Towards Holistic Concept Representations: Embedding Relational Knowledge, Visual Attributes, and Distributional Word Semantics

Steffen Thoma(✉), Achim Rettinger, and Fabian Both

Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{steffen.thoma,rettinger}@kit.edu, fabian.both@student.kit.edu

Abstract. Knowledge Graphs (KGs) effectively capture explicit relational knowledge about individual entities. However, visual attributes of those entities, like their shape and color and pragmatic aspects concerning their usage in natural language are not covered. Recent approaches encode such knowledge by learning latent representations ('embeddings') separately: In computer vision, visual object features are learned from large image collections and in computational linguistics, word embeddings are extracted from huge text corpora which capture their distributional semantics. We investigate the potential of complementing the relational knowledge captured in KG embeddings with knowledge from text documents and images by learning a shared latent representation that integrates information across those modalities. Our empirical results show that a joined concept representation provides measurable benefits for (i) semantic similarity benchmarks, since it shows a higher correlation with the human notion of similarity than uni- or bi-modal representations, and (ii) entity-type prediction tasks, since it clearly outperforms plain KG embeddings. These findings encourage further research towards capturing types of knowledge that go beyond today's KGs.

Keywords: Knowledge fusion · Multimodality · Entity embeddings · Visual features · Distributional semantics · Entity-type prediction

1 Introduction

In recent years, several large, cross-domain, and openly available knowledge graphs (KGs) have been created. They offer an impressively large collection of cross-domain, general knowledge about the world, specifically instantiated relations between individual entities (statements). However, there is a lack of other types of information like visual object features or distributional semantics about the usage of those entities in the context of textual descriptions of real-world events.

Consider for instance the entity 'baseball' as depicted in Fig. 1: Images of baseballs provide basic visual information about the shape and color, something that is not present in KGs. While it is theoretically possible to make such information explicit with a graph-based formalism, it is not the obvious choice, since

C. d'Amato et al. (Eds.): ISWC 2017, Part I, LNCS 10587, pp. 694–710, 2017.
DOI: 10.1007/978-3-319-68288-4_41

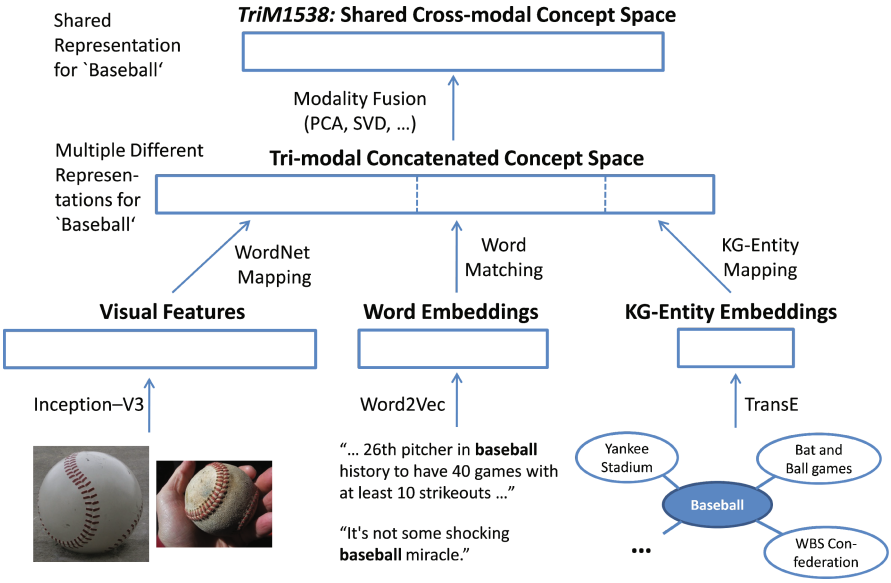


Fig. 1. Approach for extracting a shared cross-modal concept space from image, text and knowledge graph (aligned on word-level).

the detailed formal modelling of a shape or texture is far less efficient than capturing this with an unstructured representation like an image.

Similarly, text documents contain another type of essential information that is not available in KGs. Texts that mention 'baseball' typically comment or analyze baseball games and players. Since there is a huge number of examples on the actual usage of terms in text, this provides distributional context which is not available via the graph-neighborhood of the entity 'baseball' in a KG. KGs contain rather stable relations between individual entities, like attributes of baseball teams, their locations, equipment and abstract categorizations such as a 'Bat and Ball Game'.

It seems obvious that the three modalities (KGs, Text, Images) contribute different types of complementing information. Considering recent results in extraction of visual and textual content, that indicate an advantage of exploiting both modalities simultaneously to represent concepts [18], there also seems to be potential for tri-modal embeddings of textual context, visual information and relational knowledge of KG concepts.

This work investigates the influence of additional modalities on concept representations by means of a tri-modal embedding space that fuses information from text documents and image collections with knowledge graphs. When evaluating the resulting latent concept representation on standard similarity benchmarks, it indeed shows a higher correlation with the human notion of concept similarity than uni- (e.g., KG only) or bi-modal representations. Also, KG embeddings fused with embeddings trained on visual and textual documents clearly

outperform their uni-modal counterparts on KG completion tasks like entity-type prediction.

