IJCNN 2019. International Joint Conference on Neural Networks. Budapest, Hungary. 14-19 July 2019

Multimodal Data Enhanced Representation Learning for Knowledge Graphs

Zikang Wang*†, Linjing Li*, Qiudan Li*, Daniel Zeng*†
*The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China
†School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China
{wangzikang2016, linjing.li, qiudan.li, dajun.zeng}@ia.ac.cn

***Abstract*—Knowledge graph, or knowledge base, plays an important role in a variety of applications in the field of artificial intelligence. In both research and application of knowledge graph, knowledge representation learning is one of the fundamental tasks. Existing representation learning approaches are mainly based on structural knowledge between entities and relations, while knowledge among entities per se is largely ignored. Though a few approaches integrated entity knowledge while learning representations, these methods lack the flexibility to apply to multimodalities. To tackle this problem, in this paper, we propose a new representation learning method, TransAE, by combining multimodal autoencoder with TransE model, where TransE is a simple and effective representation learning method for knowledge graphs. In TransAE, the hidden layer of autoencoder is used as the representation of entities in the TransE model, thus it encodes not only the structural knowledge, but also the multimodal knowledge, such as visual and textural knowledge, into the final representation. Compared with traditional methods based on only structural knowledge, TransAE can significantly improve the performance in the sense of link prediction and triplet classification. Also, TransAE has the ability to learn representations for entities out of knowledge base in zero-shot. Experiments on various tasks demonstrate the effectiveness of our proposed TransAE method.**

***Index Terms*—representation learning, knowledge graph, multimodal**

I. Introduction

Knowledge plays an important role in artificial intelligence, as it provides a basis for machines to comprehend and accomplish complex tasks. Knowledge graph, or knowledge base, is used to store knowledge in a structured way. Commonly used knowledge graphs include WordNet [1], Freebase [2], NELL [3], etc. These knowledge graphs provide solid data support for various applications like question answering system, information retrieval, and machine comprehension.

Knowledge is typically represented as triplets in knowledge graph. A triplet, often denoted as (h, r, t), indicates the head entity h and tail entity t have relation r. Representation learning for knowledge is one of the most essential tasks in knowledge graph. Various approaches have been explored to learn knowledge representations. Translation-based approach [4]–[7] is one of the most popular methods, it maps entities and relations into a low-dimensional continuous vector space, where embeddings for entity h, t, and relation r satisfy equation $h + r \approx t$. Translation-based methods have received huge success in the field of knowledge representation.

(a) Explicit knowledge

(b) Implicit knowledge

Fig. 1
(a) illustrates that images can encode explicit knowledge in knowledge graph, like (chair, has part, leg). (b) shows images can also contain implicit knowledge, like "butterfly are highly related with flowers".

However, most of the current approaches only consider the relational knowledge between entities, which we also refer to "structure" knowledge [8], [9] in this paper. In this case, embeddings are only trained to satisfy relations with each other. Information of entities themselves has merely been taken into account, while these information, like visual or textual descriptions, are easy to access and contain lots of valuable knowledge, including both the knowledge contained explicitly in knowledge graph and even the knowledge which are hard to be described as triplets. Fig. 1(a) is an example of images representing knowledge explicitly stored in knowledge graph as (h, r, t), which is (chair, has part, leg) in this case; in Fig. 1(b), butterfly and flower all occur at the same time, which indicates entity "butterfly" and "flower" are highly related [1]. Information like this is obvious for human while hard to be captured and represented in knowledge graph. Explicitly encode knowledge of different modalities into entity

[1]All images in Fig. 1 are from ImageNet [10].

978-1-7281-2009-6/$31.00 ©2019 IEEE paper N-19826.pdf

representations can overcome this drawback to some extent, and can also bring significant improvements to the quality of learned representations according to our experiments.

Entities can have knowledge of multiple modalities stored in the knowledge graph, such as pictures, text descriptions, audios, videos, etc. However, existing attempts to integrate extra knowledge all focus on single modality [8], [9]. In this paper, we try to learn structural knowledge and knowledge of different modalities jointly, which guarantees the flexibility to integrate various modalities.

To learn structure knowledge and multimodal knowledge jointly, we propose a new knowledge representation learning model, TransAE, by combing multimodal autoencoder(AE) and TransE. Based on this model, we can learn joint representations for entities given triplets and their corresponding descriptions of different modalities. In this paper, we consider two modalities: visual and textual. During experiments, we harvest the knowledge base WN9-IMG [9] with textual information, to construct a knowledge graph contains not only triplets, but also visual and textual knowledge for each entity. We use feature vectors of text and images to represent multimodal knowledge and feed them into our model to learn the desired representation.

