Multimodal

A Multimodal Translation-Based Approach for Knowledge Graph Representation Learning

PART 1

Introduction & related work

- Introduction
- Related work

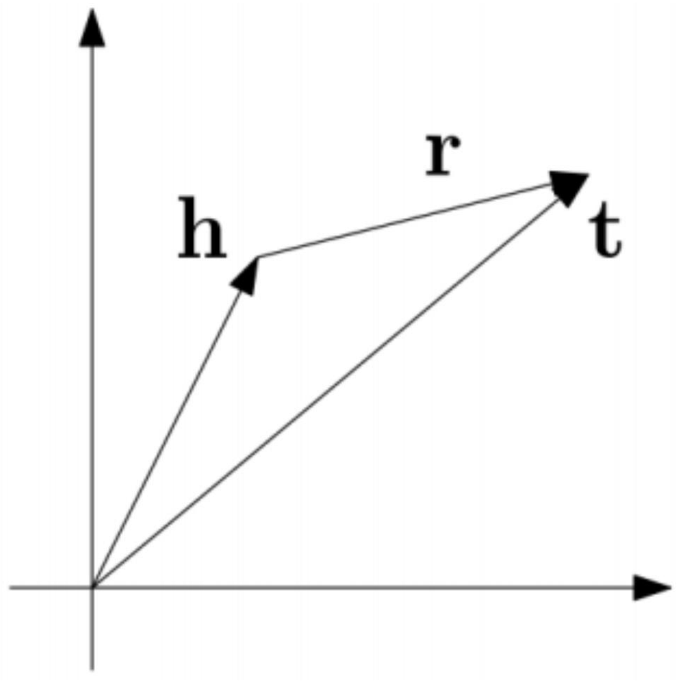
Introduction

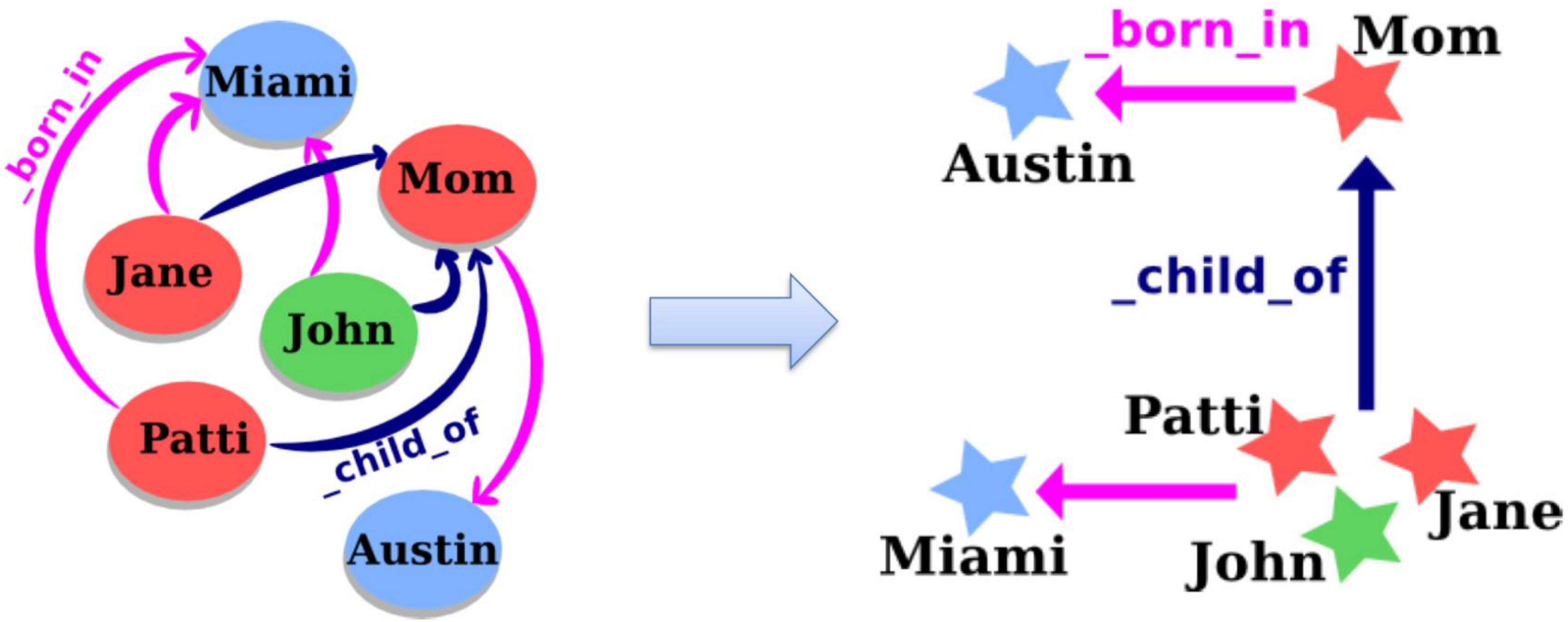
- Knowledge graph
  - Each node = an entity
  - Each edge = a relation
- Fact: (head, relation, tail)
- To solve: the incompleteness of the KGs
- Translation-based approaches:
  - model the entities and their relation as low-dimensional vector representations (embeddings)
- Problem:
  - rely on the rich structure of the KG
  - generally ignore any type of external information about the included entities