This convincingly demonstrates the great potential of joining latent knowledge representations constructed from multiple modalities, as detailed in the following sections. First, we discuss related work (Sect. 2), introduce existing uni-modal embeddings (Sect. 3), before explaining how they are aligned (Sect. 4) and fused (Sect. 5). We demonstrate its potential on similarity benchmarks (Sect. 6.1) and analyze its fusion effects (Sect. 6.2). In Sect. 6.3, we look into entity segmentation and assess entity-type prediction in Sect. 6.4, before we summarize our findings (Sect. 6.5) and conclude (Sect. 7).

2 Related Work on Fusion of Learned Representations

Recently, several researchers have tried to transfer learned knowledge from one task to another or to combine different approaches. In image classification, it is important that also new images can be classified so that visual representations from one image classification task can be transferred to another with different classes. To this end, Oquab et al. [27] learn and transfer mid-level image representations of CNNs. Kiela and Bottou [18] test the combination of visual and textual representations via vector stacking which is similar to [33] which uses a stacked auto-encoder to combine visual and textual input. In contrast to our approach they only evaluate simple vector stacking and neither evaluate more sophisticated combination techniques nor the incorporation of structured resources like KGs.

In contrast, Goikoetxea et al. [11] use textual information from a text corpus and WordNet. For this purpose, WordNet is transferred to text by performing random walks on the synset hierarchy and hereby storing the traversal path to text [12]. But, they neither use visual representations nor do they work with the information of an expressive KG directly. The transformation of a traversal path to text might lose characteristics of the underlying graph structure which is why we used latent vector representations from an explicit KG model, learned on a complete KG. Furthermore, they only combine vectors of equal size to circumvent the dimensionality bias while we introduce an appropriate normalization and weighting scheme.

Our approach also goes beyond current retrofitting ideas like [9]. They adjust learned word embeddings by incorporating information from lexical databases. Firstly, we do not slightly adapt one representation but learn a completely new combined representation. Secondly, we use much more information from a large expressive KG (DBpedia) instead of a smaller lexical database. Lastly, we also use visual information.

The closest work to our word-level alignment to concept space is [30][1]. They used autoencoders with rank 4 weight tensors to create vector representations

[1] Please note, that they did not consider any combinations with visual or KG embeddings.

for synsets and lexemes in WordNet for which there was no learned vector representation before. They achieve this by treating a word and a synset as the sum of its lexemes.

The closest work to our approach are [6,14]. Hill et al. [14] add explicit image tag information into the textual representation by adding the image tags into the training data. By placing the tags next to the words, they include the connection between word and its explicit visual features (tags). Then again, [6] concatenate and fuse latent 'visual words' and textual representations with singular value decomposition (SVD). Their results on bi-modal experiments indicate that multi-modal information is useful and can be harnessed. In addition to [6], we also consider relational knowledge from a KG, test further combination methods and evaluate on different tasks.

3 Uni-Modal Vector Representations

Latent vector representations of various types have become quite popular in recent years. The most common ones are latent *textual representations*, which are also referred to as *word embeddings*, *distributional word semantics* or *distributed word representations*. Created with unsupervised methods, they only rely on a huge text corpus as input. The information of co-occurrences with other words is encoded in a dense vector representation and by calculating the cosine similarity between two representations, a similarity score between two words is obtained. Examples for such *textual representations* are [2], SENNA [7], hierarchical log-bilinear models [24], word2vec [21–23], and GloVe [28]. Word embeddings are able to capture the *distributional knowledge* of how words are used across huge document collections.

Similarly, images can be encoded in a latent vector space. For *image representations*, deep convolutional neural networks (CNNs) have shown promising results in recent years. Deep CNNs transfer an image into a low dimensional vector space representation e.g. for image classification by applying a softmax function. The latent vector representation for images correspond to layers in the deep CNN before applying the softmax. For image classification with CNNs, Inception-V3 [34] which is used in TensorFlow [1] has shown good results on the ImageNet classification task [31]. Image embeddings are able to capture abstract *visual attributes* of objects, like their abstract shape.

The term 'Knowledge Graph' was revived by Google in 2012 and is since then used for any graph-based knowledge base, the most popular examples being DBpedia, Wikidata, and YAGO (see [8] for a survey). Similarly, *knowledge graph embeddings* can be learned on those graphs consisting of entities and typed predicates between entities and abstract concepts. These entities and predicates can be encoded in a low dimensional vector space, facilitating the computation of probabilities for relations within the knowledge graph which can be used for link prediction tasks [29]. Examples for learning latent vector representations of knowledge graphs are SE [5], RESCAL [26], LFM [17], TransE [4], SME [3], HolE [25], ComplEx [35], and the SUNS framework [16]. KG embeddings are

obtained by collective learning which is able to capture the *relational structure* of related entities in a KG.

4 Tri-Modal Concatenated Concept Space

The aim of this paper is to assess the potential of integrating *distributional*, *visual*, and *relational knowledge* into one representation. For obtaining such a consolidated tri-modal space, an embedding across all modalities is needed. Most existing bi-modal approaches rely on manually aligned document collections. Thus, an explicit reference (i.e., DBpedia URI) to the mentioned or depicted concept cannot be established, since a whole document is embedded and no individual concepts. This is not suitable for our investigations, since we want to assess how representations of single concepts can benefit from multi-modal embeddings. Instead, we build on pre-trained uni-modal representations (KG entities, words and visual objects) and align them across modalities.

