A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning

Hatem Mousselly-Sergieh[†], **Teresa Botschen**[§†], **Iryna Gurevych**[§†], **Stefan Roth**[§‡]
† Ubiquitous Knowledge Processing (UKP) Lab
‡ Visual Inference Lab
§ Research Training Group AIPHES
Department of Computer Science, Technische Universität Darmstadt
h.m.sergieh@gmail.com
{botschen@aiphes, gurevych@ukp.informatik, stefan.roth@visinf}.tu-darmstadt.de

Abstract

Current methods for knowledge graph (KG) representation learning focus solely on the structure of the KG and do not exploit any kind of external information, such as visual and linguistic information corresponding to the KG entities. In this paper, we propose a multimodal translation-based approach that defines the energy of a KG triple as the sum of sub-energy functions that leverage both multimodal (visual and linguistic) and structural KG representations. Next, a ranking-based loss is minimized using a simple neural network architecture. Moreover, we introduce a new large-scale dataset for multimodal KG representation learning. We compared the performance of our approach to other baselines on two standard tasks, namely knowledge graph completion and triple classification, using our as well as the WN9-IMG dataset.[1] The results demonstrate that our approach outperforms all baselines on both tasks and datasets.

1 Introduction

Knowledge Graphs (KGs), e.g., Freebase (Bollacker et al., 2008) and DBPedia (Auer et al., 2007), are stores of relational facts, which are crucial for various kinds of tasks, such as question answering and information retrieval. KGs are structured as triples of head and tail entities along with the relation that holds between them. Factual knowledge is virtually infinite and is frequently subject to change. This raises the question of the incompleteness of the KGs. To address this problem, several methods have been proposed for automatic KG completion (KGC, for a survey refer to Wang et al., 2017). In recent years, translation-based approaches have witnessed a great success. Their main idea is to model the entities and their relation as low-dimensional vector representations (embeddings), which in turn can be used to perform different kinds of inferences on the KG. These include identifying new facts or validating existing ones. However, translation-based methods rely on the rich structure of the KG and generally ignore any type of external information about the included entities.

In this paper, we propose a translation-based approach for KG representation learning that leverages two different types of external, multimodal representations: *linguistic* representations created by analyzing the usage patterns of KG entities in text corpora, and *visual* representations obtained from images corresponding to the KG entities. To gain initial insights into the potential benefits of external information for the KGC task, let us consider the embeddings produced by the translation-based TransE method (Bordes et al., 2013) on the WN9-IMG dataset (Xie et al., 2017). This dataset contains a subset of WordNet synsets, which are linked according to a predefined set of linguistic relations, e.g. *hypernym*. We observed that TransE fails to create suitable representations for entities that appear frequently as the head/tail of one-to-many/many-to-one relations. For example, the entity *person* appears frequently in the dataset

Embedding Space	Top Similar Synsets
Linguistic	n02472987_world, n02473307_Homo_erectus, n02474777_Homo_sapiens, 02472293_homo, n00004475_organism, n10289039_man
Visual	n10788852_woman, n09765278_actor, n10495167_pursuer n10362319_nonsmoker, n10502046_quitter, n09636339_Black
Structure (TransE)	_hypernym, n00004475_organism, n03183080_device, n07942152_people, n13104059_tree, n00015388_animal, n12205694_herb, n07707451_vegetable

Table 1: Closest synsets to the person synset (n00007846) according to different embedding spaces.

[1] Code and datasets are released for research purposes: https://github.com/UKPLab/starsem18-multimodalKB

*Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 225–234
New Orleans, June 5-6, 2018.

as a head/tail of the *hyponym/hypernym* relation; the same holds for entities like *animal* or *tree*. TransE represents such entities as points that are very close to each other in the embedding space (cf. Tab. 1). Furthermore, the entity embeddings tend to be very similar to the embeddings of relations in which they frequently participate. Consequently, such a representation suffers from limited discriminativeness and can be considered a main source of error for different KG inference tasks.

