Integrating Image-Based and Knowledge-Based Representation Learning

Ruobing Xie, Stefan Heinrich®, Zhiyuan Liu®, Cornelius Weber, Yuan Yao, Stefan Wermter®, and Maosong Sun

Abstract—A variety of brain areas is involved in language understanding and generation, accounting for the scope of language that can refer to many real-world matters. In this paper, we investigate how regularities among real-world entities impact emergent language representations. Specifically, we consider knowledge bases, which represent entities and their relations as structured triples, and image representations, which are obtained via deep convolutional networks. We combine these sources of information to learn representations of an image-based knowledge representation learning (IKRL) model. An attention mechanism lets more informative images contribute more to the image-based representations. Evaluation results show that the model outperforms all baselines on the tasks of knowledge representation learning (KRL) and refers to these concepts during communication in natural language learning.

Index Terms—Attention mechanisms and development, embodied cognition, generation of representation during development.

I. INTRODUCTION

MODELS for development and language acquisition in the human brain must consider the fact that a child has acquired several sensory and motor skills before it makes use of language. In fact, a child develops the ability to learn concepts from its perceptions and active interactions and refers to these concepts during communication in natural language [1]. For the human brain, it is overall seen plausible that concepts and thus language is represented in a network which is distributed over large parts of the cortex and comprises multiple sensory and motor areas that are not primarily language-related [2], [3]. Neural representations in the brain must consider the fact that a child has acquired several sensory and motor skills before it makes use of language. In fact, a child develops the ability to learn concepts from its perceptions and active interactions and refers to these concepts during communication in natural language [1]. For the human brain, it is overall seen plausible that concepts and thus language is represented in a network which is distributed over large parts of the cortex and comprises multiple sensory and motor areas that are not primarily language-related [2], [3]. Neural representations in the brain must consider the fact that a child has acquired several sensory and motor skills before it makes use of language. In fact, a child develops the ability to learn concepts from its perceptions and active interactions and refers to these concepts during communication in natural language [1]. For the human brain, it is overall seen plausible that concepts and thus language is represented in a network which is distributed over large parts of the cortex and comprises multiple sensory and motor areas that are not primarily language-related [2], [3]. Neural representations in...
allows integrating the rich information contained in images of entities. Information for an entity can be obtained from multiple images, each potentially providing different aspects of the appearances or functional characteristics (compare Fig. 2).

In this paper, we investigate the image-based KRL (IKRL) model,\(^1\) which utilizes the rich information in images by combining translation-based KRL models with brain-inspired visual representation learning. The image processing part of this model encodes the images in a deep neural network, where the first layers are taken from a pretrained AlexNet [14], followed by a trainable projection layer, which generates the vector representation of each image. Representations of multiple instances of an entity are then combined using an attention mechanism, which assigns high attention values to the most representative images of a given instance. Finally, the resulting image-based vector representation and a structure-based vector representation are adapted to jointly optimize a translation-based energy function. This way, the aspects of the different modalities shape the representation of both the entities and the relations.

We show that the IKRL model achieves state-of-the-art performance on KG completion and triple classification by integrating image information into structured knowledge representations. Furthermore, we present detailed analyses revealing the impact of attention in selecting informative images and the regularities underlying the formed representations. Our results are relevant for language learning in humans and in robots since language representations are constrained by relations between real-world entities, which are contained in knowledge bases or which exist as image similarities.

\(^1\) This paper is an extension of the Image-embodied Knowledge Representation Learning model proposed in [13].

II. RELATED WORK

A. Translation-Based Methods

Most currently available KRL methods build embeddings from the structured information such as triple facts. To measure the plausibility of a fact, the translation mechanism has been introduced and has achieved great success in KRL in recent years. These methods [11], [15]–[17] are inspired by translation patterns in the word representation learning field, such as “king”−“man”=“queen”−“woman” [12]. One of the successful translation-based methods are TransE [11], which embeds entities as well as relation into one low-dimensional continuous vector space. In this space, the relations describe translating operations between the head and tail entities and thus, TransE assumes that the embedding of a tail entity \(t\) is supposed to be close to \(h + r\).

In order to realize this relation, those embedding parameters will be optimized to minimize the following energy function of TransE:

\[
E(h, r, t) = ||h + r - t||
\]

under the condition that embedding vectors are normalized

\[
||h|| = ||t|| = 1
\]

in order to avoid the degenerate solution of them becoming zero. TransE is both effective and efficient, while the simple assumption may result in conflicts when modeling complicated entities and relations, such as 1-to-N, N-to-1, and N-to-N relations. To address this problem, several extensions of TransE have been proposed, which can be broadly divided into two categories.

Some extensions of TransE assign different roles to entities according to the relations involved. For instance, relations are expressed by hyperplanes in TransH [15], relations are expressed in different spaces than entities in TransR [16], or multiple representations of an entity are dynamically mapped to account for the diversity of relations in TransD [17].