We chose the most established approaches from their respective fields[2]: For *textual embeddings* we picked the word2vec model and Inception-V3 for *visual embeddings*. For *knowledge graph embeddings*, we trained representations using the TransE model [19]. To establish which embeddings represent the same concept in the different modalities we align them on a word-level:

Matching of Word Embeddings: We identified the intersection of word2vec embeddings that are represented by all modalities.

Concept Mapping of KG Embeddings: The latent vectors of TransE are representing concepts in the DBpedia graph. Each concept is uniquely addressable through a DBpedia URI and several labels (surface forms) are provided. We use the most commonly used label for referring to the concept.

WordNet Mapping of Visual Objects: For visual representations, we use the images from *ImageNet 1k* [31] which consists of 1000 categories. Each category has a set of at least 1300 images for the respective concept and is linked to synsets in WordNet. By combining all image representations for a given synset, we obtain a visual representation for the synset. Alike to [18] we combine the image representations by taking the max-value for each vector index as this yielded better results compared to mean values. Additionally, we build more abstract synset representations by utilizing the WordNet hierarchy, e.g. an embedding of *'instrument'* can be created by combining embeddings of *'violin'*, *'harp'*, etc. We build hierarchical subtrees in WordNet for each missing synset in *ImageNet 1k*. All synset representations in such a subtree with a visual representation from *ImageNet 1k* are then combined with a feature-wise max operator to form an abstract synset representation. In total, we abstract 396 additional synset representations.

[2] Please note, that any other embedding approach (see Sect. 3), could be plugged into our approach. We are not aiming to compete on uni-modal benchmarks but investigate the impact of additional modalities regardless of the original embedding approach.

The alignment of the synset representations to a shared set of concepts are performed with the WordNet lexemes which are assigned to at least one synset in WordNet. In the end, we extract 2574 lexeme representations by averaging the synset representations related to a given lexeme.

The intersection of Inception-V3 with word2vec and TransE embeddings leads to an aligned tri-modal concept space containing 1538 concepts. For each shared concept, the representations from all modalities are concatenated so that fusion techniques for the resulting concept space *TriM1538* can be applied next (see Fig. 1).

5 Shared Cross-Modal Concept Space

For fusing *distributional*, *visual*, and *relational knowledge* from the respective modalities, we used several methods which are described in the following paragraphs. Apart from simple concatenation we build on methods like SVD and PCA by proposing a *normalization (N)* and *weighting (W)* scheme for embeddings from multiple modalities. Our tri-modal concept space of 1538 different concepts is represented in three matrices: text T, knowledge graph G, and visual V. For combination techniques, we use the whole information of all three modalities and define matrix $M \in \mathbb{R}^{(t+g+v)\times 1538}$ as the vertically stacked matrices of T, G, and V. The dimensionality of these three matrices varies drastically: *Visual representations* tend to have more than 1000 dimensions while *knowledge graph representations* typically have around 50 to 100 dimensions. Thus, the representations with higher dimensionalities tend to dominate the combination techniques. Furthermore, the value range of features can differ depending on the underlying training objective and method. To address these problems we propose pre-processing steps, comprising *normalization (N)* of each column vector of T, G, and V to unit length as well as *weighting (W)* of the normalized matrices with weights w_T, w_G, and w_V before stacking. Thus, we can take into account that certain representations are more informative and condensed than others.

AVG: The averaging method uses the cosine similarity of all three modalities which are calculated separately. By averaging these three values, we get a combined similarity measure which is also robust with respect to different vector dimensionalities.

CONC: The similarity for the concatenated vectors of the single representations can be calculated with the cosine similarity. The similarities of the following techniques are also calculated with cosine similarity.

SVD: Singular value decomposition factorizes the input matrix M into three matrices such that $M = U\Sigma V^T$. U and V are unitary matrices and Σ is a diagonal matrix with the singular values of M in descending order on its diagonal. By taking the first k columns of U and the k biggest singular values of Σ, we get a new combined k-dimensional representation: $M \leftarrow M_k = U_k\Sigma_k$.

PCA: Principal Component Analysis uses an orthogonal transformation to convert the correlated variables into linearly uncorrelated variables. Fixing the number of uncorrelated principal components results in a projection into a lower dimensional vector space. By taking the principal components with the highest variance, we create a representation with the most distinctive features. We also tested canonical correlation analysis (CCA) but in our tests PCA always performed superior which is consistent with [11]. Thus, we omitted further attempts based on CCA.

AUTO: Autoencoders are neural networks for learning efficient encodings (representations). Autoencoders consist of an encode and a decode function for transforming an input vector to a lower dimensional encoding which can be decoded again. The neural network variables are learned by reducing the reconstruction error between the encoded and subsequently decoded columns of M compared to its original column.

6 Experiments

To investigate if our joint embedding approach is able to integrate *distributional*, *visual*, and *relational knowledge* from the respective modalities and ultimately if common tasks benefit from that, we conducted qualitative and quantitative empirical tests. In our assessments, we use pre-trained representations for text and images as well as trained knowledge graph representations. For the *textual representation* we use word2vec[3]. Its vectors have 300 dimensions and were trained on the Google News corpus containing about 100 billion words. For *visual representations*, the Inception-V3 model[4], pre-trained on the *ImageNet 1k* classification task, was applied to compute representations with 2048 dimensions. *Knowledge graph representations* were obtained with the TransE model [4] which we trained by running TransE on DBpedia. We trained TransE with a local closed word assumption for type constraints, rank = 50, gamma = 0.3, learningrate-embeddings = 0.2 and learningrate-parameters = 0.5 on the latest DBpedia dump (April 2016). We made all used embeddings available online[5].