To understand how multimodal representations may help to overcome this issue, we performed the same analysis by considering two types of external information: linguistic and visual. The linguistic representations are created using word embedding techniques (Mikolov et al., 2013), and the visual ones, called visual embeddings, are obtained from the feature layers of deep networks for image classification (e.g., Chatfield et al., 2014) on images that correspond to the entities of the dataset. For the same category of entities discussed above, we observed that both the visual and the linguistic embeddings are much more robust than the structure-based embeddings of TransE. For instance, *person* is closer to other semantically related concepts, such as *Homo_erectus* in the linguistic embedding space, and to concepts with common visual characteristics (e.g., *woman*, *actor*) in the visual embedding space (cf. Tab. 1). Furthermore, the linguistic and the visual embeddings seem to complement each other and hence are expected to enhance KG representations if they can be leveraged during the representation learning process.

The contributions of this paper can be summarized as follows: *(1)* We propose an approach for KG representation learning that incorporates multimodal (visual and linguistic) information in a translation-based framework and extends the definition of triple energy to consider the new multimodal representations; *(2)* we investigate different methods for combining multimodal representations and evaluate their performance; *(3)* we introduce a new large-scale dataset for multimodal KGC based on Freebase; *(4)* we experimentally demonstrate that our approach outperforms baseline approaches including the state-of-the-art method of Xie et al. (2017) on the link prediction and triple classification tasks.

2 Related Work

2.1 Translation Models

TransE (Bordes et al., 2013) is among the earliest translation-based approaches for KG representation learning. TransE represents entities and relations as vectors in the same space, where the relation is considered a translation operation from the representation of the head to that of the tail entity. For a correct triple, TransE assumes that $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$, where $\boldsymbol{h}$, $\boldsymbol{r}$, $\boldsymbol{t}$ are the vector representations of the head, relation, and tail, respectively. Additionally, TransE use a dissimilarity measure d to define the energy of a given triple as $d(\boldsymbol{h}+\boldsymbol{r}, \boldsymbol{t})$. Finally, the representations of KG entities and relations are learned by minimizing a margin-based ranking objective that aims to score positive triples higher than negative triples based on their energies and a predefined margin.

TransE is a simple and effective method, however, the simple translational assumption constrains the performance when dealing with complex relations, such as one-to-many or many-to-one. To address this limitation, some extensions of TransE have been proposed. Wang et al. (2014) introduced TransH, which uses translations on relation-specific hyperplanes and applies advanced methods for sampling negative triples. Lin et al. (2015b) proposed TransR, which uses separate spaces for modeling entities and relations. Entities are projected from their space to the corresponding relation space by relation-specific matrices. Moreover, they propose an extension called CTransR, in which instances of pairs of head and tail for a specific relation are clustered such that the members of the clusters exhibit similar meanings of this relation. Lin et al. (2015a) proposed another extension of TransE, called PTransE, that leverages multi-step relation path information in the process of representation learning.

The above models rely only on the structure of the KG, and learning better KG representations is dependent upon the complexity of the model. In this paper, however, we follow a different approach for improving the quality of the learned KG representation and incorporate external multimodal information in the learning process, while keeping the model as simple as possible.

2.2 Multimodal Methods

Recent advances in natural language processing have witnessed a greater interest in leveraging

multimodal information for a wide range of tasks. For instance, Shutova et al. (2016) showed that better metaphor identification can be achieved by fusing linguistic and visual representations. Collell et al. (2017) demonstrated the effectiveness of combining linguistic and visual embeddings in the context of word relatedness and similarity tasks. Regarding KG representation learning, the first and, to the best of our knowledge, only attempt that considers multimodal data is the work of Xie et al. (2017). Their IKRL approach extends TransE based on visual representations extracted from images that correspond to the KG entities. In IKRL, the energy of a triple is defined in terms of the structure of the KG as well as the visual representation of the entities. Our work, while building upon the foundations of Xie et al. (2017), sets itself apart based on the following properties: *(1)* in addition to images, our model integrates another kind of external representation, namely linguistic embeddings for KG entities – thus, adding multimodal information; *(2)* we base our approach on a simple and easily extensible neural network architecture; *(3)* we introduce an additional energy function that considers the multimodal representation of the KG entities; *(4)* we introduce a new large-scale dataset for multimodal KG representation learning.