Besides, there are also some works that model complicated relations by relaxing the over-strict translation assumption \(h + r = t\). TransM [18] associates each triple fact with a weight, which represents the degree of mapping, and assigns lower weights to complicated relations. TransF [19] proposes a flexible translation mechanism which only constrains the direction of \(h + r\) to be the same as \(t\), but allows flexible magnitude. TransA [20] proposes an adaptive metric approach for flexible representation learning. StarSpace [21] proposes a general-purpose embedding method which is capable of solving various problems, and achieves comparable performance with TransE on KG embedding.

There are also some works that learn knowledge representations via tensor factorization, such as Tucker [22] and RESCAL [23], [24]. Compared to tensor factorization-based methods, translation-based methods achieve both better performance and computation efficiency. Besides, translation-based methods are capable of explicitly modeling complex semantic relationships between entities using a translation mechanism.

The above-mentioned methods, however, use only the relational information from KGs. The structured information in KGs is usually over-simplified and incomplete, which will hurt the performance when the knowledge representation is applied to downstream tasks, such as KG completion and triple classification. In this paper, we propose to consider the important side information of entity images on the basis of TransE. It is in principle possible to use translation-based settings to combine representations obtained from multiple sources, such as images, structured KBs, or text.

B. Multisource Information Learning

In addition to structured triple facts in KGs, there are many other sources of information about entities and relations that can be incorporated to benefit KRL. Textual description, for example, can provide rich information about entities in and out of KGs. Jointly utilizing the multisource information is significant for KRL. To utilize rich textual information, Wang et al. [25] project both entities and words into a joint vector space with alignment models. Xie et al. [26] directly construct entity representations from entity descriptions, which is capable of modeling new entities. There are also other KRL methods utilizing additional information besides textual
descriptions as well. SSE [27] incorporates entity types in KRL and requires entities belonging to the same semantic category to stay close to each other in the embedding space. PTransE [28] introduces path-based TransE which learns representations of entities and relations considering relation paths. Wang et al. [29] utilized logical rules to benefit KRL by viewing inference as an integer linear programming problem.

As for visual information, multimodal representations based on words and images are widely used for various tasks like image-sentence ranking [30], metaphor identification [31], and visual QA [32]. However, it has not been fully explored how we can effectively incorporate image information into KRL. IKRL explicitly encodes visual information from images into KRL.

C. Vision-Based Structured Information Extraction

Recent years have witnessed the tremendous success of convolutional neural networks (CNNs) on various computer vision tasks such as object detection and image classification. LeNet [33] is the first successful application of CNN, which is designed for handwritten and machine-printed character recognition. Many models have been proposed to improve the performance of CNN on various computer vision tasks. AlexNet [34] proposes a deeper CNN architecture and achieves significant improvement on the task of image classification. VGG16 [35] further demonstrates that depth is critical for CNNs for good performance. GoogLeNet [36] introduces an inception module and replaces the fully connected layers with average pooling at the top of CNN, which substantially reduces the number of parameters. ResNet [37] proposes shortcut connections, and surpasses human-level performance of image classification on ImageNet.

An image or video also contains rich structured relations between objects. On the basis of the development of the CNN architecture, many models have been proposed recently to extract structured information from visual information. Yao and Li [38] regarded relations as hidden variables in visual relation detection. Visual relation extraction models can be divided into two categories. Those joint models consider a triple as a unique class [39], [40], while the separate models detect subjects, objects and predicates individually [41], [42]. VTransE [43] proposes a visual translation embedding model by utilizing the translation mechanism for visual relation detection from images. Shang et al. [44] proposed a visual relation detection model from videos, which consists of three components including object tracklet proposal, short-term relation prediction and greedy relational association. Lu et al. [42] further exploited language priors to boost visual relation extraction.

Leveraging KRL might benefit visual relation extraction. Incorporating structured facts extracted from visual information might also be conducive to KRL. However, in the current phase, the problem of visual relation extraction is still poorly understood, and the performance of these techniques needs to be improved before they can be directly applied in this paper presented here.

III. METHODOLOGY

We first introduce the terms and notations used in this paper. Knowledge facts are represented as triples, in the form of \((h, r, t) \in T\), which consist of a head entity \(h \in E\), a tail entity \(t \in E\) and a relation \(r \in R\). \(T\) describes the whole training set of triples, while \(E\) is the set of entities and \(R\) the set of relations, both in \(d_e\)-dimensional vector spaces \(\mathbb{R}^{d_e}\).

To include brain-inspired encoding and entity image information in KRL, we associate each entity with two types of representations. First, we define \(h_S, t_S\) as the structure-based representations (SBRs) of head and tail entities, which are trained with conventional KRL models. Second, we utilize a novel image-based representation (IBR) that is constructed from the corresponding images of entities, with head entities \(h_I\), and tail entities \(t_I\).