TransE

- For each triple (head, relation, tail), relation as a translation from head to tail
- Learning objective: h + r = t

$$f_r(h,t) = |\boldsymbol{l}_h + \boldsymbol{l}_r - \boldsymbol{l}_t|_{L_1/L_2}$$





Introduction

- Propose a model that leverages two different types of external, multimodal representations :
  - linguistic representations: created by analyzing the usage patterns of KG entities in text corpora
  - visual representations: obtained from images corresponding to the KG entities.
- Compare with TransE model:
  - Datasets:WN9-IMG dataset (2017)
  - TransE fails to create suitable representations for entities that appear frequently as the head/tail of one-to-many/many-to-one relations.

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

Contributions

- Propose an approach for KG representation learning that incorporates multimodal (visual and linguistic) information in a translation-based framework
- Investigate different methods for combining multimodal representations and evaluate their performance;
- Introduce a new large-scale dataset for multimodal KGC based on Freebase;
- The approach outperforms baseline approaches including the state-of-the-art method of Xie et al. (2017) on the link prediction and triple classification tasks.

Related work——Translation Models

- TransE (2013)
  - represents entities and relations as vectors in the same space
  - $h + r \approx t$
  - minimizing a margin-based ranking objective
- TransH (2014)
  - uses translations on relation-specific hyperplanes
  - applies advanced methods for sampling negative triples
- TransR (2015)
  - uses separate spaces for modeling entities and relations
- PTransE (2015)
  - leverages multi-step relation path information in the process of representation learning

Related work——Multimodal methods

- Shutova et al. (2016)
  - better metaphor identification can be achieved by fusing linguistic and visual representations
- Col-lell et al. (2017)
  - demonstrated the effectiveness of combining linguistic and visual embeddings in the context of word relatedness and similarity tasks
- IKRL (2017)
  - extends TransE based on visual representations extracted from images that correspond to the KG e-ntities
  - the energy of a triple is defined in terms of the structure of the KG as well as the visual representatio-n of the entities

PART 2

Algorithm

- Definition
- Model
- Objective function
- Combining Multimodal Representations

Definition

- Knowledge graph :
  - G = (E , R, T)
  - E is the set of entities, R is the set of relations, and T = {(h, r, t)| h, t ∈ E, r∈ R}
- three kinds of representations (head & tail)
  - Structural : $\boldsymbol{h}_{\boldsymbol{s}}^{\boldsymbol{I}}, \boldsymbol{t}_{\boldsymbol{s}}^{\boldsymbol{I}} \in \mathbb{R}^N$
  - linguistic : $\boldsymbol{h}_{\boldsymbol{w}}^{\boldsymbol{I}}, \boldsymbol{t}_{\boldsymbol{w}}^{\boldsymbol{I}} \in \mathbb{R}^M$
  - Visual : $\boldsymbol{h}_{\boldsymbol{i}}^{\boldsymbol{I}}, \boldsymbol{t}_{\boldsymbol{i}}^{\boldsymbol{I}} \in \mathbb{R}^P$
- Relation representation: $\boldsymbol{r}_{\boldsymbol{s}}^{\boldsymbol{I}} \in \mathbb{R}^N$
- Transform into common space : multi-layer model
- translational assumption $\boldsymbol{h}_{\boldsymbol{s}} + \boldsymbol{r}_{\boldsymbol{s}} \approx \boldsymbol{t}_{\boldsymbol{s}}.$

Model

- sample specific kinds of negative triples :