6.1 Word Similarity

For evaluating whether a joint embedding captures the human notion of similarity better than uni-modal embeddings, we utilize various word similarity datasets. These datasets were created by several persons that rated the similarity of word pairs like 'cheetah - lion'. Since *TriM1538* does not cover all words in

[3] https://code.google.com/archive/p/word2vec/.
[4] http://download.tensorflow.org/models/image/imagenet/inception-v3-2016-03-01.tar.gz.
[5] https://people.aifb.kit.edu/sto/TriM1538.

Table 1. Spearman's rank correlation on subsets and complete datasets for word2vec.

	MEN	WS-353	SimLex-999	MTurk-771	weighted ∅
Complete data	0.762	0.700	0.442	0.671	0.682
Subset	0.740	0.694	0.441	0.608	0.672

the evaluation datasets[6], we evaluate on the covered subsets and provide them online (See Footnote 5). To ensure that the subsets used for evaluation are not easier to align we compared the word2vec performance to the full set. Table 1 shows the performance of word2vec on the respective datasets *MEN* [6], *WS-353* [10], *SimLex-999* [15], and *MTurk-771* [13]. We also report the average performance over all evaluation datasets, weighted by their respective size. Table 1 confirms that similarities in the subsets are equally hard to predict.

In Table 2, the Spearman's rank correlation on all subsets for raw stacking, *normalization (N)* and *weighting (W)* is reported. Normalized representations allow for a fixed combination ratio, resembling an equal weight of information from all modalities. We conducted experiments with different dimension parameters for SVD, PCA, and AUTO. Our results indicate that 100 dimensions are sufficient to encode the information for the word similarity task. In case of simple stacking (second block in Table 2), none of the combination methods is significantly better than the uni-modal text representation on the *MEN*, *MTurk-771*, and *WS-353* subset. Also, combination methods with *normalization (N)* are not significantly and consistently outperforming the textual representation.

To investigate if modalities are equally informative or provide complementary information we use *weighting (W)* of representations after normalization in order to quantify the impact of different proportions of information induced by each representation. With grid search and a step size of 0.05 we investigated the modality composition on the weighted average of all evaluation sets. The optimal weights for (w_T, w_G, w_V) are: AVG (0.15, 0.05, 0.8), CONC (0.25, 0.15, 0.6), SVD (0.3, 0.05, 0.65), and PCA (0.25, 0.05, 0.7)[7]. While some of the weighting schemes only include small proportions of the KG representations, the extracted complementary information from KGs still improves the performance in every approach significantly. In Fig. 2, you can see the weighted average of Spearman's rank correlation scores for different weightings between normalized *visual*, *textual*, and *KG representations*. It clearly shows that the combination of the three fused and weighted modalities produces better results than any single modality[8]. Weighted combination methods substantially outperform uni- and bi-modal embeddings while best results are obtained with SVD and PCA. Applying the dimension reduction methods SVD and PCA (100 dimensions) only on the initial

[6] Naturally, the limiting factors are verbs, abstract words, and named entities (e.g. persons) for which no visual representation is available.

[7] Due to high computational costs we omitted the autoencoder.

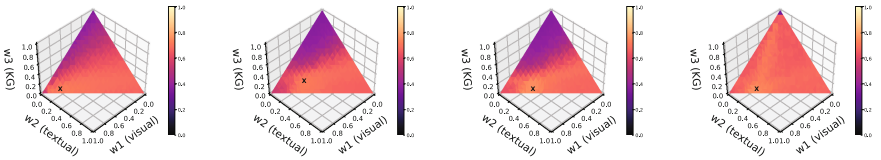
[8] Otherwise the optimum (depicted with a black cross) would be in a corner (uni-modal) or edge (bi-modal) of the triangle.

Table 2. Spearman's rank correlation on subsets of evaluation datasets.

	MEN	WS-353	SimLex-999	MTurk-771	Weighted ∅
Visual	0.619	0.526	0.522	0.308	0.546
Textual	0.740	0.707	0.423	0.594	0.669
KG	0.452	0.433	0.284	0.097	0.369
AVG	0.738	0.595	0.460	0.485	0.643
CONC	0.620	0.520	0.518	0.317	0.546
SVD	0.739	0.646	0.591	0.352	0.646
PCA	0.710	0.595	0.663	0.354	0.634
AUTO	0.456	0.672	0.485	0.294	0.456
AVG-N	0.738	0.595	0.460	0.485	0.643
CONC-N	0.738	0.595	0.460	0.485	0.643
SVD-N	0.724	0.555	0.422	0.440	0.618
PCA-N	0.769	0.601	0.452	0.558	0.673
AUTO-N	0.742	0.607	0.473	0.527	0.655
AVG-W	0.795	0.726	0.592	0.577	0.724
CONC-W	0.795	0.726	0.598	0.574	0.724
SVD-W	0.826	0.722	0.633	**0.667**	**0.762**
PCA-W	**0.831**	**0.758**	**0.688**	0.567	0.760

uni-modal embeddings did show improvements for the visual embeddings to an averaged Spearman's rank of 0.619 (SVD) and 0.639 (PCA) (weighted average). For comparison, the best reported result for uni-modal models on SimLex-999 is [32] with Spearman's rank correlation of 0.563. The bi-modal approach [6] reported Spearman's rank correlations of 0.78 on MEN and 0.75 on WS-353 while their model covered 252 word pairs of WS-353. Please note, our results on subsets of *MEN*, *WS-353*, *SimLex-999*, and *MTurk-771* are competitive but not directly comparable to the numbers reported by state-of-the-art uni-modal approaches as they are evaluated on the complete datasets and ours cannot. However, since this paper is about relative performance gains through additional modalities we do not compete with, but are complementary to the state-of-the-art uni- and bi-modal approaches.