We evaluate our model on traditional knowledge graph representation tasks, link prediction and triplet classification. On both two tasks, our model outperforms all methods concentrated only on structural knowledge with significant improvements, which demonstrates the usefulness of multimodal knowledge and the effectiveness of our method on learning relational knowledge between entities. Furthermore, we also perform these tasks on out-of-knowledge-base(OOKB) entities, our approach can learn valid representations for them in zero-shot, which cannot be done by most existing representation learning methods. Also, by assigning different weights to autoencoder loss in the loss function, we show that extra knowledge like text and images can improve the representation performance in both efficiency and accuracy. Last but not least, we perform multimodal query retrieval as a case study, which shows that our model can group similar entities together in the embedding space.

Our contributions are three-fold:

- We propose TransAE, a knowledge representation learning model, which can learn representations based on both structural knowledge and multimodal knowledge of triplets. The introduction of multimodal knowledge leads to great improvements on model performance compared to traditional approaches.
- Our model learns a joint representation based on all kinds of knowledge, thus can be easily adapted to various kinds and amounts of modalities, which is difficult for pervious models.
- Apart from receiving promising results on knowledge graph representation tasks, our method can also learn valid representations for entities out of knowledge base in zero-shot, which cannot be done by traditional methods.

II. Related Work

Representation learning is well studied in recent years, several methods have been proposed to learn representations based on knowledge graphs.

Translation-based approach is one of the most popular methods for representation learning. First proposed in [4], TransE model achieves good results with simple, intuitive assumptions, leading to a series of work such as TransH [5], TransR [6], TransD [7], etc. TransE maps entities and relations into a low-dimensional continuous space, under the assumption that head entity h, tail entity t, and the relation r satisfy $h + r \approx t$. Since TransE cannot handle complicated relationships very well, like 1-to-N, N-to-1, and N-to-N relations, TransH is proposed in [5], by mapping entities into a relation-specific hyperplane, where the same entity represented differently under different relations. For TransE and TransH, entities and relations are all projected into the same embedding space, TransR [6], however, creates a separate relation space and maps triplets into it to satisfy $h_r + r \approx t_r$. Other translation-based methods include TransD [7], TransG [11], etc.

Apart from translation-based method, many other effective approaches have also been explored. RESCAL [12] is an approach getting low-dimensional representations by matrix factorization, HolE [13] uses correlation as compositional operator to learn compositional vector space representations. All these methods can get representative embeddings for entities and relations.

However, all these methods mainly learn representations based on the structural information between entities and relations, paying little attention to entities themselves. Rich knowledge can be obtained from images and textual descriptions of entities. There are some works attempting to improve representations by introducing extra knowledge to knowledge graph from various sources. IKRL [9] is the first work integrating visual knowledge to existing models to help learn representations. IKRL learns two separate representations for each entity, one based on structural knowledge and the other based on visual knowledge. Both representations yield much better results in experiments than previous models. Integrating textual knowledge into knowledge graph has also been explored. Proposed in [8], DKRL introduces entity descriptions from Freebase [2] to knowledge graph, it extracts feature vectors of text using CBOW [14] model, and learns knowledge from text and structure simultaneously. Similar as IKRL, DKRL also learn representations based on structural knowledge and textual knowledge separately, then integrate the two representations into one. These two models are both limited by the number of modalities, it is hard for them to integrate knowledge of more than one modalities, as their model complexity is largely depended on the number of modalities.

III. Methodology

To learn from multimodal knowledge and structural knowledge jointly, we first get knowledge from each single modality