A. Architecture

In the overall architecture we integrate the SBR and the IBR into one coherent IKRL model and define a joint energy function as follows:

\[
E(h, r, t) = E_{SS} + E_{SI} + E_{IS} + E_{II}.
\]  

\[
E_{SS} = ||h_S + r - t_S||
\]

is identical to the energy function of TransE [11], which only depends on the SBRs. Analogously, \(E_{II} = ||h_I + r - t_I||\) captures IBRs that are learned from corresponding images. Both functions provide to learn the two kinds of entity representations in a relatively independent manner and will embed them into two different semantic spaces. In order to integrate the two kinds of entity representations, we introduce \(E_{SI} = ||h_S + r - t_I||\) and \(E_{IS} = ||h_I + r - t_S||\) to facilitate that both structure-based and IBRs are learned into the same semantic vector space. \(E_{SI}\) and \(E_{IS}\) can also benefit the SBRs by incorporating visual information. Note that the entity vectors \(h_S, h_I, t_S, t_I\) are normalized but the relation vectors are not. It is also possible to learn structure-based and IBRs for relations. However, this is not necessary since unlike entity representations, relation representations do not directly depend on image information. Besides, using shared relation representations as translations between two kinds of entity representations can also naturally help to integrate them into the same semantic space.

The overall architecture of the IKRL model is presented in Fig. 3. For the \(h_I\) and \(t_I\) entities, multiple images are considered to provide significant visual input and are processed as follows. First, every entity image is fed into a neural image decoder that is designed to construct the image representations in entity space. Second, an attention-based learning step calculates how the attention is distributed over different images in each entity. Finally, the aggregated IBRs are learned jointly with the SBRs under the overall energy function.

B. Image Encoder

Crucial inputs to the IKRL model are images since they potentially provide important aspects of the appearances as well as functional or behavior-related characteristics of the entities. Particularly because images can depict entities from vastly changing perspectives or contexts, multiple image instances \(I_k = \{img_1^{(k)}, img_2^{(k)}, \ldots, img_{n_k}^{(k)}\}\) are included for each entity \(e_k\).

The proposed image encoder consists of an image representation module and an image projection module in order to effectively encode the image information into knowledge representations. We utilize a deep CNN within the image representation module to extract visual features from images, and to construct image feature representations for each image. The
have more than one image in different poses and various scenarios. Visual information from images is intuitive but also noisy. It is essential but also challenging to select informative image representations for the corresponding entities. Simply summing up all the image representations may suffer from noises and loose detailed information. Instead, to construct the aggregated IBR for each entity from multiple instances, we propose an attention-based multi-instance learning method.

Humans are capable of selecting representative instances and ignoring irrelevant instances by an attention mechanism. The attention-based methods have been shown to be beneficial in automatically selecting informative instances from multiple candidates. It has been widely utilized in various fields, such as image classification [45], machine translation [46], and abstractive sentence summarization [47]. For example, machine translation and image captioning aim to generate parallel natural language descriptions for a source sentence/image. Instead of simply encoding the whole sentence/image, it is shown to be beneficial to select relevant parts from the source sentence/image to predict a target word using attention mechanisms [46], [48]. Utilizing attention mechanisms not only achieves better performance but also gives results that better agree with human intuition [46].

In IKRL, instance-level attention is obtained by jointly considering each image representation and the SBR of its corresponding entity. For the $i$th image representation $p^{(k)}_i$ of the $k$th entity, the attention is defined as follows:

$$\text{att}(p^{(k)}_i, e^{(k)}_S) = \frac{\exp(p^{(k)}_i \cdot e^{(k)}_S)}{\sum_{j=1}^{n} \exp(p^{(k)}_j \cdot e^{(k)}_S)}$$ (5)

where $e^{(k)}_S$ represents the SBR of the $k$th entity.

Intuitively, attention-based methods select informative instances and de-emphasize noisy instances by assigning different weights to different candidate instances. The weights are determined by the similarities between the candidates and attention vector. Specifically, we adopt the SBR of the corresponding entity as the attention vector. High attention indicates that the image representation is similar to its corresponding SBR, and thus should contribute more to the aggregated IBR of the entity according to the energy function. The aggregated

---

**Fig. 3.** Overall architecture of the IKRL model. Each entity has two kinds of representations: the SBRs and the IBRs. $h_1$ and $t_1$ are SBRs of head and tail entities, while $h_2$ and $t_2$ are IBRs. The entity representation and relation representation are associated by the translation mechanism. The attention-based method with attention values $\text{att}_i$ automatically selects informative instances from multiple representation candidates.

**Fig. 4.** Image encoder consisting of a pretrained representation module and an adaptive projection module.

The image projection module finally projects those image features from the image to the entity space (compare Fig. 4 for the overall pipeline of the encoder).