For combinations via AVG and CONC as shown in Fig. 2a and b, we observe similar behavior on all evaluation datasets in terms of optimal weights. SVD and PCA exploit information from KG representations with very low weight, but the combined representation of all three modalities is significantly better than a combination of only two modalities. The best bi-modal combinations were AVG (0.15, 0, 0.85) with 0.709, CONC (0.3, 0, 0.7) with 0.709, PCA (0.3, 0, 0.7) with 0.749, and SVD (0.3, 0, 0.7) with 0.759 Spearman's rank correlation (weighted average).



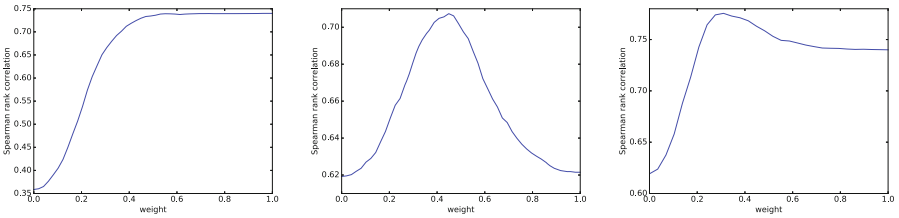
(a) AVG (b) CONC (c) SVD 100 dim (d) PCA 100 dim

Fig. 2. Averaged plots over all evaluation datasets for weighting with normalization. The colorbar indicates Spearman's rank correlation and the black cross marks the optimum.

The key finding is, that optimal weights always include all three modalities, so indeed make use of *visual*, *distributional*, and *relational knowledge*. Further experiments with different TransE model parameterizations revealed that this finding is not depending on a specifically trained TransE embedding, but can be attributed to information extracted from the knowledge graph. Thus, we can improve concept representations from other modalities with complementary information encoded in Inception-V3, word2vec, and TransE embeddings.

6.2 Noise Induced Errors Vs. Complementary Information Gain

In a further step, we investigated the fusion effects in more detail. Every meaningful representation encodes useful information which is defined by the model's learning objective. Before combining models for a certain task, one has to verify that the model encodes information for that specific task. Also, the representation quality for a certain task might vary greatly. While complementary information of various models and modalities can lead to an improvement when combined, a weak model for the specific task might induce noise. Adding a model to a combined representation is only beneficial if the gain through complementary information is greater than the information loss induced by noise.



(a) No complementary information between noisy word2vec (w=0) and noise-free word2vec (w=1).

(b) Complementary information between visual (w=0) and noisy word2vec (w=1) representations.

(c) Superposition of both effects, visual (w=0) and noise-free word2vec (w=1).

Fig. 3. Concatenation effects

To illustrate these two effects, we evaluate representations with noisy models on the *MEN* dataset. To isolate effects of *noise induced errors*, we combine two textual representations after normalization and compute Spearman's rank correlations. Pre-trained word2vec representations served as the first high quality model. A second textual representation was generated by artificially adding noise to the word2vec model. For that reason, we added 100 dimensions with uniformly distributed random values and tuned representation quality by scaling the distribution interval. Following this procedure, we can observe the fusion effects between two concatenated representations with no complementary information in Fig. 3a.

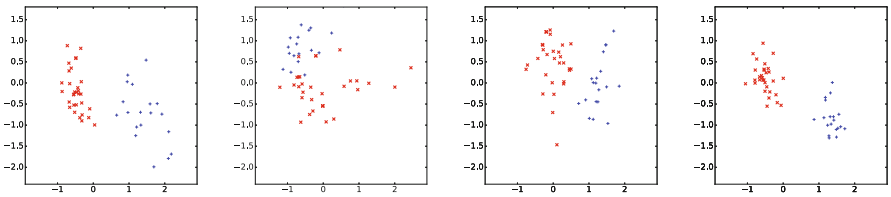
For showing the information gain of *complementary information* in Fig. 3b, we combine Inception-V3 representations ($w = 0$) with another noisy word2vec version ($w = 1$). Following the procedure above, we added noise to the word2vec representations and scaled the distribution interval until performance was similar to Inception-V3. One can observe the performance peak close to a weighting ratio of 1:1 between visual and textual representations which indicates that the visual and textual embeddings indeed hold complementary information.

In Fig. 3c one can observe the superposition of both effects during concatenation. While the visual model performs worse than the textual model on its own, the information gain through complementary information is larger than the information loss due to noise. Understanding the exact position of the maximum requires further research. Overall, combining two representations via concatenation improves results, if the performance gap between both models is not too large and both models encode complementary information (which is the case in our experiments).