3 Proposed Approach

We denote the knowledge graph as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is the set of entities, $\mathcal{R}$ is the set of relations, and $\mathcal{T} = \{(h, r, t)|h, t \in \mathcal{E}, r \in \mathcal{R}\}$ the set of KG triples. For each head and tail entity $h, r \in \mathcal{E}$, we define three kinds of representations (embeddings): structural $\boldsymbol{h_s^I}, \boldsymbol{t_s^I} \in \mathbb{R}^N$, linguistic $\boldsymbol{h_w^I}, \boldsymbol{t_w^I} \in \mathbb{R}^M$, and visual $\boldsymbol{h_i^I}, \boldsymbol{t_i^I} \in \mathbb{R}^P$, where N, M and P are the corresponding numbers of dimensions. Furthermore, we represent each relation $r \in \mathcal{R}$ as a vector $\boldsymbol{r_s^I} \in \mathbb{R}^N$ in the space of the structural information. The superscript $\boldsymbol{I}$ denotes that these are the input embeddings. Since the different embeddings do not live in the same space, we assume from now on that they can be transformed into a common space using a multilayer network (e.g., $\boldsymbol{h_s^I}$ into $\boldsymbol{h_s}$, cf. Fig. 1). Following the translational assumption, given a triple (h, r, t), we have

$$\boldsymbol{h_s} + \boldsymbol{r_s} \approx \boldsymbol{t_s}. \tag{1}$$

3.1 Model

In general, previous works such as (Bordes et al., 2013) start from Eq. (1) and build models for minimizing a ranking loss between positive and negative triples that are sampled from the KG. Conventionally, negative triples are sampled by corrupting the head, the tail, or the relation of correct triples. We follow this idea and make it explicit by taking two different "views" on the translational assumption. Apart from the first view through Eq. (1), we can also rewrite the translational assumption as

$$\boldsymbol{t_s} - \boldsymbol{r_s} \approx \boldsymbol{h_s}. \tag{2}$$

We will learn the two views jointly. For each view, we sample specific kinds of negative triples according to which part of the triple has to be predicted. For the head-centric view, we define $\mathcal{T}'_{\text{tail}}$, a set of negative triples that is sampled by corrupting the tail of gold triples. Similarly, for the tail-centric view, we define $\mathcal{T}'_{\text{head}}$, a set of negative triples sampled by corrupting the head of the gold triples:

$$\mathcal{T}'_{\text{tail}} = \{(h, r, t')|h, t' \in \mathcal{E} \wedge (h, r, t') \notin \mathcal{T}\} \tag{3a}$$
$$\mathcal{T}'_{\text{head}} = \{(h', r, t)|h', t \in \mathcal{E} \wedge (h', r, t) \notin \mathcal{T}\}. \tag{3b}$$

Next, we extend the definition of triple energy in order to integrate both the structural and the multimodal representations of the KG entities. For each kind of representation as well as their combination, we define a specific energy function. Subsequently, the final energy of a triple is defined as the sum of the individual energies defined below.

Structural Energy: The structure-based energy of a triple is defined in terms of the structure of the KG as proposed by the TransE approach (Bordes et al., 2013). Accordingly, we define

$$E_S = \|\boldsymbol{h_s} + \boldsymbol{r_s} - \boldsymbol{t_s}\|. \tag{4}$$

Multimodal Energies: The multimodal representation of a KG entity is defined by combining the corresponding linguistic and visual representations. Let $\oplus$ denote the combination operator (more details in Section 3.2). Now, we define the multimodal representations $\boldsymbol{h_m}$ and $\boldsymbol{t_m}$ of the head and the tail entities, respectively, as

$$\boldsymbol{h_m} = \boldsymbol{h_w} \oplus \boldsymbol{h_i} \tag{5a}$$
$$\boldsymbol{t_m} = \boldsymbol{t_w} \oplus \boldsymbol{t_i}. \tag{5b}$$

