Transfer Learning. In *Proceedings of EMNLP*, 2195–2204.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *Proceedings of ECCV*, 852–869. Springer.

Lv, S.; Guo, D.; Xu, J.; Tang, D.; Duan, N.; Gong, M.; Shou, L.; Jiang, D.; Cao, G.; and Hu, S. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of AAAI*, volume 34, 8449–8456.

Malaviya, C.; Bhagavatula, C.; Bosselut, A.; and Choi, Y. 2020. Commonsense knowledge base completion with structural and semantic context. In *Proceedings of AAAI*, volume 34, 2925–2933.

Miller, G. A. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ALC-IJCNLP*, 1003–1011.

Narasimhan, M.; Lazebnik, S.; and Schwing, A. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *NeurIPS*, 31.

Nguyen, T. H.; and Grishman, R. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, 39–48.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paik, C.; Aroca-Ouellette, S.; Roncone, A.; and Kann, K. 2021. The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In *Proceedings of EMNLP*, 823–835.

Peng, H.; Gao, T.; Han, X.; Lin, Y.; Li, P.; Liu, Z.; Sun, M.; and Zhou, J. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. In *Proceedings of EMNLP*, 3661–3672.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of EMNLP-IJCNLP*, 2463–2473.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*, 148–163. Springer.

Sadeghi, F.; Kumar Divvala, S. K.; and Farhadi, A. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of ICCV*, 1456–1464.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of AAAI*, volume 33, 3027–3035.

Schuster, S.; Krishna, R.; Chang, A.; Fei-Fei, L.; and Manning, C. D. 2015. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. In *Proceedings of the Fourth Workshop on Vision and Language*, 70–80.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2556–2565.

Shwartz, V.; and Choi, Y. 2020. Do neural language models overcome reporting bias? In *Proceedings of COLING*, 6863–6870.

Soares, L. B.; Fitzgerald, N.; Ling, J.; and Kwiatkowski, T. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of ACL*, 2895–2905.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*.

Speer, R.; Havasi, C.; and Lieberman, H. 2008. AnalogySpace: Reducing the Dimensionality of Common Sense Knowledge. In *Proceedings of AAAI*, volume 8, 548–553.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of CVPR*, 3716–3725.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

Vedantam, R.; Lin, X.; Batra, T.; Zitnick, C. L.; and Parikh, D. 2015. Learning common sense through visual abstraction. In *Proceedings of ICCV*, 2542–2550.

Wen, H.; Lin, Y.; Lai, T.; Pan, X.; Li, S.; Lin, X.; Zhou, B.; Li, M.; Wang, H.; Zhang, H.; et al. 2021. Resin: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of NAACL*, 133–143.

Wu, Q.; Shen, C.; Wang, P.; Dick, A.; and Van Den Hengel, A. 2017. Image captioning and visual question answering based on attributes and external knowledge. *TPAMI*, 40(6): 1367–1381.

Wu, R.; Yao, Y.; Han, X.; Xie, R.; Liu, Z.; Lin, F.; Lin, L.; and Sun, M. 2019. Open Relation Extraction: Relational Knowledge Transfer from Supervised Data to Unsupervised Data. In *Proceedings of EMNLP*, 219–228.

Wu, S.; Li, Y.; Zhang, D.; Zhou, Y.; and Wu, Z. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of ACL*, 5811–5820.

Xu, D.; Zhu, Y.; Choy, C. B.; and Fei-Fei, L. 2017. Scene graph generation by iterative message passing. In *Proceedings of ICCV*, 5410–5419.

Xu, F. F.; Lin, B. Y.; and Zhu, K. 2018. Automatic Extraction of Commonsense LocatedNear Knowledge. In *Proceedings of ACL*, 96–101.

Yao, Y.; Du, J.; Lin, Y.; Li, P.; Liu, Z.; Zhou, J.; and Sun, M. 2021a. CodRED: A Cross-Document Relation Extraction Dataset for Acquiring Knowledge in the Wild. In *Proceedings of EMNLP*, 4452–4472.

Yao, Y.; Ye, D.; Li, P.; Han, X.; Lin, Y.; Liu, Z.; Liu, Z.; Huang, L.; Zhou, J.; and Sun, M. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the ACL*, 764–777.

Yao, Y.; Zhang, A.; Han, X.; Li, M.; Weber, C.; Liu, Z.; Wermter, S.; and Sun, M. 2021b. Visual distant supervision for scene graph generation. In *Proceedings of the ICCV*, 15816–15826.

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2021c. CPT: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

Yatskar, M.; Ordonez, V.; and Farhadi, A. 2016. Stating the Obvious: Extracting Visual Common Sense Knowledge. In *Proceedings of NAACL*, 193–198.

Zellers, R.; Yatskar, M.; Thomson, S.; and Choi, Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of ICCV*, 5831–5840.

Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, 1753–1762.

Zhang, A.; Yao, Y.; Chen, Q.; Ji, W.; Liu, Z.; Sun, M.; and Chua, T.-S. 2022. Fine-Grained Scene Graph Generation with Data Transfer. In *Proceedings of ECCV*.

Zhang, K.; Yao, Y.; Xie, R.; Han, X.; Liu, Z.; Lin, F.; Lin, L.; and Sun, M. 2021a. Open Hierarchical Relation Extraction. In *Proceedings of NAACL*, 5682–5693.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of CVPR*, 5579–5588.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, 4623–4629.

Zhou, X.; Zhang, Y.; Cui, L.; and Huang, D. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of AAAI*, volume 34, 9733–9740.