6.3 Entity Segmentation

Besides showing that a joint concept embedding comes closer to the human notion of similarity, we can also demonstrate improvements in semantic entity segmentation. In Fig. 4, we exemplarily show that the *TriM1538* space is better suited for segmenting entities when compared to the *textual*, *visual*, and *KG embedding space*. Entities represented with *red crosses* are *land vehicles* and various *birds* are plotted with *blue plus symbols*. We computed the first two principal components of all three modalities and of *TriM1538*. For *TriM1538* we used normalization with weighted concatenation and the respective weights are taken from our previous experiments on the evaluation datasets: (w_T, w_G, w_V) = (0.25, 0.15, 0.6). In order to compare the first two principal components of different embeddings, we normalize the PCA-vectors for each embedding to unit length. This is important since we are interested in a relative separation while the variance explained by the first two components might vary greatly between different representations.

All three single representations show the ability to separate the DBpedia categories *birds* from *land vehicles*. In the textual domain, clustering of *land vehicles* is clearly observable and *birds* are separated but do not show an equally



(a) Textual (b) Visual (c) KG (d) CONC

Fig. 4. Segmentation results for birds (blue '+') and land vehicles (red '×'). (Color figure online)

condensed cluster. Visually, *birds* are clustered relatively close together, but vehicles are mixed into the cluster.

Similarly to text, the KG separates *birds* and *land vehicles* almost perfectly, but does not create clean clusters. When combined in *TriM1538*, clustering and separation is better than in all others modalities. Apparently, exploiting *distributional*, *visual* and *relational knowledge* results in a clearer semantic entity segmentation.

6.4 Entity-Type Prediction

Finally, we show that established KG tasks can also benefit from embedding *distributional semantic* and *visual attributes* into *relational knowledge*. Entity-type prediction is such a common KG completion task, similar to link prediction [20].

In order to test our *TriM1538* embeddings in the context of entity-type prediction, we use the following experimental setup: for a given KG entity $e \in E$ and the set of available categories C, we predict to which of the categories $c \in C$ the entity belongs (e.g. http://dbpedia.org/page/Category:Mammals). We define the subgraph of DBpedia that contains entities covered by *TriM1538* and their relations as the *TriM-KG* and denote the complete set of KG entities as $E^* = C \cup E$. Overall, *TriM-KG* contains 3220 triples and 1955 entities of which 634 are categories and 1321 entities with multi-modal information. Embeddings trained on *TriM-KG* are named *locally* trained embeddings, while embeddings trained on the whole KG are referred to as *globally* trained. In the following, we refer to entity, predicate and category vector embeddings with e, p and c.

We utilize the standard link prediction procedure of TransE as a baseline: For an entity e of interest, we train TransE on *TriM-KG* and exclude all triples connecting that entity to its category in C. The training parameters are the same as for the *globally* trained TransE except for a reduced rank in order to circumvent overfitting. The translation operation of TransE is then defined as the vector operation:

$$sim(e, p, c) = \|c - (e + p)\|_2 \tag{1}$$

Similar to [19], we compute $sim(e, p, c)$ for all possible $c \in C$ and get the rank of the true triples by ignoring all other true triples to prevent distortion (since

an entity might be correctly related to multiple categories). As the similarity measure we use the L2-norm within TransE and report mean ranks as well as the ratio of hits in the top 10 (hits@10). As an additional benchmark, we compare the *locally* trained TransE embeddings with the *global* TransE embeddings, for which the entity-category relations (which have to be predicted) were present during training. Finally, we report results for *locally* trained RESCAL embeddings with the same setup as for *local* TransE training (for details see [19]).

Category memberships of the multi-modal entities can also be directly computed with multi-modal embeddings of *TriM1538*. For this, we construct category embeddings from entity embeddings related to that category: For a given category, we compute its embedding with $\frac{1}{N} \cdot \sum_{i=1}^{N} e_i$ for all N multi-modal embeddings e_i related to category c. Please note, for predicting category memberships of an entity, that specific entity is not considered as being related to any category during the category construction process. Thus, we obtain different category embeddings for each related entity. In *TriM-KG*, all considered categories have connections to at least two different multi-modal entities to ensure the construction of the category embedding. We name this procedure *hierarchic construction (HC)* and use $d = \|e - c\|_2$ as the similarity measure.

Finally, we combine the entity-type prediction schemes from above. Since TransE performs superior to RESCAL (see Table 3), we introduce an enrichment procedure for TransE, which could similarly be adapted to RESCAL. We concatenate *locally* trained TransE representations e_{loc} with *TriM1538* entities e_{tri} after normalizing the respective embeddings to unit length. Similarly, we concatenate TransE category representations with embeddings obtained by *HC*. With these extended embeddings $e_{ext} = (e_{loc}, e_{tri})$, $c_{ext} = (c_{loc}, c_{tri})$ we reformulate Eqs. 1 to 2:

$$sim(e_{ext}, p, c_{ext}) = \|(c_{loc}, c_{tri}) - (e_{loc} + p, e_{tri})\|_2 \tag{2}$$

For the fusion techniques, the modality weights have to be optimized. To this end, we create training and test sets with a 0.5:0.5 split of our data and optimize on the training set. This resulted in (w_T, w_G, w_V): PCA (0.2, 0.4, 0.4), SVD (0.45, 0.55, 0), and CONC (0.4, 0.6, 0) for *Trans* E_{loc}+ *HC* and PCA (0.2, 0.55, 0.25), SVD (0.4, 0.6, 0), and CONC (0.4, 0.6, 0) for *HC*. As we have discussed in Sect. 6.2, weighting is task and model dependent which implies that the usefulness of the different types of knowledge from the respective modalities varies across different tasks. Further, the performance of a model, which is enriched with multi-modal information, greatly impacts the optimal modality composition. Thus, adapting the modality composition for new tasks is necessary.