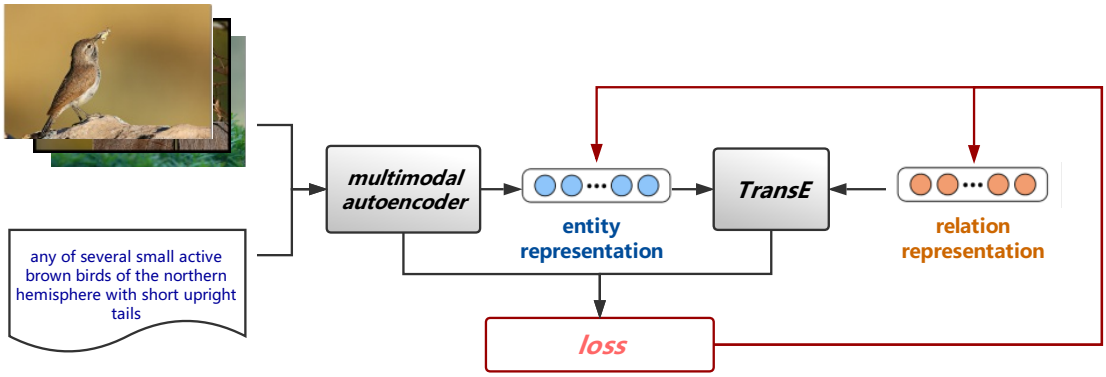

Fig. 2: The overall architecture of TransAE. We feed images and text into multimodal autoencoder to learn entity embeddings. We want representations of entities and relations satisfy $h + r \approx t$, as proposed in TransE. Representations are trained simultaneously based on both autoencoder loss and TransE loss.

by learning feature vectors respectively. Then we feed these feature vectors into multimodal autoencoder. For entities, we let the middle-hidden layer of an autoencoder satisfy the TransE assumption $h + r \approx t$ with corresponding relations. In this paper, we consider two modalities: visual and textual. Note that our model can easily be adapted to any other modalities as well. The overall architecture of the TransAE model is shown in Fig. 2.

A. Knowledge Extraction

We use feature vectors of images and text as visual and textual knowledge respectively. For visual knowledge, we use VGG16 Net [15] pre-trained on ImageNet [10]. VGG16 Net consists of several convolution layers each followed by a ReLU layer and a max pooling layer, there are two fully connected layers before the final softmax layer classifying images into 1000 categories. We use the vector from the last fully connected layer as our desired image feature vector.

To extract textual features, we use PV-DM [16] model, one of the Doc2Vec [16] models, to learn paragraph vectors in an unsupervised way. PV-DM can learn a fixed-length representation from variable-length sentences, which is the definition of entities in our case. PV-DM model is like Word2Vec [14] model, it maps the paragraph to a vector, then calculate a joint vector by averaging or concatenating the paragraph vector with word vectors, and predict the next word according to the joint vector. Different from Word2Vec model, Doc2Vec model takes the word order into consideration, encoding both semantic knowledge and context information into embeddings at the same time.

B. Knowledge Representation Learning

Multimodal autoencoder

The multimodal autoencoder we used in our paper is a feedforward neural network consisting of an input layer, three hidden layers and a output layer. The input layer takes input from two different modalities, which are image feature vectors $v_i^{(1)}$ and text feature vectors $v_t^{(1)}$ we get in the previous initialization step. A fully-connected hidden layer with much less units is followed the input layer for each modality, we map the input to these hidden layers separately, and denote them as $v_i^{(2)}$ and $v_t^{(2)}$. Then we concatenate $v_i^{(2)}$ and $v_t^{(2)}$, map it to the second hidden layer jointly, which is also the middle layer of autoencoder network, and the desired joint multimodal entity embedding $v^{(3)}$.

The decoder stage has exactly symmetrical structure with encoder, taking the embedding $v^{(3)}$ as input, first maps it to two separate hidden layers $v_i^{(4)}$ and $v_t^{(4)}$, each shares the same dimension with the corresponding hidden layer in the encoder part, $v_i^{(2)}$ and $v_t^{(2)}$. Then the decoder maps the two layers to output layer $v_i^{(5)}$, $v_t^{(5)}$. The output layer and input layer are of the same dimension for each modality. The output layer is also called reconstruction layer, which aims to reconstruct the input feature vectors.

The whole architecture of the proposed multimodal autoencoder is defined as:

$$v_i^{(2)} = f(W_i^{(1)} \times v_i^{(1)} + b_i^{(1)}) \quad (1)$$
$$v_t^{(2)} = f(W_t^{(1)} \times v_t^{(1)} + b_t^{(1)}) \quad (2)$$
$$v^{(3)} = f(W^{(2)} \times (v_i^{(2)} \oplus v_t^{(2)}) + b^{(2)}) \quad (3)$$
$$v_i^{(4)} = f(W_i^{(3)} \times v^{(3)} + b_i^{(3)}) \quad (4)$$
$$v_t^{(4)} = f(W_t^{(3)} \times v^{(3)} + b_j^{(3)}) \quad (5)$$
$$v_i^{(5)} = f(W_i^{(4)} \times v_i^{(4)} + b_i^{(4)}) \quad (6)$$
$$v_t^{(5)} = f(W_t^{(4)} \times v_t^{(4)} + b_j^{(4)}) \quad (7)$$

where f is an activation function, which is set to sigmoid in this paper. $W_i^{(j)}$ and $W_t^{(j)}$ denote the weight matrix mapping

embeddings from layer j to layer $(j+1)$ of modality image and text respectively, $b_i^{(j)}$ and $b_t^{(j)}$ are the bias items for image and text in the j-th hidden layer. Symbol $\oplus$ represents the concatenate operation.