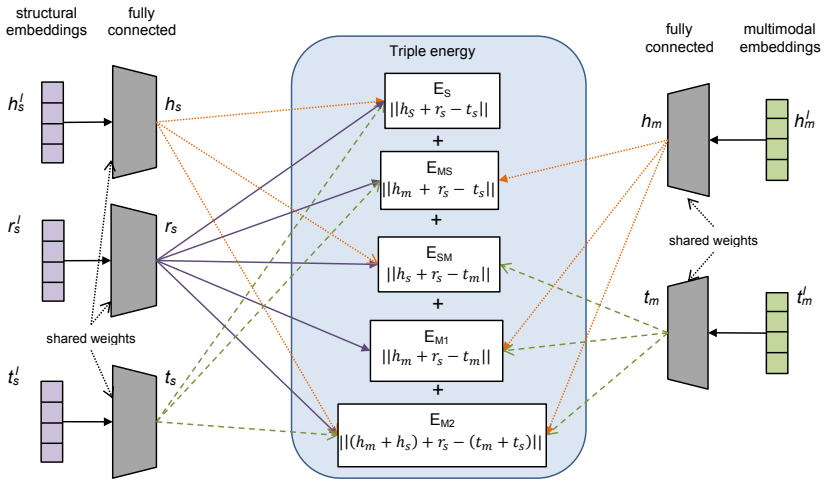

Figure 1: Overview of the neural network architecture for calculating the total triple energy from the different models. The fully connected networks transform the respective input embeddings into a common space.

Next, we transfer the structure-based energy function from Eq. (4) to the multimodal case where it incorporates the multimodal representations under the translational assumption, i.e.

$$E_{M1} = \|\boldsymbol{h_m} + \boldsymbol{r_s} - \boldsymbol{t_m}\|. \quad (6)$$

We then extend the previous energy from Eq. (6) to define another energy function that considers the structural embeddings in addition to the multimodal ones as follows:

$$E_{M2} = \|(\boldsymbol{h_m} + \boldsymbol{h_s}) + \boldsymbol{r_s} - (\boldsymbol{t_m} + \boldsymbol{t_s})\|. \quad (7)$$

These presented multimodal energies can be understood as additional constraints for the translation model. $M1$ states that the relation corresponds to a translation operation between the multimodal representation of the head and the tail entities once projected into the structural space. $M2$ enforces that same constraint, however, on the sum of the multimodal and the structural embeddings of the head and the tail entities. While Eqs. (4), (6), and (7) cannot be fulfilled at the same time, we found that combining these complementary energies makes the results more robust.

Structural-Multimodal Energies: Next, to ensure that the structural and the multimodal representations are learned in the same space, we follow the proposal of Xie et al. (2017) and define the following energy functions:

$$E_{SM} = \|\boldsymbol{h_s} + \boldsymbol{r_s} - \boldsymbol{t_m}\| \quad (8a)$$
$$E_{MS} = \|\boldsymbol{h_m} + \boldsymbol{r_s} - \boldsymbol{t_s}\|. \quad (8b)$$

Finally, the overall energy of a triple for the head and the tail views are defined as

$$E(h,r,t) = E_S + E_{M1} + E_{M2} + E_{SM} + E_{MS}. \quad (9a)$$

Objective Function: For both the head and the tail view, we aim to minimize a margin-based ranking loss between the energies of the positive and the negative triples. The corresponding loss functions are finally defined as

$$\mathcal{L}_{head} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h,r,t')\in\mathcal{T}'_{\text{tail}}} \max\left(\gamma + E(h,r,t) - E(h,r,t'), 0\right) \quad (10)$$

$$\mathcal{L}_{tail} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t)\in\mathcal{T}'_{\text{head}}} \max\left(\gamma + E(t,-r,h) - E(t,-r,h'), 0\right). \quad (11)$$

Here, γ is a margin parameter, which controls the amount of energy difference between the positive and the negative triples. Finally, we aim to minimize the global loss

$$\mathcal{L} = \mathcal{L}_{\text{head}} + \mathcal{L}_{\text{tail}}. \quad (12)$$

To bring the different representations (structural, linguistic, visual) into the same space, we employ a simple feed-forward neural network architecture. The input of the network consists of the structural and the multimodal embeddings of the heads, the tails, and the relations (Fig. 1); the fully-connected layers map these inputs into a common space. Furthermore, we share the weights between

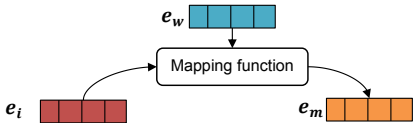

Figure 2: The DeViSE method (Frome et al., 2013).