Results for all methods are shown in Table 3. Consistent with observations in [19], the TransE-based baseline performs better than RESCAL. Interestingly, the *globally* trained TransE embeddings perform worse than the *locally* trained TransE, although the links to be predicted were present during its training and it has more information available. However, this is not surprising when comparing the size of the concept space of DBpedia ($7 \cdot 10^6$ concepts) with *TriM-KG* (1955 concepts).

Table 3. Results for type predictions with multi-modal embeddings on the right side. Results for TransE$_{loc}$ enriched with multi-modal embeddings on the left side. *TransE*, *RESCAL*, and *Random* at the bottom are baseline predictors without any multi-modal information or enhanced construction scheme.

	TransE$_{loc}$ + *HC*				Hierarchic Construction			
	Train		Test		Train		Test	
	Mean rank	hits@10	mean rank	hits@10	Mean rank	hits@10	Mean rank	hits@10
PCA	**10.401**	**0.828**	**10.251**	**0.824**	**12.274**	**0.863**	**14.680**	**0.869**
SVD	14.310	0.749	14.637	0.716	17.424	0.762	19.420	0.762
CONC	14.086	0.765	13.696	0.742	17.254	0.807	19.595	0.806
Word	14.297	0.763	14.215	0.741	24.157	0.784	28.107	0.764
Visual	32.982	0.475	33.477	0.462	96.805	0.581	96.763	0.575
KG	15.609	0.744	14.009	0.730	33.129	0.671	30.732	0.699
	Baselines							
TransE$_{loc}$	35.641	0.442	36.408	0.422				
TransE$_{glob}$	58.493	0.382	57.075	0.392				
RESCAL$_{loc}$	116.640	0.286	115.275	0.261				
Random	317.000	0.016	317.000	0.016				

The *HC* method even yields good results for entity type-predictions with uni-modal embeddings as shown in Table 3. Visual attributes alone are obviously not suited for predictions of type relations within the KG. Consistent with our observations in the word similarity task, embeddings from different modalities incorporate complementary information which can be exploited. With our modality fusion techniques, we achieve substantially superior results compared to uni-modal embeddings. Further, PCA is the best suited method for incorporating the sparse and rather noisy visual information in this setup and shows a significant performance boost compared to CONC and SVD.

Combining TransE$_{loc}$ with *HC* improves the mean rank even further. Utilizing uni- and multi-modal information enhances the predictions while PCA dominates all other methods. Compared to the standard TransE predictions, we improve the mean rank by 255% with multi-modal enrichment via *HC*.

6.5 Key Findings

- All our empirical evidence suggests that each modality encodes complementing information that is conceptually different: text provides *distributional*, images *visual* and KGs *relational knowledge*. Information encoded in the structure of embeddings can be useful for vastly different tasks and training objectives, even in other domains, as long as concepts can be aligned.
- Complementing information can be embedded in a joint representation which is closer to the human notion of similarity (see Sect. 6.1), as well as the human intuition in entity segmentation tasks (see Sect. 6.3).
- When enriching KG embeddings with *distributional* and *visual* knowledge from text and images, the performance of entity-type predictions is considerably improved (see Sect. 6.4). This indicates that those types of knowledge

are missing in today's KGs and KGs would greatly benefit if this could be integrated.
– The weighting of the influence for each modality before joining them across modalities is crucial and task dependent since the type of knowledge needed for each task varies. For improved performance, the positive effects created by the complementarity of information has to outweigh negative effects induced by noise in the original embeddings (see Sect. 6.2).

7 Conclusion and Future Work

The intention of this research was to find out if essential types of information, like *distributional* and *visual knowledge*, are not sufficiently represented in today's KGs (here DBpedia). This was investigated by embedding knowledge from text corpora, image collection and KG entities into a joint concept space. Comparing the performance of the joint cross-modal representation to uni-modal representations on various benchmark tasks allowed a quantitative and qualitative assessment. Our proposed two-step approach starts with pre-trained uni-modal concept representations created with established embedding methods from computer vision, natural language processing and semantic technologies. Next, the obtained concept embeddings were aligned across the three modalities, normalization and weighting schemes were devised, before the embeddings were fused into one shared space. Our novel cross-modal concept representation was evaluated in four sets of experiments by comparing it to uni-modal representations.

The main finding of this work is that the fused tri-modal embeddings reliably outperform uni- and bi-modal embeddings. This indicates that complementing information is available in the three investigated content representations and that the types of knowledge represented in text and images is conceptually different (*distributional* and *visual*) to the knowledge represented in KGs (*relational*). On the one hand, the performance gains were observed in tasks that optimize for the human notion of semantic similarity. It appears that the more modalities are considered the closer the knowledge representations come to a human-like perception. On the other hand, we investigated type-prediction in KGs and outperformed existing uni-modal methods by 255%. Again, the shared concept representation performed best when information from all three modalities was included.

