Extending Knowledge Bases Using Images

Vincent P. A. Lonij, Ambrish Rawat, Maria-Irina Nicolae
IBM Research – Ireland
Mulhuddart, Dublin 15, Ireland
{vincentl, ambrish.rawat}@ie.ibm.com, maria-irina.nicolae@ibm.com

1 Introduction

Traditional knowledge base generation extracts structured knowledge from text. However, many other sources of information, such as images, audio, or time series, can be leveraged into this process to build richer and more complete knowledge bases.

In this paper, we show how images can be converted to a structured semantic representation of their content. We achieve this by designing a mutual embedding space for images and knowledge graphs. From this mutual embedding space we can then obtain relationships between an image and known entities in a knowledge graph. Based on these obtained relationships, we show how labeled images can be used to add new concepts to the knowledge base. Furthermore, this mutual embedding helps visual recognition systems generalize better to unknown object classes.

One key finding of this paper is that the smoothness constraint we impose on the semantic embedding space has significant impact in multimodal tasks, while leaving the performance unchanged on unimodal tasks.

Vision Tasks involving multiple modalities like texts and images often use joint-semantic embeddings. However, our finding that these semantic spaces can be optimized for specific tasks suggests that a single semantic space cannot serve as the nexus between all modalities simultaneously. We hypothesize that a knowledge graph provides for a more universal semantic representation.

We envision an architecture where a knowledge graph forms the central component enabling perception through multiple modes (Figure 1). The knowledge base becomes the single representation of knowledge about the environment of the system. Each mode of perception would require bridging the gap between only that mode and the knowledge graph. The work presented in this paper fills the role of knowledge generation through images. Because the unified representation takes the form of a graph, it can then be used in reasoning systems to decide on an action.

Furthermore, this approach is a step towards bridging the gap between statistical frameworks that learn from large amounts of data and reasoning-based frameworks that can infer new knowledge from a limited set of predicates.

Our finding that knowledge graphs can improve visual recognition (by enabling recognition of unknown classes), and that visual recognition can be used to extend knowledge graphs points to the potential of building a continuous learning system based on these two components. This is a step towards building more generally applicable human-like learning systems.

2 Related Work

The method proposed in this paper is situated at the crossroads of multiple domains, including semantic embedding, multi-relational knowledge bases, link prediction and zero-shot learning. In this section, we position our work with respect to the most relevant approaches from these fields.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

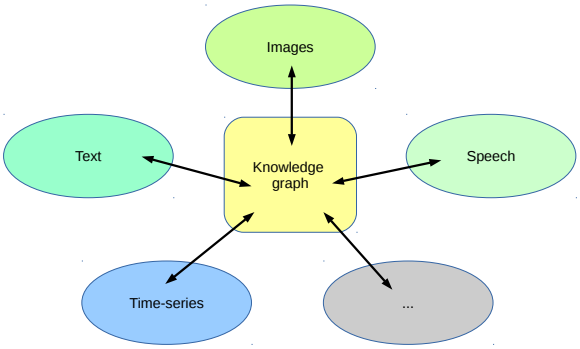

Figure 1: A knowledge base as the central component of a system capable of knowledge acquisition through multiple modes. Each arrow represents the capability to extract knowledge from a particular type of data. In this work, we demonstrate extraction of knowledge from images.

Semantic image embeddings A significant amount of work has been done to embed images into semantic vector spaces. Deep visual-semantic embedding (DeViSE) (Frome et al., 2013) leverages images as well as unannotated text to provide a common representation. Our setup closely relates to Frome et al. (2013), as well as Socher et al. (2013b) where they make class label predictions about unknown classes. In our setup, we focus on the prediction of semantic relationships, as opposed to class labels for new images. To our knowledge, visual recognition using such a structured semantic space has not explicitly been explored.

In Xie et al. (2016), knowledge representations are learned with both triples and images from a joint objective. However, they only consider knowledge generation for known entities, and do not add additional nodes to the knowledge base. A parallel thread for predicting zero-shot classes is to use their attribute signatures (Lampert et al., 2009; Shi et al., 2014; Huang et al., 2015). Although this has proven to be a successful approach, it requires explicit attribute labeling of the images and is therefore not suitable for knowledge base extension.