The whole multimodal autoencoder is trained to minimize the reconstruction error, which is the sum of dissimilarities between the input layer and the output layer for the two modalities:

$$L_a = \left\| v_i^{(1)} - v_i^{(5)} \right\|_2^2 + \left\| v_t^{(1)} - v_t^{(5)} \right\|_2^2 \tag{8}$$

TransE

The TransE model maps entities and relations in low-dimensional continuous vector space, where embeddings of head entity h, tail entity t, and relation r satisfy:

$$h + r \approx t \tag{9}$$

TransE tries to minimize the distance between $h+r$ and t for triplets in knowledge graph, and maximize distance for corrupted ones which are out of knowledge graph. While training TransE, we minimize the loss L_e:

$$L_e = \sum_{(h,r,t)\in S} \sum_{(h',r',t')\in S'} \max(0, [\gamma + d(h,r,t) - d(h',r',t')]) \tag{10}$$

where γ is a margin hyperparameter, and $\max(0,x)$ aims to get the positive part of x, d is the dissimilarity function, can be L1 norm or L2 norm. S denotes the original dataset consists of triplets existed in knowledge graph, while S' is the corrupted dataset consisted of triplets not in knowledge graph. S' is considered as negative samples while training, it is constructed by replacing head or tail entity for each triplet, following:

$$S'_{h,r,t} = (h', r, t)|h' \in E \cup (h, r, t')|t' \in E \tag{11}$$

TransAE

In the TransAE model, we combine the above two models to learn multimodal knowledge and structural knowledge simultaneously. There are several images and a sentence description for each entity in the knowledge graph, we first extract their visual and textual feature vectors, then feed these vectors into multimodal autoencoder to get the joint embedding as the entity representation. Relation embeddings are initialized randomly in the beginning of the training. These entity and relation representations are used to train our model. Structure loss $L_e^{'}$ for (h,r,t) can be represented as:

$$L'_e = \sum_{(h,r,t)\in S} \sum_{(h',r',t')\in S'} \max(0, [\gamma + d(v_h^{(3)}, r, v_t^{(3)}) - d(v_{h'}^{(3)}, r', v_{t'}^{(3)})]) \tag{12}$$

where $v_h^{(3)}$ and $v_t^{(3)}$ represent the representation for h and t embeddings respectively. For better generalization, we also add a regularizer $\Omega(\theta)$ for parameter set θ, with α as its weight. Thus we can train our model by minimizing the overall loss L:

$$L = L_a + \beta L'_e + \alpha\Omega(\theta) \tag{13}$$

Parameters $W_i^{(j)}$, $W_t^{(j)}$, $b_i^{(j)}$, $b_t^{(j)}$ in autoencoder and relation embeddings are initialized randomly at the beginning of training. Weight parameters β and α are chosen to balance the magnitude and importance of losses L_a, L'_e and regularizer $\Omega(\theta)$.

At each training round, a set of triplets are randomly selected, combing with the constructed corrupted triplets, together serve as batch data. We train our model on the training set and select optimal parameters based on the validation set. Performance of our model is illustrated in the next section in detail.

IV. Results

We empirically evaluate our model on two tasks: link prediction [4] and triplets classification [17]. These two tasks evaluate whether the learned representations can embody the structural knowledge. Also, we preform these two tasks on out-of-knowledge-base(OOKB) entities to show that TransAE can learn valid representations based on multimodal knowledge for OOKB entities in zero-shot. Furthermore, we analyze the role of multimodal knowledge during training, showing it brings improvements in both efficiency and accuracy. In the case study part, examples show that our model can group embeddings of similar entities together, indicating our model can capture the entity knowledge as well.

A. Datasets

We harvest the existing knowledge graph WN9-IMG [9] with textual information to create a new knowledge base WN9-IMG-TXT. WN9-IMG-TXT consists of three parts: entity-relation triplets, images of entities, and textual descriptions of entities, among which the triplets and images are the same with WN9-IMG dataset. To be more specific, triplets are a subset of WN18 [18], which is built on knowledge graph WordNet [1]. For visual knowledge, images are all extracted from ImageNet [10], a large image database built according to WordNet hierarchy, containing images for entities in WordNet. All entities in our dataset have images mapped to them with amount up to 10 as in WN9-IMG [9]. Textual description is also extracted from WordNet, where definition is provided for each entity. This new dataset contains 9 relation types and 6555 entities. Specific statistics is shown in Table I.