Our findings raise fundamental questions and open up a large number of future research directions. First and foremost, it became obvious that knowledge graphs, and likely any knowledge representation that aims to provide a holistic view on entities and concepts, would benefit from integrating distributional and visual knowledge. Fusing embeddings from multiple modalities is an initial step to achieve that. Our approach is currently limited to the concept intersection of all modalities. While we do not need aligned training data, the obtained multi-modal concept space is relatively small. The most pressing issue for future work is to find ways to scale to a larger number of entities e.g. by including visual representations of tagged images, and to include relations.

Investigating approaches which harness multi-modal information for concepts outside of this intersection is also part of our future research. Beyond knowledge representation and representation learning research, findings in this area would impact numerous cross-disciplinary fields like sensory neuroscience, philosophy of perception, and multimodality research.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv preprint arXiv:1603.04467 (2016)
2. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
3. Bordes, A., Glorot, X., Weston, J., Bengio, Y.: A semantic matching energy function for learning with multi-relational data. Mach. Learn. **94**(2), 233–259 (2014)
4. Bordes, A., Usunier, N., García-Durán, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: NIPS 26, pp. 2787–2795 (2013)
5. Bordes, A., Weston, J., Collobert, R., Bengio, Y.: Learning structured embeddings of knowledge bases. In: AAAI 2011, pp. 301–306 (2011)
6. Bruni, E., Tran, N., Baroni, M.: Multimodal distributional semantics. J. Artif. Intell. Res. (JAIR) **49**, 1–47 (2014)
7. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: ICML 2008, pp. 160–167 (2008)
8. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked Data Quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semant. Web J. (2017, to be published)
9. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E.H., Smith, N.A.: Retrofitting word vectors to semantic lexicons. In: NAACL HLT 2015, pp. 1606–1615 (2015)
10. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. In: WWW 2001, pp. 406–414 (2001)
11. Goikoetxea, J., Agirre, E., Soroa, A.: Single or multiple? combining word representations independently learned from text and WordNet. In: AAAI 2016, pp. 2608–2614 (2016)
12. Goikoetxea, J., Soroa, A., Agirre, E.: Random walks and neural network language models on knowledge bases. In: NAACL HLT 2015, pp. 1434–1439 (2015)
13. Halawi, G., Dror, G., Gabrilovich, E., Koren, Y.: Large-scale learning of word relatedness with Constraints. In: ACM SIGKDD 2012, pp. 1406–1414 (2012)
14. Hill, F., Korhonen, A.: Learning abstract concept embeddings from multi-modal data: since you probably can't see what i mean. In: EMNLP 2014, pp. 255–265 (2014)
15. Hill, F., Reichart, R., Korhonen, A.: SimLex-999: evaluating semantic models with (genuine) similarity estimation. Comput. Linguist. **41**(4), 665–695 (2015)
16. Huang, Y., Tresp, V., Nickel, M., Rettinger, A., Kriegel, H.: A scalable approach for statistical learning in semantic graphs. Semant. Web **5**(1), 5–22 (2014)
17. Jenatton, R., Roux, N.L., Bordes, A., Obozinski, G.: A latent factor model for highly multi-relational data. In: NIPS 25, pp. 3176–3184 (2012)

18. Kiela, D., Bottou, L.: Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In: EMNLP 2014, pp. 36–45 (2014)
19. Krompaß, D., Baier, S., Tresp, V.: Type-constrained representation learning in knowledge graphs. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9366, pp. 640–655. Springer, Cham (2015). doi:10.1007/978-3-319-25007-6_37
20. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: AAAI 2015, pp. 2181–2187 (2015)
21. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 (2013)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS 26, pp. 3111–3119 (2013)
23. Mikolov, T., Yih, W., Zweig, G.: Linguistic regularities in continuous space word representations. In: NAACL HLT 2013, pp. 746–751 (2013)
24. Mnih, A., Hinton, G.E.: A scalable hierarchical distributed language model. In: NIPS 21, pp. 1081–1088 (2008)
25. Nickel, M., Rosasco, L., Poggio, T.A.: Holographic embeddings of knowledge graphs. In: AAAI 2016, pp. 1955–1961 (2016)
26. Nickel, M., Tresp, V., Kriegel, H.: A three-way model for collective learning on multi-relational data. In: ICML 2011, pp. 809–816 (2011)
27. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR 2014, pp. 1717–1724 (2014)
28. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP 2014, pp. 1532–1543 (2014)
29. Rettinger, A., Lösch, U., Tresp, V., d'Amato, C., Fanizzi, N.: Mining the semantic web - statistical learning for next generation knowledge bases. Data Min. Knowl. Discov. **24**(3), 613–662 (2012)
30. Rothe, S., Schütze, H.: AutoExtend: extending word embeddings to embeddings for synsets and lexemes. In: ACL 2015, pp. 1793–1803 (2015)
31. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
32. Schwartz, R., Reichart, R., Rappoport, A.: Symmetric pattern based word embeddings for improved word similarity prediction. In: CoNLL 2015, pp. 258–267 (2015)
33. Silberer, C., Lapata, M.: Learning grounded meaning representations with autoencoders. In: ACL 2014, pp. 721–732 (2014)
34. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR 2016, pp. 2818–2826 (2016)
35. Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., Bouchard, G.: Complex embeddings for simple link prediction. In: ICML 2016, vol. 48, pp. 2071–2080 (2016)