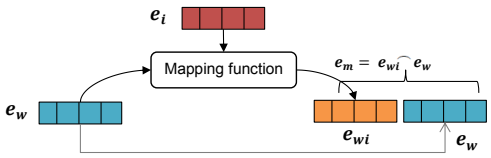

Figure 3: The Imagined method (Collell et al., 2017).

those fully-connected layers that receive the same kind of input. Additionally, the weights are also shared across the head and the tail views.

3.2 Combining Multimodal Representations

To complete the description of our approach, we still need to define the $\oplus$ operator used in Eq. (5) to combine the linguistic and visual embeddings into a single one. To that end, we identified three methods for multimodal representation learning and adapted them to KG entities.

Concatenation Method: The simplest method to create multimodal representations for KG entities is to combine the multimodal embedding vectors by concatenation. Given the linguistic $\boldsymbol{e_w}$ and the visual $\boldsymbol{e_i}$ embeddings of an entity e, we define the multimodal representation $\boldsymbol{e_m} = \boldsymbol{e_w} \frown \boldsymbol{e_i}$, where $\frown$ is the concatenation operator.

DeViSE Method: Next, we consider the deep visual-semantic embedding model (DeViSE) of Frome et al. (2013), which leverages textual data to explicitly map images into a rich semantic embedding space. Given the visual representation of some concept, the goal is to learn a mapping into the linguistic (word) embedding space. The mapped representation can then be used as a multimodal representation for the target entity. Fig. 2 illustrates the application of DeViSE to generating multimodal representations for KG entities.

Imagined Method: Finally, we consider the Imagined method of Collell et al. (2017) for creating multimodal representations of concepts based on their linguistic and visual embeddings. Imagined is similar to DeViSE, however, it applies the reverse procedure. That is, for a given concept Imagined aims to learn a mapping from the linguistic embedding space of that concept into the visual embedding space. The mapping can be formulated as a linear or nonlinear transformation using a simple neural network, and the objective is to minimize the distance between the mapped linguistic representation and the visual representation of the entities. Subsequently, a multimodal representation is created by applying the learned mapping function on the linguistic representation of the entity and then concatenating the resulting vector with the original linguistic embedding (Fig. 3).

Dataset	#Rel	#Ent	#Train	#Valid	#Test
WN9-IMG	9	6555	11 741	1337	1319
FB-IMG	1231	11 757	285 850	29 580	34 863

Table 2: Datasets statistics

4 Experiments

4.1 Datasets

WN9-IMG: This dataset provided by Xie et al. (2017) is based on WordNet. It contains a collection of triples, where the entities correspond to word senses (synsets) and the relations define the lexical relationships between the entities. Furthermore, for each synset a collection of up to ten images obtained from ImageNet (Deng et al., 2009) is provided.

FB-IMG: To demonstrate the scalability of our approach to larger datasets, we created another dataset based on FB15K (Bordes et al., 2013), which consists of triples extracted from Freebase. For each entity, we crawled 100 images from the web using text search based on the entity labels. To ensure that the crawled images are representative of the corresponding entities, we applied an approach for image filtering based on the PageRank algorithm (Page et al., 1999). First, we created a vector representation (embedding) for each image by feeding it into a pre-trained VGG19 neural network for image classification (Simonyan and Zisserman, 2014). The image embeddings consist of the 4096-dimensional activation of the last layer (before the softmax). Next, for each entity we create a similarity graph for the corresponding images based on the cosine similarity between their embedding vectors. Finally, we calculated the PageRank score for each image in the

graph and kept the top 10 results. Tab. 2 gives basic statistics of the two datasets.

4.2 Representations

We now discuss the procedure we followed to obtain different kinds of representations for the entities and relations of the two evaluation datasets.