TABLE I: Statistics of dataset.

Dataset	#Rel	#Ent	#Train	#Valid	#Test
WN9-IMG-TXT	9	6,555	11,741	1,337	1,319

Besides the statistics shown in Table I, there are also 6555 sentences and 63225 images in it. We train and evaluate our

model on WN9-IMG-TXT, and simply ignore the visual and textual descriptions when evaluating all baseline methods.

B. Experimental Settings

Baselines

Since the triplet and image parts in our dataset are the same with WN9-IMG, which is also used in [9], we directly adopt the experimental results of model TransE [4], TransR [6] and IKRL [9] as our baseline results from paper [9].

For models not evaluated in [9], such as TransH [5], TransD [7], RESCAL [12] and HolE [13], we train them on triplets of our dataset using OpenKE[2], following the optimal experimental settings claimed in their original papers.

Implementation

During training, we select learning rate λ among $\{0.0001, 0.001, 0.01\}$, margin γ among $\{0.1, 0.5, 1.0, 2.0\}$, weight β among $0.1 \sim 0.9$. Select optimizer among SGD, RMSProp and Adam. We choose dissimilarity measure d among L1 and L2 measures. We choose the way replacing entities while creating corrupted triplets between "uniform" and "bernoulli". "Bernoulli" is a way to reduce false negative labels proposed in [5], we strictly follow the instructions in paper [5]. The dimensions of input layer for visual and textual modality are both fixed, $v_i^{(1)}$ and $v_t^{(1)}$ are set to be 4096 and 100 respectively. We select the first hidden layer dimension for visual modality $v_i^{(2)}$ among $\{512, 1024, 2048\}$ and $\{50, 100\}$ for textual modality $v_t^{(2)}$. We select hidden dimension d among $\{20, 50, 100\}$, which is also the dimension of relation representations. Note decoder and encoder have exactly the same structure and dimensions for corresponding layers in autoencoder.

The optimal configurations we used are: $\lambda = 0.001$, $\gamma = 0.5$, $\beta = 0.4$. We choose the optimizer to be RMSProp, use L1 as dissimilarity measure, and construct corrupted entities by the way "bernoulli". For dimensions, we set $v_i^{(2)} = 512$, $v_t^{(2)} = 50$, $d = 50$. We limit the training times in 1000 rounds.

C. Link Prediction

Evaluation protocol

Link prediction is a task first proposed in [19] to predict the missing entities in triplets, and has been widely used to evaluate the learned representations. We follow the same protocol as described in [19]. We first construct corrupted triplets by removing head entity in each triplet and replace it with every other entities in dataset. Then rank these triplets in ascending order according to their dissimilarity scores. We can get the rank for tail entities by removing tail entities instead of head following the same procedure.

Two metrics are used to evaluate: mean rank (the average of predicted ranks for correct entities) and hits@10(proportion of original golden entities ranked in top 10). These metrics are also referred to as "raw" mean rank and "raw" hits@10. Corrupted triplets constructed this way may turn out to be in the knowledge graph, as the triplets are 1-to-N, N-to-1 or N-to-N, not 1-to-1 in dataset. To eliminate this error, we remove all the candidates existed in the knowledge graph before evaluation, this leads to "filtered" mean rank and "filtered" hits@10. For both raw and filtered settings, a smaller mean rank and a higher hits@10 indicates a representation with more capability.

[2]https://github.com/thunlp/OpenKE

TABLE II: Link Prediction Results.

Metric	Mean Rank		Hits@10(%)	
	Raw	Filter	Raw	Filter
TransE	170	165	73.2	87.1
TransH	275	267	56.7	65.4
TransR	172	168	62.3	85.4
TransD	261	252	70.1	83.2
RESCAL	364	356	54.4	61.5
HolE	282	274	78.9	90.4
DistMult	326	310	74.5	80.4
ComplEx	319	291	82.9	87.1
DKRL(CNN)	148	89	79.0	92.1
IKRL(SBR)	41	34	81.1	92.9
IKRL(IBR)	29	22	80.2	93.3
IKRL(UNION)	**28**	21	80.9	93.8
TransAE	29	**17**	**83.2**	**94.2**

Experimental results

The experimental results are shown in Table II. Our model outperforms all baseline models that only use structural knowledge significantly, and receives competitive results with IKRL and DKRL.