1) **Image Representation Module:** The image representation module constructs image feature representations for each image. For this, we obtain the feature representations from a pretrained AlexNet, a widely used deep CNN that consists of five convolution layers, two fully connected layers and a soft-max layer [14]. We reshape the images to $224 \times 224$ from the center, corners and their horizontal reflections. Lastly and in accordance with [31], we retrieve the 4096-D embeddings, which are the outputs of the second fully connected layer (called “fc7”), as the representation of the image feature.

2) **Image Projection Module:** After obtaining the compressed feature representations for each image, we associate images with the corresponding entities via a trainable image projection module. Specifically, we transfer the image feature representations from image space to entity space with a shared projection module. The IBR $p_i$ in the entity space for the $i$th image is defined as

$$p_i = M \cdot f(\text{img}_i)$$ (4)

where $M \in \mathbb{R}^{d_i \times d_e}$ is a trainable projection matrix, $d_i$ represents the dimension of image features in image space, $d_e$ represents the dimension of entities in entity space, and $f(\text{img}_i)$ stands for the $i$th image feature representation in image space, which is constructed by the image representation module.

**C. Attention-Based Multi-Instance Learning**

The image encoder takes images as inputs and then constructs IBRs for every single image. However, the most entities
IBR for the $k$th entity is defined as follows:

$$e_j^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \text{att}(p_i \cdot e_x^{(k)}) \cdot p_i^{(k)}.$$  

(6)

Besides the attention-based method, we also implement two alternative combination methods for further comparisons. AVG is a simple combination method that averages over all image representations, supposing that each image has equal contributions to the aggregated IBR. MAX is a simplified version for attention, which only considers the image representations with the highest attention.

D. Objective Formalization

We utilize a margin-based score function as our training objective, which is defined as follows:

$$L = \sum_{(h, r, t) \in T} \sum_{(h', r', t') \in T'} \max(\gamma + E(h, r, t) - E(h', r', t'), 0)$$

(7)

where $\gamma$ is a margin hyperparameter. $E(h, r, t)$ is the overall energy function stated above, where both head and tail entities have two kinds of representations, including SBRs and IBRs. $T'$ stands for the negative sample set of $T$ that we define as follows:

$$T' = \{(h', r, t) | h' \in E \} \cup \{(h, r', t') | t' \in E \}$$

$$\cup \{(h, r', t) | r' \in R \}, \quad (h, r, t) \in T$$

(8)

which means that a negative sample is obtained by randomly replacing one of the entities or relations in a triple. We also wipe out all generated negative triples that are already in $T$ to assure triples in $T'$ are truly negative. The training in translation-based methods is based on a pair-wise energy function, where positive triples shall lead to minimal energies, negative triples to large energies. This avoids a degenerate solution in which all vectors $\mathbf{h}$ and $\mathbf{t}$ would become the same and where $\mathbf{r}$ would become zero.

E. Optimization and Implementation Details

We formalize the IKRL model as a parameter set $\theta = (\mathbf{E}, \mathbf{R}, \mathbf{W}, \mathbf{M})$. In this set, $\mathbf{E}$ stands for the structure-based embedding set of entities, which consists directly of the embedding vectors (used as $\mathbf{h}_s$ or $\mathbf{ts}$, respectively). $\mathbf{R}$ stands for the embedding set of the relations. $\mathbf{W}$ and $\mathbf{M}$ represent the parameters of the image encoder: $\mathbf{W}$ are the parameters of the image representation module, which are pretrained and fixed during training and $\mathbf{M}$ is the projection matrix used in the image projection module.

We utilize mini-batch stochastic gradient descent (SGD) to optimize our model, with chain rule applied to update the parameters. $\mathbf{M}$ is initialized randomly. $\mathbf{E}$ and $\mathbf{R}$ are initialized from pretrained embeddings by TransE, while they could also be initialized randomly. In the image representation module, we utilize the AlexNet implemented by a deep learning framework Caffe [49] to construct image representations. In our experiments, AlexNet was pretrained on ILSVRC 2012 with a minor variation from the version described in [14]. For efficiency reasons, we use a GPU to accelerate the image representation and employ a multithread version for training.

| Relation                  | Train | Valid | Test | Total  |
|---------------------------|-------|-------|------|--------|
| hypernym                  | 5,162 | 583   | 538  | 6,283  |
| hyponym                   | 5,120 | 560   | 603  | 6,283  |
| part                      | 613   | 74    | 76   | 763    |
| has part                  | 598   | 91    | 72   | 761    |
| member of domain topic    | 72    | 11    | 5    | 88     |
| synset domain topic       | 66    | 11    | 11   | 88     |
| derivationally related form | 68  | 5     | 10   | 83     |
| member meronym            | 21    | 1     | 2    | 24     |
| member holonym            | 21    | 1     | 2    | 24     |
| Total                     | 11,741| 1,337 | 1,319| 14,397 |

### IV. Evaluation and Analysis

In order to investigate the effectiveness, we evaluate the performance of our proposed model on the task of KG completion and triple classification and analyze the resulting representations in-depth.