Structural Representation: This baseline representation is created based on the structure of the KG only, without any external information. In our experiments we created structure representations for the entity and the relations of the two datasets using the TransE algorithm. For both datasets, we trained TransE with 100 dimensions and used the same values for the other hyperparameters as recommended by Bordes et al. (2013).

Linguistic Representation: The linguistic representations of the entities are obtained by applying word embedding techniques. For the FB-IMG dataset, we used a pre-trained word embedding model for Freebase entities as provided by the word2vec framework (Mikolov et al., 2013). The provided embeddings are 1000 dimensional and are trained using the skipgram model over the Google 100B token news dataset. We applied L_2-normalization on the generated embeddings.

The entities of the WN9-IMG dataset correspond to word senses rather than to individual words. In order to create embeddings for the synsets, we used the AutoExtend framework (Rothe and Schütze, 2015), which enables creating sense embeddings for a given sense based on the embeddings of the contained lemmas. For this purpose, we initialized AutoExtend with pre-trained 300-dimensional GloVe embeddings (Pennington et al., 2014). In case where no pre-trained embeddings are found for the sense lemmas, AutoExtend generates zero initialized vectors for the corresponding synsets. In order to provide better representations, we define the embeddings of such synsets by copying the embeddings of the first hyperonym synset that has non-zero AutoExtend embeddings. The linguistic embeddings of WN9-IMG entities (synsets) are 300-dimensional vectors, which were also L_2-normalized.

Visual Representation: For each image of a given KG entity, we created a visual embedding vector using the same procedure as for creating the FB-IMG dataset. This was done using a pre-trained VGG model (Simonyan and Zisserman, 2014). For the WN9-IMG dataset, we used the VGG19 model and extracted the 4096-dimensional vector of the last fully-connected layer before the softmax. For the FB-IMG dataset, which contains much more data than WN9-IMG and in order to speed up the training, we used the more compact VGG-m-128 CNN model (Chatfield et al., 2014), which produces 128-dimensional embedding vector for each image. Next, the visual embeddings are L_2-normalized. We investigated two ways of combining the embedding vectors corresponding to images of a given entity. The first method defines the visual embedding of an entity as the average of the embeddings of all corresponding images. The second method uses the dimension-wise maximum. In our experiments we observed that averaging the embedding vectors outperforms the maximum method. Hence, we only report the results obtained with averaging.

4.3 Experimental Setup

We investigated different sets of hyperparameters for training the model. The best results were obtained using the Adam optimizer (Kingma and Ba, 2014) with a fixed learning rate of 0.001 and batch size of 100. We used the hyperbolic tangent function ($tanh$) for the activation and one fully-connected layer of 100 hidden units. We observed that regularization has a minor effect. In the case of WN9-IMG, we used dropout regularization (Srivastava et al., 2014) with a dropout ratio of 10%; we applied no regularization on the FB-IMG dataset. Regarding the margin of the loss function, we experimented with several values for both datasets $\gamma \in \{4, 6, 8, 10, 12\}$. The best results for both datasets were obtained with $\gamma = 10$.

We investigated different configurations of our approach: *(1) Ling* considers the linguistic embeddings only, *(2) Vis* considers the visual embeddings only, *(3) multimodal* where the visual and the linguistic embeddings are considered according to the presented multimodal combination methods: *DeViSE*, *Imagined*, and the *Concatenation* methods (cf. Sec. 3.2), and *(4) only head* in which we use the head view only and the *concatenation* method for combining the multimodal representations. Here, negative samples are produced by randomly corrupting the head, the tail, or the relation of gold triples.

We compared our approach to other baseline

methods including TransE (Bordes et al., 2013) and IKRL (Xie et al., 2017). For TransE, we set the size of the embeddings to 100 dimensions and followed the recommendations of Bordes et al. (2013) regarding the other hyperparameters. We also implemented the IKRL approach and the best results were achieved by using margins of 8 and 4 for the WN9-IMG and the FB-IMG datasets, respectively. We tested two configurations of IKRL: *(1) IKRL (Vis)* uses the visual representation only (as in the original paper) and initializes the structural representations with our learned TransE embeddings, and *(2) IKRL (Concat)*, which uses the concatenation of the linguistic and the visual embeddings. Please note that we do not apply the attention mechanism for creating image representations as proposed in the IKRL paper (Xie et al., 2017). However, we include that model, referred to as *IKRL (Paper)*, in the comparison.