From Table II we can conclude that visual and textual knowledge bring large improvement in the results, IKRL, DKRL and our model all achieve results far beyond others, which proves the effectiveness of external multimodal knowledge. IKRL [9] learns two separate representations based on structural knowledge and images, referred to as "SBR" and "IBR" respectively. Baseline mentioned as "UNION" is calculated by combining "SBR" and "IBR" representations. Our method can get competitive results with IKRL while learning one unified representations jointly. Also, our model is quite simple and flexible, can be combined with other more complicated models than TransE, which may lead to better performance.

D. Triplet Classification

Evaluation protocol

Triplet classification aims to classify whether a triplet (h, r, t) is correct or not based on the assumption that a triplet is correct if and only if it exists in the known knowledge graph. Following the same procedure in NTN [17], we construct a corrupted triplet which is not in knowledge graph for each triplet, then we calculate a relation-specific threshold σ_r for classification based on dissimilarity scores of all triplets. We say a triplet is positive if its dissimilarity score is below the threshold σ_r for the given relation r, and consider it negative otherwise.

TABLE III: Triple Classification Results. TransAE outperforms all the baselines.

Model	Accuracy(%)
TransE	95.0
TransR	95.3
IKRL(MAX)	96.3
IKRL(AVG)	96.6
IKRL(ATT)	96.9
TransAE	**97.9**

Experimental results

We listed the results in Table III. It shows that our method gets better result than all of the pervious approaches, also, the accuracy is high enough to discriminate positive triplets in knowledge graph from corrupted ones in most cases.

E. Zero-shot Learning for OOKB Entities

Most existing models can only perform link prediction task and triplet classification task on entities within the given knowledge graph, learning representations for out-of-knowledge-base(OOKB) entities has always been a challenging problem [22]. Making use of both textual and visual knowledge, our method can easily learn representations for all entities, whether they are in knowledge graph or not.

We perform both link prediction task and triplet classification task on OOKB entities. We split the WN9-IMG-TXT dataset into 4 datasets: new training set as the new knowledge base, which is a subset of the original WN9-IMG-TXT dataset; the first test set where only head entity is out of knowledge base, the second test set where only tail entity is out of knowledge base, and the third test set where head and tail entities are both out of knowledge base. There are 6000 triplets in the new training set, while 1000 triplets each test set. We perform both link prediction and triplet classification tasks on all these three datasets.

Zero-shot Link Prediction

In this task, we follow the same protocol as in the normal link prediction task, except that not all the entity representations are learned during training. We learn representations for OOKB entities based on the multimodal representation model, whose parameters are learned based on the known entities of knowledge base. We report the filtered mean rank among all 6555 entities in Table IV, from which we can see that our method can get good results for entities not known before, the results are even competitive to some traditional approaches on known entities according to Table II.

Zero-shot Triplet Classification

For this task, we use the same dataset as in zero-shot link prediction, following the same procedure as in the previous section. Experiment results are shown in Table V, where the classification accuracy is only slightly lower than the case where all entities are known.

These experiments show that TransAE can learn valid representations for unknown entities in zero-shot and get satisfactory results, which is unachievable for most existing models.

TABLE IV: Entity prediction for OOKB entities.

	rank of h	rank of t
h is OOKB	316	311
t is OOKB	301	311
h, t are OOKB	259	248
Overall	292	290

TABLE V: Classification result for OOKB entities.

	Accuracy(%)
h is OOKB	88.95
t is OOKB	90.50
h, t are OOKB	91.75
Overall	90.4

F. Further Analysis on Multimodal Knowledge

To further analyze the influence of external knowledge, we adjust the relative weight between autoencoder loss L_a and TransE loss L_e, and train our model under different weight settings respectively. Denote w as the relative weight autoencoder loss compared to TransE loss, we plot the learning curves under $w = 0.05, 0.15$ and 0.30 in Fig. 3. It can be observed that as the weight of autoencoder increases, model converges to a better result more quickly, showing that external multimodal knowledge of entities can improve both the representation quality and learning efficiency.