A. Data Set

For evaluation and analysis tasks in this paper, we constructed a new data set called WN9-IMG, combining the KG with images. First, we included triples from a subset of the KG data set WN18 [50], which was originally developed based on WordNet [51]. Second, we included 63,225 images, extracted from ImageNet [52], which is a large image database organized in accordance with the WordNet hierarchy, in order to provide a reasonable image quality. Here, we made sure that all entities have images and resulted in 6,555 entities and nine types of relations between them. The relations and the numbers of their occurrences are listed in Table I and the semantic categories according to WordNet are given in Table II.

| Entity     | Total |
|------------|-------|
| Plant      | 1280  |
| Geo. formation | 43   |
| Natural object | 149  |
| Sport      | 53    |

| Entity     | Total |
|------------|-------|
| Artifact   | 2134  |
| Person     | 797   |
| Animal     | 1489  |
| Misc       | 610   |

### B. Experimental Settings

The implementation of the IKRL model was trained using the mini-batch SGD, setting the margin $\gamma$ among $\{1.0, 2.0, 4.0\}$. For the learning rate $\lambda$, good values have been empirically identified among $\{0.0002, 0.0005, 0.001\}$, but a flexible, adaptively decreasing learning rate is feasible as well. In our experiments, we found setting $\gamma = 4.0$ and using a linear decline for $\lambda$ from 0.001 to 0.0002 as the optimal configuration. The dimensionality of the image feature embeddings was set to $d_i = 4096$ in order to ease comparison with the structure-based embeddings, while the dimensionality of the relation and entity embeddings was set to $d_r = 50$. In order to balance diversity and efficiency, we used an image number $n$ of up to 10 for all entities. For our baseline, we implemented TransE [11] and TransR [16] and used the experimental settings that have been reported in the respected publication but kept the dimensionality of the relation and entity embeddings set to 50 as well.
TABLE III
EVALUATION RESULTS ON ENTITY PREDICTION

| Metric  | Mean Rank | Hits@10 (%) |
|---------|-----------|-------------|
|         | Raw | Filter | Raw | Filter |
| TransE  | 143  | 137  | 79.9 | 91.2 |
| TransR  | 147  | 140  | 80.1 | 91.7 |
| IKRL (SBR) | 41  | 34  | 81.1 | 92.9 |
| IKRL (IBR) | 29  | 22  | 80.2 | 93.3 |
| IKRL (UNION) | 28  | 21  | 80.9 | 93.8 |

C. Knowledge Graph Completion

We conduct an experiment on KG completion, a typical task for KGs to evaluate the quality of knowledge representation. We also demonstrate the effectiveness of attention-based methods by comparing them to several other combination strategies.

1) Evaluation Protocol: KG completion aims to complete a triple \((h, r, t)\) when one of \(h, r, t\) is missing. Here, we focus on entity prediction, as this is commonly used to evaluate the quality of knowledge representations [11], [53]. The prediction is determined via the dissimilarity function \(|| h + r - t ||\). Since the IKRL model has two kinds of representations, we will report three prediction results based on our models: 1) SBR only utilizes structure-based representations to predict the missing component of a triple; 2) IBR only utilizes image-based representations in KG completion; while 3) UNION combines both entity representations by weighted concatenation.

Following the same settings in TransE [11], we use two measures as our evaluation metrics in entity prediction: 1) mean rank of correct entities (Mean Rank), which measures the overall rank of ground-truth entities and 2) proportion of correct entity results in top 10-ranked entities (Hits@10). We note that Mean Rank and Hits@10 are strict evaluation metrics, considering the large number of candidates. For example, there are 6555 possible candidate entities in the task of entity prediction and random chance is 3277.5 for Mean Rank and 0.0015 for Hits@10. Thus, low Mean Rank and high Hits@10 values strongly indicate good performance of models under evaluation. We also follow the two evaluation settings named “Raw” and “Filter” used in [11]. In this section, we first demonstrate the results of entity prediction, and then implement another experiment for further discussions on the power of attention.

2) Entity Prediction: Table III demonstrates the results of entity prediction. From the table, we observe that the following.

1) On all variants, the IKRL models outperform all baselines on both evaluation metrics of Mean Rank and Hits@10, among which UNION achieves the best performance. This indicates the successful integration of visual information and structured information, which is significant when building knowledge representations.

2) For SBR and IBR the performance indicates that including visual information enables building images-based representations but also benefits the SBRs.

3) All IKRL models outperform the baselines significantly on Mean Rank, seemingly because Mean Rank is depending on the quality of the knowledge representations and therefore sensitive to results that are wrongly predicted. TransE and other conventional translation-based methods are based on structured information only and may fail on the KG completion task in case of particularly sparse corresponding information. Though, since the IKRL includes visual information into the representation, the results of this model are much better on Mean Rank.

