Neural, Multimodal, Energy-based Approach for Knowledge Graph Completion

Hatem Mousselly-Sergieh, Iryna Gurevych & Stefan Roth –TU Darmstadt, Germany

Introduction

Knowledge Graphs (KGs) are stores of facts represented as triples (h, r, t) of **h**ead and **t**ail entities as well as a **r**elation that holds between them. Although KGs with high coverage already exist, KGs are still far from complete.

Several approaches for automatic KG completion have been proposed recently (e.g. (Bordes et al., 2013), (Wang et al., 2014)). In general, most approaches rely on the structure of the KG (represented by the included triples) and use variations of the translation model *TransE* proposed by (Bordes et al., 2013). Given a triple (h, r, t), *TransE* models the head, the tail and the relation as vectors (embeddings) in a continuous space. Thereby, the relation vector **r** is considered as a translation from the head vector **h** to the tail vector **t**. For a gold triple, *TransE* assumes that $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$. Accordingly, for each triple an energy score is defined as $d(\mathbf{h} + \mathbf{r}, \mathbf{t})$ where d is a dissimilarity measure. To learn the representations of KG entities and relations, margin-based ranking criterion over the training set is minimized.

Most recently, (Xie et. al., 2016) proposed an approach called *IKRL* that extends *TransE* based on external information obtained from images about KG entities. To the best of our knowledge this is the first and the only work which considers multimodal data for the KG completion task. *IKRL* builds upon *TransE* and defines the energy of a triple based on the structure of the KG (as in *TransE*), entity image information, as well as a combination thereof. The model is trained using the same loss function as in *TransE*. The authors experimentally demonstrated that combining image and structure information not only outperforms KG completion methods that leverage structure information only, but also results in a better incorporation of structure information for creating KG representations.

In this work, we propose an approach for KG completion that extends the approach of (Xie et. al., 2016). Our approach leverages multimodal information on KG entities including 1) visual features which are obtained using state-of-the-art convolutional neural network models for image classification and 2) textual representations which are learned using word embedding techniques. Moreover, we propose an additional energy function that combines multimodal features. Finally, we use a neural network architecture in order to learn the corresponding KG representations. We experimentally demonstrate the effectiveness of our approach and compare its performance to other baseline models.

Approach

We denote with $\mathcal{G} = (E, R, T)$ the knowledge graph, where E is the set of entities, R is the set of relations and $T = \{(h, r, t) | h, t \in E \wedge r \in R\}$ the set of KG triples. For each head and tail entity $h, r \in E$, we define three kinds of representations (embeddings), structure-based $\mathbf{h}_s, \mathbf{t}_\mathbf{s} \in \mathbb{R}^N$, text-based $\mathbf{h}_\mathbf{w}, \mathbf{t}_\mathbf{w} \in \mathbb{R}^M$ and image-based $\mathbf{h}_\mathbf{i}, \mathbf{t}_\mathbf{i} \in \mathbb{R}^P$. Furthermore, we represent each relation $r \in R$ as a vector $r_s \in \mathbb{R}^N$. We propose a model that leverages the presented kinds of representations by defining a set of energy functions based on the idea of the translation model (Bordes et al., 2013). Figure 1 shows the overall architecture of the proposed model.

Structure-based Energy E_s: E_s calculates the energy of a given triple (h, r, t) based on the structural information only as in (Bordes et al., 2013). We use the cosine similarity as a scoring function: $E_s = cos(\mathbf{h}_\mathbf{s} + \mathbf{r}_\mathbf{s}, \mathbf{t}_\mathbf{s})$.

Multimodal Energies E_{m1}, E_{m2}: we create multimodal representations of the head $\mathbf{h}_\mathbf{m}$and the tail $\mathbf{t}_\mathbf{m}$entities by concatenating the corresponding textual and visual representations: $\mathbf{h}_\mathbf{m} = \mathbf{h}_\mathbf{w} \oplus \mathbf{h}_\mathbf{i}$ and $\mathbf{t}_\mathbf{m} = \mathbf{t}_\mathbf{w} \oplus \mathbf{t}_\mathbf{i}$ where $\oplus$ is the concatenation operator. Next, we define $E_{m1} = cos(proj(\mathbf{h}_\mathbf{m}) + \mathbf{r}_\mathbf{s}, proj(\mathbf{t}_\mathbf{m}))$, where $proj$ is a projecting function that maps the multimodal representation of the entities into the relation space. We model $proj$ using a dense neural network layer. E_{m2} is similar to E_{m1}, however, instead of using the sum, it concatenates the head and relation representations. This type of combination can be seen as a compensation for possible information loss caused by the sum: $E_{m2} = \cos(\mathbf{h}_\mathbf{m} \oplus \mathbf{h}_\mathbf{s} \oplus \mathbf{r}_\mathbf{s}, \mathbf{t}_\mathbf{m})$.

Structure-Multimodal Energies E_{sm} and E_{ms}: To ensure that the structural and multimodal representations are learned in the same space, we follow the proposal of (Xie et al., 2016) and define the two energy functions: $E_{sm} = cos(\mathbf{h}_\mathbf{s} + \mathbf{r}_\mathbf{s}, proj(\mathbf{t}_\mathbf{m}))$ and $E_{sm} = cos(proj(\mathbf{h}_\mathbf{m}) + \mathbf{r}_\mathbf{s}, \mathbf{t}_\mathbf{s})$.