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

- energy function
  - Structural Energy : $E_S = \|\boldsymbol{h_s} + \boldsymbol{r_s} - \boldsymbol{t_s}\|.$
  - Multimodal Energies : $E_{M1} = \|\boldsymbol{h_m} + \boldsymbol{r_s} - \boldsymbol{t_m}\|.$ $E_{M2} = \|(\boldsymbol{h_m} + \boldsymbol{h_s}) + \boldsymbol{r_s} - (\boldsymbol{t_m} + \boldsymbol{t_s})\|.$
  - Structural-Multimodal Energies : $$E_{SM} = \|\boldsymbol{h_s} + \boldsymbol{r_s} - \boldsymbol{t_m}\|$$ $$E_{MS} = \|\boldsymbol{h_m} + \boldsymbol{r_s} - \boldsymbol{t_s}\|$$
- overall energy : $E(h, r, t) = E_S + E_{M1} + E_{M2} + E_{SM} + E_{MS}$
  - cannot be fulfilled at the same time, but combining these energies makes the results more robust.

Model

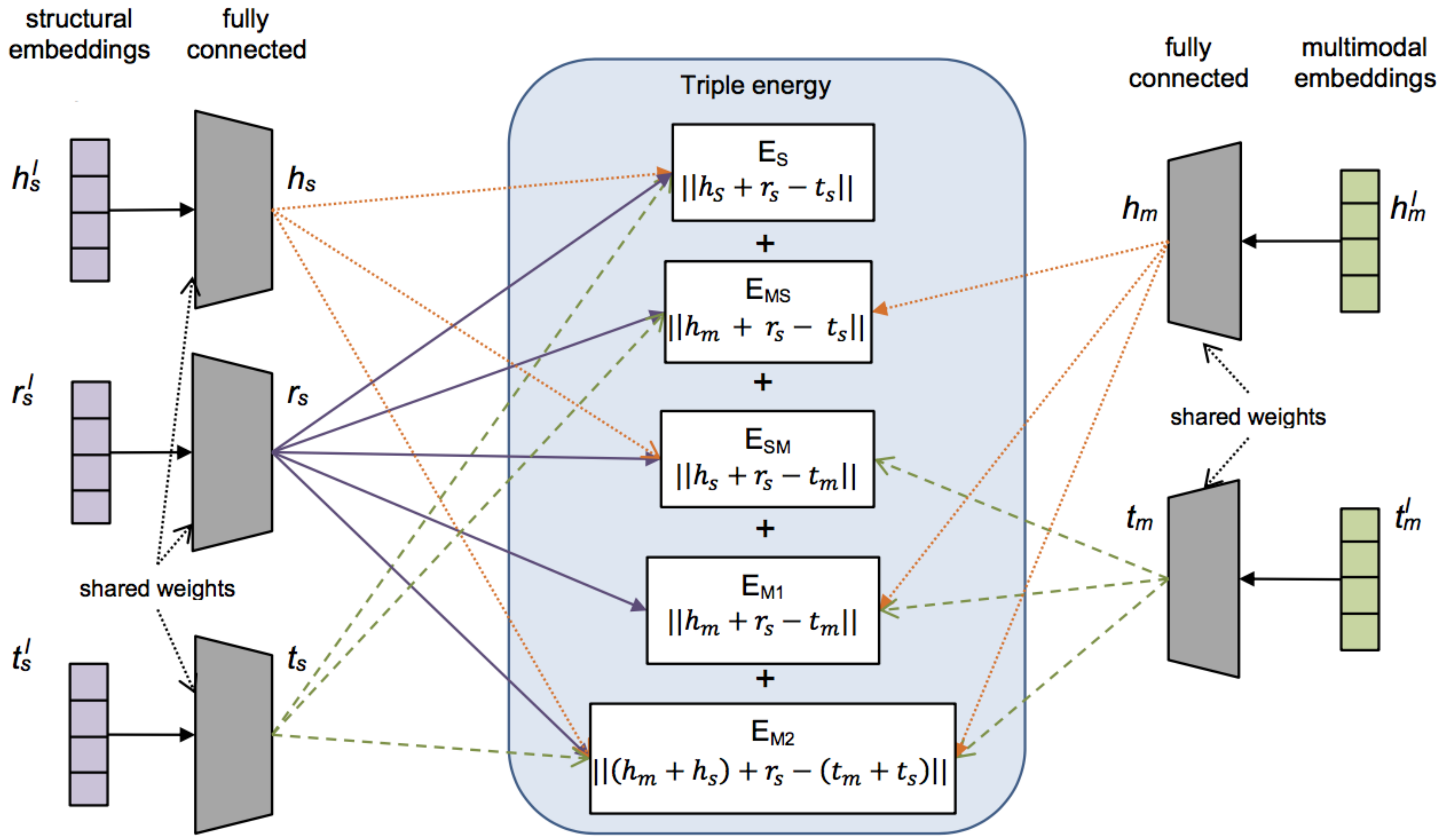


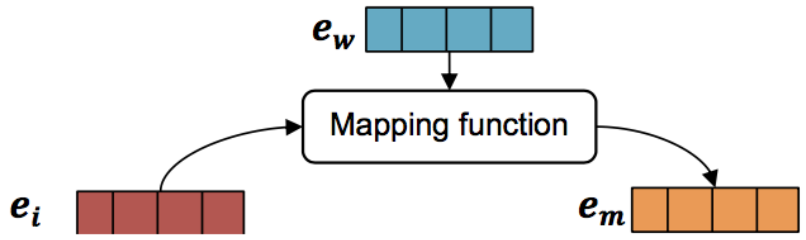
Figure 1: Overview of the neural network architecture for calculating the total triple energy from the different models. The fully connected networks transform the respective input embeddings into a common space.

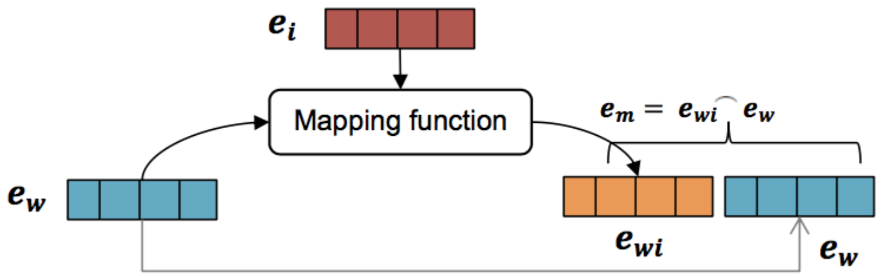
Objective function

- Head view: $$\mathcal{L}_{head} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h,r,t')\in\mathcal{T}'_{\text{tail}}} \max\big(\gamma + E(h,r,t) - E(h,r,t'), 0\big) \quad (10)$$
- Tail view: $$\mathcal{L}_{tail} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t)\in\mathcal{T}'_{\text{head}}} \max\big(\gamma + E(t,-r,h) - E(t,-r,h'), 0\big). \quad (11)$$
- global loss $$\mathcal{L} = \mathcal{L}_{\text{head}} + \mathcal{L}_{\text{tail}}.$$

Combining Multimodal Representations

- Concatenation Method : $e_m = e_w \frown e_i$
- DeViSE Method (2013): map images into a semantic embedding space