3) Varying Attention Strategies: In order to study the capability of the attention-based method, we compare three combination strategies that differ in how the multiple image instances are considered.

1) As the basic model, the IKRL (AVG) strategy chooses the average embedding of all available images instances for the entity representation.

2) The IKRL (MAX) strategy considers only the image instance that has the highest attention value to determine the entity representation.

3) The IKRL (ATT) strategy includes images into the entity representation based on the similarity to the SBR [compare (6)].

All results, comparing these strategies on both, structure-based as well as IBRs, are shown in Table IV.

From these results we can observe the following.

1) The baseline methods are outperformed by IKRL models using any of the combination strategies on Mean Rank and Hits@10. This emphasizes that introducing visual information into the encoding of the knowledge representation alone already improves the outcomes.

2) Overall, the performance is best for the ATT strategy. This indicates that the ATT strategy is successfully automatically selecting these images instances that are most representative of the corresponding entities.

3) The MAX strategies perform considerably worse than the AVG strategies, showing that including only images with high attention leads to losing information in other instances that might be important as well.

4) The ATT strategy shows only slight advantages over the AVG strategy. This seems to be caused by the data set construction, where we especially focused on including high-quality images and thus may have narrowed down the need for the selective nature of the ATT strategy.

Although these results provide evidence for the strength of the IKRL model using the ATT strategy, a qualitative analysis could reveal whether this is caused by a successful differentiation of poor and good image candidates and will be provided in a case study.

D. Triple Classification

Another typical task for KGs is triple classification. Experiment results on triple classification task demonstrate the effectiveness of the proposed KRL method.

1) Evaluation Protocol: In triple classification, a method is evaluated based on a dissimilarity function over all triples. In the basic form of binary classification it is determined, whether a triple \((h, r, t)\) is correct or incorrect [54]. To enable such an evaluation we added negative instances to our data set by replacing head or tail entities of correct instances by random other entities, as proposed in [54]. In particular, a triple \((h, r, t)\) is evaluated as positive in case the dissimilarity function \(|| h + r - t ||\) results below a threshold \(\delta\), which was optimized beforehand by maximizing the cumulated classification accuracy on the validation set. For the IKRL model, we focus on calculating the dissimilarity function on the IBR in order to provide a comparison with the baseline methods.
TABLE IV
Evaluation Results on Different Combination Strategies

| Type       | Structure-based representation | Image-based representation |
|------------|--------------------------------|---------------------------|
|            | Mean Rank | Raw Filter | Hits@10 (%) | Mean Rank | Raw Filter | Hits@10 (%) |
| IKRL (MAX) | 62        | 55         | 81.0        | 92.3       | 59        | 52         | 79.8        | 92.1       |
| IKRL (AVG) | 43        | 36         | 80.7        | 92.8       | 29        | 22         | 79.3        | 92.9       |
| IKRL (ATT) | 41        | 34         | 81.1        | 92.9       | 29        | 22         | 80.2        | 93.3       |

Fig. 5. Analysis of the entity representations: similarity matrices for represented entities of different categories.

TABLE V
Evaluation Results on Triple Classification

| Methods     | Accuracy (%) |
|-------------|--------------|
| TransE      | 95.0         |
| TransR      | 95.3         |
| IKRL (MAX)  | 96.3         |
| IKRL (AVG)  | 96.6         |
| IKRL (ATT)  | 96.9         |

2) Experimental Results: From Table V, we can obtain the following.

1) Compared to the baseline methods, all IKRL model variants reach higher accuracies, indicating higher robustness and effectiveness when integrating structure-based and image-based information. Since the baseline model TransE was used for initializing the SBR, the improvements are seemingly introduced by the images.

2) The IKRL model using attention in aggregating the representation (ATT) results in the best performance. Compared to other strategies, this shows that taking multiple instances into account but choosing the most informative image from all candidates in a smart fashion leads to the relatively best representation formation.

E. Representation Analysis

In order to understand how the model representations were formed while integrating structured knowledge information and visual information, we analyzed the resulting representations for the entities and relations.

At first, we inspected the entity representations that are based on the structure and the image-projection, respectively. Computing the covariance between all individual entities of the whole data set, we can visualize how both representations contribute to different aspects of the entities to the model with respect to their meaning. For this, we performed a principal component analysis (PCA) and a representational similarity analysis (RSA) [55] on both representations. Fig. 5 shows the similarity matrices and Fig. 6 provides the plots of the projections of all the data points’ representation onto the first two principal components (PC1 and PC2). The similarity matrices, as well as the plots, differentiate into eight major categories as suggested for ImageNet and WordNet, respectively, (compare Table II).