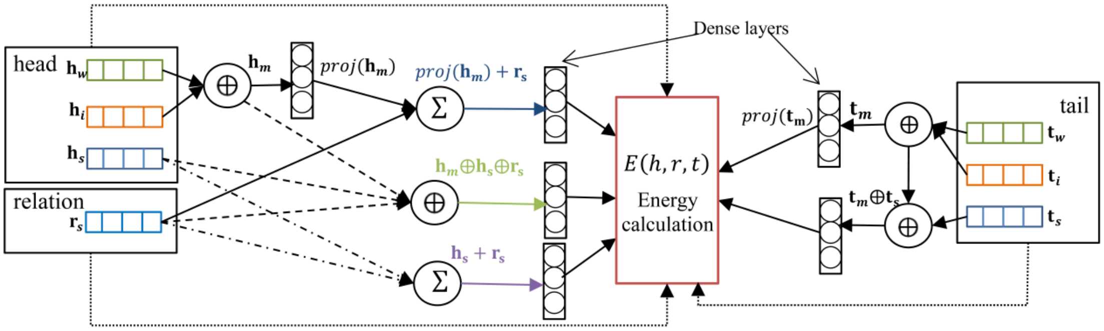
Finally, the overall energy function is defined as: $E(h, r, t) = E_s + E_{m1} + E_{m2} + E_{sm} + E_{ms}$

Training Objective: We define our objective as the hinge loss between the energies of positive and negative triples. For this purpose, we create a set of negative triples T' by corrupting the head or tail entities of the triples in T and ensuring that the new triples are not contained in the KG: $T' = \{(h', r, t) \cup (h, r, t') | h', t' \in E \wedge (h', r, t) \notin T \wedge (h, r, t') \notin T\}$. The corresponding objective function is then given as:

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

γ is a margin parameter which controls the amount of energy difference between the positive and the negative triples.

Figure 1: Overall architecture of the proposed model



Experiments

We used the WN9-IMG dataset (Xie et al., 2016) which contains triples that link a subset of WordNet synsets (entities) according to 9 different relations. For each entity, a maximum of 10 images is collected from ImageNet.

- *Structural representation:* we trained the TransE (Bordes et al., 2013) system on the WN9-IMG dataset to create the structural representation of the relations and the entities. We used 100 embedding dimensions and set all other hyperparameters to the values that were recommended by (Bordes et al., 2013).
- *Visual representation:* for each image of a given KG entity, we extracted visual features using the pre-trained VGG-m-128 CNN model (Chatfield et al. 2014). The image embeddings consist of the 128-dimensional activation of the last layer (before the softmax). Subsequently, we take the average of the embeddings of the images corresponding to each entity and apply L2-normalization to create the final visual representation.
- *Textual representation:* we used the AutoExtend (Rothe et al., 2015) framework to construct word embeddings for each synset based on the GloVe embeddings (Jeffrey et al., 2014) of the synset lemmas. We also apply L2-normalizatoin on the generated synset embeddings.

Finally, we trained the model using Adam optimizer. We set the learning rate to 0.001 and the margin $\gamma = 2$.

Results

A standard procedure to evaluate KG completion approaches is the link prediction task. Given a pair of a head/tail and a relation, the goal is to predict the missing tail/head. For each test triple, we replaced the head/tail by all entities in the KG, calculated the corresponding scores and ordered the results in the descending order of scores (energies). In a similar manner 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 the ranking. We compared our approach to *transE* (Bordes et al., 2013) and *IKRL* (Xie et. al., 2016). The results in Table 1 show that, in general, multimodal information leads to a significant improvement, especially, in terms of the mean rank. Our model outperforms *TransE* in terms of the mean rank and the hits@10 for both the raw and the filter settings. Although our model fails to beat *IKRL* in terms of hits@10, it significantly outperforms it in terms of the mean rank. According to (Xie et al., 2016), the mean rank metric is more sensitive to incorrect predictions and depends in the first place on the quality of the generated KG representations. A better mean rank indicates the ability of the system to deal with missing structural information in the KG. Accordingly, we conclude the superiority of our approach in creating more stable KG representations than the other compared approaches.

Tabel 1: Evaluation results

Method	**Mean Rank**		**Hits@10 (%)**	
	Raw	**Filter**	**Raw**	**Filter**
TransE	160	152	78.77	91.21
IKRL	28	21	80.9	**93.8**
Our	**19**	**12**	79.80	91.55

References

- Bordes, Antoine, et al. "Translating embeddings for modeling multi-relational data." *Advances in neural information processing systems*. 2787-2795. (2013).
- Xie, Ruobing, et al. "Image-embodied Knowledge Representation Learning." *IJCAI 2017*. 3140-3146 (2016).
- Rothe, Sascha et al. "Autoextend: Extending word embeddings to embeddings for synsets and lexemes." *Proceedings of ACL* (2015).
- Chatfield, Ken, et al. "Return of the devil in the details: Delving deep into convolutional nets." *arXiv preprint arXiv:1405.3531* (2014).
- Pennington, Jeffrey et al. "Glove: Global vectors for word representation." *Proceedings of EMNLP*. 1532-1543 (2014).
- Zhang Wang Z et al."Knowledge graph embedding by translating on hyperplanes". *Proceedings of AAAI.* 1112-1119. (2014)