4.4 Link Prediction

Evaluation Protocol: Given a pair of a head/tail and a relation, the goal of link prediction is to identify the missing tail/head. For each test triple, we replaced the head/tail by all entities in the KG and calculated the corresponding energies in ascending order. Similar to Bordes et al. (2013), we calculated two measures: *(1)* the mean rank (MR) of the correctly predicted entities and *(2)* the proportion of correct entities in the top-10 ranked ones (Hits@10). We also distinguished between two evaluation settings, "Raw" and "Filter". In contrast to the "Raw" setting, in the "Filter" setting correct triples included in the training, validation, and test sets are removed before ranking.

Results: Tab. 3 shows the results on the WN9-IMG dataset. First, we can observe that leveraging multimodal information leads to a significant improvement compared to the structure-only based approach TransE, especially in terms of the mean rank. This conclusion is in accordance with our intuition: although the structural representations become less discriminative after the training for certain kinds of entities (such as the one discussed in Sec. 1), the multimodal representations compensate for this effect, thus the prediction accuracy increases. Regarding the multimodal representations, combining the linguistic and the visual embeddings seems to outperform models that rely on only one kind of those representations. This holds for our approach as well as for IKRL. Regarding the multimodal combination method, we surprisingly noticed that the simple concatenation method outperforms other advanced methods like DeViSE (Frome et al., 2013) and Imagined (Collell et al., 2017). This suggests that translation-based approaches for KG representation learning profit more from the raw representations than general purpose pre-combined ones, which are not necessarily tuned for this task.

Method	**MR**		**Hits@10 (%)**	
	Raw	**Filter**	**Raw**	**Filter**
TransE	160	152	78.77	91.21
IKRL (Paper)	28	21	80.90	93.80
IKRL (Vis)	21	15	81.39	92.00
IKRL (Concat)	18	12	82.26	93.25
Our (Ling)	19	13	80.78	90.79
Our (Vis)	20	14	80.74	92.30
Our (DeViSE)	19	13	81.80	93.21
Our (Imagined)	19	14	81.43	91.09
Our (Concat)	**14**	**9**	**83.78**	**94.84**
Our (only head)	19	13	82.37	93.21

Table 3: Link prediction results on WN9-IMG.

The evaluation also shows that our approach with the concatenation method outperforms the best IKRL model, IKRL (Concat), which was trained on the same representations as our approach. Additionally, our model outperforms the best performing IKRL model reported in (Xie et al., 2017) with less than half the MR and more than one point in Hits@10. This shows the benefit of our additional energy term coupling structural and multimodal embeddings. To assess the benefit of taking two separate views on the translational assumption, we evaluated the performance of using the head view only. We observe a considerable drop in performance. The MR becomes 5 points higher and the Hits@10 drops by more than one percentage point compared to the same model that is trained using both the head and the tail views.

Compared to WN9-IMG, the FB-IMG dataset has a much larger number of relations, entities, and triples (cf. Tab. 2), thus it better resembles the characteristics of real KG. On the FB-IMG dataset, the superiority of our model compared to the baselines, especially IKRL, becomes even more evident (cf. Tab. 4). Our model performs best and achieves a significant boost in MR and Hits@10 compared to the baselines, while IKRL slightly outperforms TransE in terms of MR only.

Method	MR		Hits@10 (%)	
	Raw	Filter	Raw	Filter
TransE	205	121	37.83	49.39
IKRL (Concat)	179	104	37.48	47.87
Our (Concat)	**134**	**53**	**47.19**	**64.50**

Table 4: Link prediction results on FB-IMG.

Therefore, the results confirm the robustness of our method for large-scale datasets.