From the plots, we can infer that the IBR stronger discriminates the entities based on their category, while the SBR interlinks entities from certain categories. On the first principal component (PC), this is particularly visible for the artifact category, which spreads out wide in structure-based PCA space, indicating that the SBR captures particularly general properties for these entities since they occur in a broad range of situations in KG data sets. The images used to obtain the IBR, however, specify use cases and particular settings and thus include a more narrow connotation. Remarkably, entities from the plants or natural object categories seem to provide less variability. Inspecting specific entity representations confirms that both image- and SBR spaces show regularities for cases of dependence, particularly visible for semantic inclusion (hypernym/hyponym, part-of/has-part, etc.) and also similarity (synonyms). The IBR implies a notably stronger correlation based on appearance properties (e.g., compare for artifact or plant), while the SBR suggests correlations because of functional links (e.g., sport). Fig. 7 provides examples for those entities. Note that for both structure-based as well as IBRs, the representational complexity is high (compare Fig. 6 on the right), which shows that the data can only be explained well if considering a large number of components. In the case of the IBRs, the majority of the data is explained with slightly fewer dimensions. In all this shows that the representations develop similarly but include subtle but important differences.

Second, we compared the relation representations that formed in the model, particularly how they are linked, based on the representation space (see Fig. 8). The PCA reveals that...
the relations are represented based on their occurrence and role in the data set. For instance, the most frequently used relations, the opposite hypernym and hyponym are represented orthogonally on the axis of the first PC, while part of and has part are represented orthogonally on the axis of the second PC. Overall, this shows that the complexity of the relation representation space is larger than necessary and was adapted to cope with our particular data set in a way that the most frequent ones are differentiated most strongly.

### F. Case Study

To further understand how the model is exploiting the representations, we provide detailed analyses for two cases. First, we present the capability of the attention. Second, we demonstrate the semantic regularities that underlay the representations. Note, the shown images are slightly chopped to ease the reader’s focus on the included main objects, although the full images were used in the tests.

#### 1) Attention

Different pairs of image instances are shown in Fig. 9 in order to showcase how the attention component is
v(part_of) \approx
\begin{array}{c}
\text{dresser} \\
\text{drawer}
\end{array}
\Rightarrow
\begin{array}{c}
\text{pianoforte} \\
\text{keyboard}
\end{array}
\Rightarrow
\begin{array}{c}
\text{cat (Felidae)} \\
\text{tiger}
\end{array}
\Rightarrow
\begin{array}{c}
\text{toothed whale} \\
\text{dolphin}
\end{array}

2) Semantic Regularities of Images: Our analyses as well as Mikolov et al. [56] showed that representations for the entities and thus some word embeddings have interesting regularities, for instance: \( v(\text{king}) - v(\text{man}) \approx v(\text{queen}) - v(\text{woman}) \). In image–text space comparable regularities have been reported [30], showing that is feasible to interpret the images-structure knowledge space in-depth. In the joint space of images-structure knowledge, we identified similar semantic translation regularities for the image versus the structure-based case with respect to the relations (compare Fig. 10). For instances, the result of \( \text{dresser minus drawer matches the specific relation part_of} \), and \( \text{cat minus tiger yields the hypernym relation} \). These concrete and meaningful matches confirm that the representations encode the semantic regularities well.

V. CONCLUSION

In this paper, we study the proposed IKLR model\(^2\) with the aim of integrating image-embedding and knowledge-based representation learning. Accounting for the distributed representations in the human brain that develop during language learning, we utilize neural networks for encoding visual information from images and structure-based encodings from established knowledge-based mechanisms. We employ a projection module to model entities, found in each image, and then construct the aggregated IBRs by combining multiple image instances based on attention. Our experimental results confirm that our model is capable of encoding image information into knowledge representations and allows for a better prediction of entities and exploration of relational facts in KG completion and classification tasks. From the detailed analysis, we learned that the representations for entities tend to merge information from image-based and structure-based encodings but contribute aspects that are specific to visual observations and the knowledge about structural links, respectively. In particular, semantic regularities, which underlie the observations from the data but are latent in the representations, are exploited by the attention.

By constraining the representations of concepts via relations, our model helps to develop meaningful amodal representations [4], [5], since the relations that are expressed in the KG are independent of modality. Simultaneously, individual concepts remain grounded in reality, because their representation can be obtained from images. With regard to concept learning in the brain and the current theories on entity representation (compare Fig. 1), the model can help to understand the dynamics of the representation formation. Since head-, relation-, and tail-representations are vectors of the same dimensionality, which are transformed via summation into each other, and since a given concept can appear as either head or tail, we would regard head-, relation- and tail representations to occupy the same neural tissue, which may correspond to a group of several language-related cortical areas. It would be possible to constrain neural activations in the model to be sparse and positive in order to yield more biologically plausible patterns and to facilitate the superposition of several patterns while reducing interference. Different patterns might also activate sequentially, such as activating the head at first, the relation at second, and the tail representations at third. Such activation sequences could be encoded by recurrent connections, which could be added to the model.

In future work, we will further explore this paper in different directions.