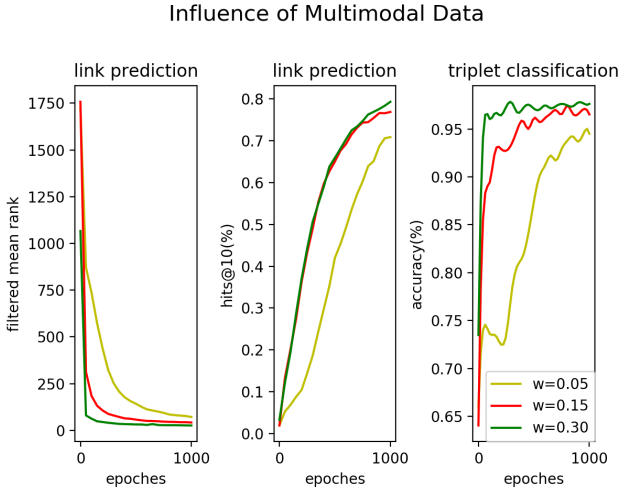

Fig. 3: Further analysis on multimodal knowledge. It shows that external multimodal knowledge can improve both representation quality and learning efficiency.

G. Case Study

In this section, we present a case study by performing multimodal retrieval experiment, it shows that our model can group similar entities together in the embedding space.

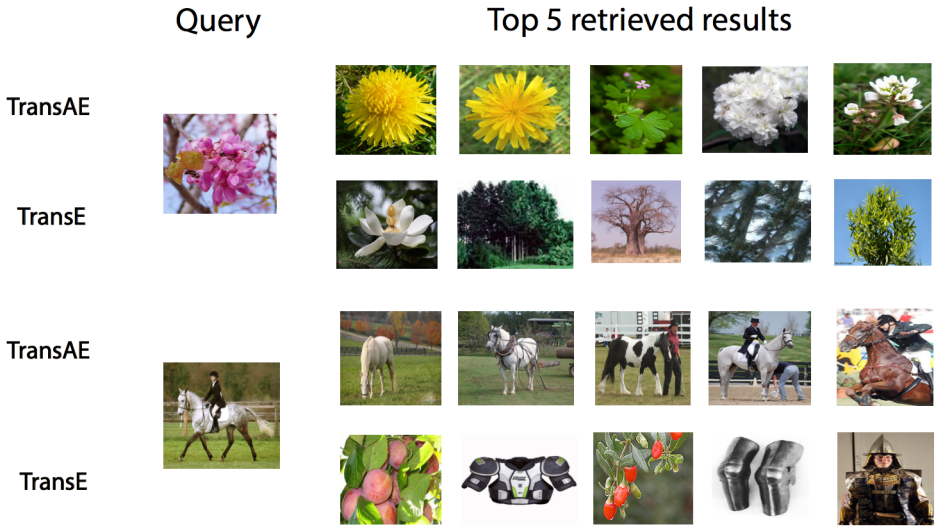

Fig. 4: Examples of multimodal query retrieval results. TransAE can aggregate similar entities together in the embedding space.

We first pre-process images and text respectively to get visual and textual representations, then feed these representations to our model to get the entity representations. We use kNN [21] to get the nearest 5 entities for an entity. To make comparison, we also use kNN to get the nearest 5 entities in representation space of TransE, with the optimal parameter setting claimed in its original paper. Examples of our results are shown in Fig. 4.

From Fig. 4, we found that the nearest entities TransAE found are also near in semantic space and similar in visual representations, indicating that when mapping entities and relations into embedding space, TransAE not only follow relational constraints, but also group similar entities together. As TransE model values little on entities themselves, very different entities can be near to each other in the embedding space. For example, for embeddings get by TransE, the nearest neighbor of entity "dressage" are either fruits or armors, which is not a desired illustration in the real world.

V. Conclusions and Future Work

In this paper, we proposed a new model, TransAE, to learn knowledge embeddings. By combining autoencoder and TransE, the proposed TransAE model can learn representations not only based on structural information between entities and relations, but visual and textual information of entities as well. It learns a joint representation based on all kinds of knowledge, thus can be easily adapted to multimodalities. Experiments on tasks such as link prediction and triplet classification validated the effectiveness of our method. Furthermore, TransAE can learn representations for entities out of the knowledge base in zero-shot.

The proposed TransAE model can be explored further in the following aspects: 1. In this paper, we integrate visual and textual knowledge from different sources, which may result in the mismatch of text and images, thus leading to the damage of representation. Matching text and images before learning representations may enhance the representation quality further. 2. Visual and textual descriptions are from existing databases in this paper, while extracting multimodal knowledge from open domain may bring more flexibility to the model. 3. We only focus on one entity in each image, ignoring relationships between entities when more than one entities are contained in the image. How to extract relations between these entities and represent them, still awaits further exploration.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2016QY02D0305 and 2017YFC0820105, the National Natural Science Foundation of China under Grants 71702181, 71621002, as well as the Key Research Program of the Chinese Academy of Sciences under Grant ZDRW-XH-2017-3.