Finally, we observe that, in general, the performance boost on the FB-IMG dataset is lower than in the case of the WN9-IMG dataset. This can be explained by the higher scale and complexity of the FB-IMG dataset. Furthermore, the visual representations of the FB-IMG entities are based on images that are automatically crawled from the Web. Accordingly, some of the crawled images may not be representative enough or even noisy, while the images in WN9-IMG have better quality since they are obtained from ImageNet, which is a manually created dataset.

4.5 Triple Classification

Evaluation Protocol: Triple classification is a binary classification task, in which the KG triples are classified as correct or not according to a given dissimilarity measure (Socher et al., 2013). For this purpose a threshold for each relation δ_r is learned. Accordingly, a triple (h, r, t) is considered correct if its energy is less than δ_r, and incorrect otherwise. Since the dataset did not contain negative triples, we followed the procedure proposed by Socher et al. (2013) to sample negative triples for both the validation and the test sets. As a dissimilarity measure, we used the total energy of the triple and determined the relation threshold using the validation set and then calculated the accuracy on the test set.

Results: We measured the triple classification accuracy of our approach using ten test runs. In each run, we sampled new negative triples for both the validation and the test sets. We report the maximum, the minimum, the average, and the standard deviation of the triple classification accuracy.

For WN9-IMG, the results (cf. Tab. 5) show that our approach outperforms the baselines with up to two points in maximum accuracy and around three points in average accuracy. Please note that a direct comparison with IKRL (Paper) is not possible since we do not have access to the same set of negative samples. Still, the maximum classification accuracy of our approach is higher than that of by IKRL (Paper). Finally, the results confirm that using separate head and tail views leads to better results than using the head view only.

Method	Accuracy(%)		
	max	min	avg $\pm$ std
TransE	95.38	89.67	93.35 ± 1.54
IKRL (Paper)	96.90	–	–
IKRL (Vis)	95.16	88.75	92.57 ± 1.78
IKRL (Concat)	95.40	91.77	93.56 ± 1.03
Our (Concat)	**97.16**	**94.93**	$\mathbf{96.10 \pm 0.87}$
Our (only head)	95.58	91.78	93.14 ± 1.09

Table 5: Triple classification results on WN9-IMG.

Method	Accuracy(%)		
	max	min	avg $\pm$ std
TransE	67.13	66.47	66.81 ± 0.21
IKRL (Concat)	66.68	66.03	66.34 ± 0.20
Our (Concat)	**69.04**	**68.16**	$\mathbf{68.62 \pm 0.25}$

Table 6: Triple classification results on FB-IMG.

Regarding the FB-IMG dataset, the results in Tab. 6 emphasize the advantage of our approach. Compared to the multimodal approach IKRL, which fails to outperform TransE, our model employs multimodal information more effectively and leads to more than one point improvement in average accuracy compared to TransE.

In conclusion, the conducted evaluation demonstrates the robustness of our approach on both evaluation tasks and on different evaluation datasets.

5 Conclusion

In this paper, we presented an approach for KG representation learning that leverages multimodal data about the KG entities including linguistic as well as visual representations. The proposed approach confirms the advantage of multimodal data for learning KG representations. In future work, we will investigate the effect of multimodal data in the context of advanced translation methods and conduct further research on combining visual and linguistic features for KGs.

Acknowledgments

The first author was supported by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1416B (CEDIFOR). Other than this, the work has been supported by the DFG-funded research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1). Finally, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the International Conference on Management of Data (ACM SIGMOD)*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 2787–2795.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Proceedings of the 31st Conference on Artificial Intelligence (AAAI)*, pages 4378–4384.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. DeViSE: A deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 2121–2129.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Yankai Lin, Zhiyuan Liu, Huan-Bo Luan, Maosong Sun, Siwei Rao, and Song Liu. 2015a. Modeling relation paths for representation learning of knowledge bases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 705–714.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015b. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI)*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1793–1803.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 160–170.

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

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Proceedings of the 26th International Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 926–934.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research (JMLR)*, 15(1):1929–1958.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the 28th Conference on Artificial Intelligence (AAAI)*, pages 1112–1119.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017. Image-embodied knowledge representation learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3140–3146.