1) We consider more advanced and complex models for better extracting the features that are relevant for visual representations, and enhance the translation-based methods by extending the image-based model.

2) Since this paper is limited to regarding entity images as a visual representation of the corresponding entity, we plan to explore learning multiple entities and their relations within a single image in combination with the IKRL model.

3) We plan to integrate further mechanisms that have been suggested to underly language learning in humans, such as embodied multimodal representations as well as training via scaffolding.

\(^2\)Both the source code and the data set of this paper can be accessed via https://github.com/thunlp/IKRL.
Interesting applications are search engines and QA tasks. While search engines can operate on query items without analyzing their relations, they will benefit when considering the relations between items. For QA, relations are essential. Moreover, the learned word embeddings of our model are constrained by relations from the KG, which endows them with semantic content and which makes them robust.

Overall, concept representations can differ if constrained differently, e.g., via knowledge bases, images, language statistics [12], or combinations thereof [57]. Thus, for concept representations in the cortex that are distributed over several modalities (see Fig. 1), this means that they may be strongly distributed but entangled because they account for different needs of visual, auditory, somato-sensory-motor and frontal areas’ functions. We are confident that combining embodied processes from human development with computational knowledge-based systems can provide both, a better insight into mechanisms and representations in humans and more robust and adaptive models of language and meaning.

REFERENCES
[1] A. Cangelosi and M. Schlesinger, Developmental Robotics: From Babies to Robots. Cambridge, MA, USA: MIT Press, 2015.
[2] F. Pulvermüller and L. Fadiga, “Active perception: Sensorimotor circuits as a cortical basis for language,” Nat. Rev. Neurosci., vol. 11, no. 5, pp. 360–370, May 2010.
[3] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, “Natural speech reveals the semantic maps that tile human cerebral cortex,” Nature, vol. 532, no. 7600, pp. 453–458, Apr. 2016.
[4] J. R. Binder, “In defense of abstract conceptual representations,” Psychonomic Bull. Rev., vol. 23, no. 4, pp. 1096–1108, 2016.
[5] M. Kiefer and F. Pulvermüller, “Conceptual representations in mind and brain: Theoretical developments and evidence for future directions,” Cortex, vol. 48, no. 7, pp. 805–825, 2012.
[6] F. Pulvermüller, “How neurons make meaning: Brain mechanisms for embodied and abstract-symbolic semantics,” Trends Cogn. Sci., vol. 17, no. 9, pp. 458–470, 2013.
[7] V. Jahn, V. Zhigulin, and H. S. Seung, “Representing part-whole relationships in recurrent neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2006, pp. 563–570.
[8] N. B. Turk-Browne, B. J. Scholl, M. M. Chun, and M. K. Johnson, “Neural evidence of statistical learning: Efficient detection of visual regularities without awareness,” J. Cogn. Neurosci., vol. 21, no. 10, pp. 1934–1945, 2009.
[9] B. Yang, W.-T. Yih, X. He, J. Gao, and L. Deng, “Embedding entities and relations for learning and inference in knowledge bases,” in Proc. ICLR, 2015, pp. 1–13.
[10] J. Yin et al., “Neural generative question answering,” in Proc. IJCAI, vol. 70, pp. 2181–2187.
[11] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data,” in Proc. NIPS, 2013, pp. 2787–2795.
[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in Proc. ICLR, 2013, pp. 1–12.
[13] R. Xie, Z. Liu, H. Luan, and M. Sun, “Image-embodied knowledge representation learning,” in Proc. IJCAI, 2017, pp. 3140–3146.
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NIPS, 2012, pp. 1097–1105.
[15] Z. Wang, J. Zhang, J. Feng, and Z. Chen, “Knowledge graph embedding via hypercomplex numbers,” in Proc. AAAI, 2014, pp. 1112–1119.
[16] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, “Learning entity and relation embeddings for knowledge graph completion,” in Proc. AAAI, 2015, pp. 2181–2187.
[17] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, “Knowledge graph embedding via dynamic mapping matrix,” in Proc. ACL, 2015, pp. 687–696.
[18] M. Fan, Q. Zhou, E. Chang, and T. F. Zheng, “Transition-based knowledge graph embedding with relational mapping properties,” in Proc. PACLIC, 2014, pp. 328–337.
[19] J. Feng et al., “Knowledge graph embedding by flexible translation,” in Proc. KR, 2016, pp. 557–560.
[20] H. Xiao, M. Huang, Y. Hao, and X. Zhu, “TransA: An adaptive approach for knowledge graph embedding,” CoRR, vol. abs/1509.05490, Apr. 2015.
[21] L. Y. Wu et al., “StarSpace: Embed all the things!” in Proc. AAAI, 2018, pp. 5586–5593.
[22] J. L. Tucker, “Some mathematical notes on three-mode factor analysis,” Psychometrika, vol. 31, no. 3, pp. 279–311, 1966.