References

[1] Miller, G. A., "WordNet: a lexical database for English," Communications of the ACM 38(11):3941, 1995.

[2] Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J., "Freebase: A collaboratively created graph database for structuring human knowledge," In Proceedings of the 2008 ACM SIGMOD International Conference on Manage- ment of Data, SIGMOD 08, 12471250, 2008.

[3] Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Hruschka, Jr., E. R.; and Mitchell, T. M., "Toward an architecture for never-ending language learning," In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI10, 13061313. Atlanta, Georgia: AAAI Press, 2010.

[4] Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O., "Translating embeddings for modeling multi-relational data," In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., Advances in Neural Information Processing Systems 26. Lake Tahoe, USA: Curran Associates, Inc. 27872795, 2013.

[5] Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z., "Knowledge graph embedding by translating on hyperplanes," In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Canada: AAAI Press, 2014.

[6] Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X., "Learning entity and relation embeddings for knowledge graph completion," In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI15, 21812187. Austin, Texas: AAAI Press, 2015.

[7] Ji, G.; He, S.; Xu, L.; Liu, K.; and Zhao, J.,"Knowledge graph embedding via dynamic mapping matrix," In ACL (1), 687696. Beijing, China: The Association for Computer Linguistics, 2015.

[8] Xie, R.; Liu, Z.; Jia, J.; Luan, H.; and Sun, M., "Representation learning of knowledge graphs with entity descriptions," In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI16, 26592665. Phoenix, Ari- zona: AAAI Press, 2016.

[9] Xie, R.; Liu, Z.; Luan, H.; and Sun, M., "Image-embodied knowledge representation learning," In Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI17, 31403146. Melbourne, Australia: AAAI Press, 2017.

[10] Deng, J.; Dong, W.; Soeher, R.; Li, L.-J.; Li, K.; and Li, F.-F., "ImageNet: A large-scale hierarchical image database," In CVPR, 248255. Florida, USA: IEEE Computer Society, 2009.

[11] Xiao, H.; Huang, M.; and Zhu, X., "TransG: A generative model for knowledge graph embedding," In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 23162325. Berlin, Germany: Association for Computational Linguistics, 2016.

[12] Nickel, M.; Tresp, V.; and Kriegel, H.-P., "A Nthree-way model for collective learning on multi-relational data," In Proceedings of the 28th International Conference on Inter- national Conference on Machine Learning, ICML11, 809 816. Bellevue, Washington, USA: Omnipress, 2011.

[13] Nickel, M.; Rosasco, L.; and Poggio, T., "Holographic embeddings of knowledge graphs," In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI16, 19551961. Phoenix, Arizona: AAAI Press, 2016.

[14] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J., "Distributed representations of words and phrases and their compositionality," In NIPS, 31113119, 2013.

[15] Simonyan, K., and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," In International Conference on Learning Representations. San Diego, CA: ICLR, 2015.

[16] Le, Q., and Mikolov, T., "Distributed representations of sentences and documents," In Xing, E. P., and Jebara, T., eds., Proceedings of the 31st International Conference on Machine Learning, Proceedings of Machine Learning Research, 11881196. Bejing, China: PMLR, 2014.

[17] Socher, R.; Chen, D.; Manning, C. D.; and Ng, A., "Reasoning with neural tensor networks for knowledge base completion," In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., Advances in Neural Information Processing Systems 26. Lake Tahoe, USA: Curran Associates, Inc. 926934, 2013.

[18] Bordes, A.; Glorot, X.; Weston, J.; and Bengio, Y., "A semantic matching energy function for learning with multi-relational data," Machine Learning 94(2):233259, 2014.

[19] Bordes, A.; Weston, J.; Collobert, R.; and Bengio, Y., "Learning structured embeddings of knowledge bases," In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI11, 301306, 2011.

[20] Srivastava, N., and Salakhutdinov, R., "Multimodal learning with deep boltzmann machines," Journal of Machine Learning Research 15:29492980, 2014.

[21] Altman, N. S., "An introduction to kernel and nearest-neighbor nonparametric regression," The American Statisti- cian 46:175185, 1992.

[22] Hamaguchi, T.; Oiwa, H.; Shimbo, M.; and Matsumoto, Y., "Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach," In Proceedings of the 26th International Joint Conference on Artificial Intelli- gence, IJCAI17, 18021808. Melbourne, Australia: AAAI Press, 2017.