    - Given the visual representation, learn a mapping into the linguistic (word) embedding space





- Image Method (2017): reverse procedure
    - learn a mapping from the linguistic embedding space of that concept into the visual embedding space
    - **objective** : minimize the distance between the mapped linguistic representation and the visual representation of the entities.

PART 3

Experiment

- Datasets
- Representations
- Link prediction
- Triple Classification

Datasets

- WN9-IMG: provided by Xie et al. (2017) is based on WordNet
  - entities : word senses (synsets)
  - relations : lexical relationships between the entities
  - for each synset a collection of up to ten images obtained from ImageNet
- FB-IMG: created based on FB15K
  - For each entity, crawled 100 images from the web using text search based on the entity labels
  - feeding images into a pre-trained VGG19 neural network for image classification (4096)
  - calculated the PageRank score for each image in the graph and kept the top 10 results

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

Representations

- Structural Representation (both datasets)
  - train TransE with 100 dimensions
  - use the same values for the other hyperparameters
- Linguistic Representation
  - FB-IMG: word2vec (1000 dimensions)
  - WN9-IMG: AutoExtend (initialized AutoExtend with pretrained 300-dimensional GloVe embeddings)
- Visual Representation
  - FB-IMG: VGG-m-128 CNN model (128 dimensions)
  - WN9-IMG: pre-trained VGG model (4096 dimensions)

Link Prediction

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.

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.

Triple Classification

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.

END

THANK YOU