Knowledge graph embeddings Finding practical representations for knowledge graphs has been the focus of an important body of work. In the standard setting, the algorithm is presented with triples encoding existing edges in the graph. Each triple (e_h, r, e_t) links the head entity e_h to the tail entity e_t through the relation r which holds true between them. A major direction of research for knowledge representation are translation models which embed information into a low-dimensional vector space. TransE (Bordes et al., 2013) is one of the simplest, yet most effective, formulations of this type.

The problem of knowledge graph embedding has also been tackled via tensor factorization for link prediction (Nickel et al., 2011, 2012), as well as latent factors models (Sutskever et al., 2009; Jenatton et al., 2012) and semantic energy matching (Bordes et al., 2012, 2014). In this work, we build on the neural tensor layer (NTL) (Socher et al., 2013a) which considers multiplicative mixing of entity vectors in addition to linear mixing. We improve on this model by proposing a new objective function that enforces local smoothness on the scoring function.

Visual knowledge creation The paradigm of never ending learning of Carlson et al. (2010) has been applied to the task of knowledge extraction from images (Chen et al., 2013). This was achieved by modeling correlation between occurrences of detected objects in images. Vedantam et al. (2015) studied how common sense knowledge could be extracted from generated clip-art. Our work generalizes this and treats images and knowledge on equal footing. This enables us to extract knowledge directly from images without explicitly modeling correlations. Furthermore the knowledge we are able to extract contains relationships between images and concepts without images.

3 Knowledge Graph Embedding for Visual Recognition

Our purpose is to find a semantic vector space in which entities from a knowledge graph and images can have a unified representation. We arrive at this common space using a two-step approach. First, a representation for the entities is learned using a knowledge-graph. Subsequently, a mapping from

images to vectors in that same space is learned. This combination enables link-prediction for images (Figure 2).

To formalize these concepts we use the following notation. We will refer to the set of real numbers as $\mathbb{R}$. Scalars are denoted in lower case (n), vectors in bold ($\mathbf{v}$), while arbitrary sets in upper case symbols (M). We define the hinge loss as $[t]_+ = \max(0, t)$, while for the L_2 norm we use the notation $||\cdot||_2$. We use the set subtraction $A \setminus B$ to indicate the elements of A that are not in B.

Consider the finite set of all knowledge, represented as a knowledge graph $\mathcal{G}$. $\mathcal{G}$ can be modeled as a set of triples of the form $\{(e_i, r, e_j) \in T, e_i, e_j \in E, r \in R\}$, where E is the set of entities in the vocabulary, and R is the set of types of relations between entities. All triples in $\mathcal{G}$ are taken to be true, while all triples not in $\mathcal{G}$ are false. Having observed a subset $T' \subset T$ of triples containing a subset of entities $E' \subset E$, the task of link prediction is finding the triples in $T \setminus T'$. A standard version of this task addresses the setting where both T' and $T \setminus T'$ are based on same entities E' and relationships R. However, we will show that models can also be brought to predict triples for unknown entities $E \setminus E'$. To do this, we use information from a set of labeled images, $D := \{(x, e), x \in I, e \in E'\}$ where I is the set of all images. As mentioned above, this is done by learning an entity embedding $g : E \to V$ and an image embedding $h : I \to V$, where V is a vector space in $\mathbb{R}^d$. We now describe how to determine g and h.

Knowledge graph embedding To train the knowledge embedding model, we have at our disposal a set of triples T'. Let g be the function mapping the entities to a vector space $g : E \to V$. Now consider a scoring function f defined over triples which attributes true triples low scores, and false triples high ones; for now, we assume it of general form $f : V \times R \times V \to \mathbb{R}$. For a pair of entities e_h and e_t related with a relationship r, the score of the triple (e_h, r, e_t) can be computed in different ways. We adopt the bilinear form used in the neural tensor layer (NTL) architecture (Socher et al., 2013a):

$$f(g(e_h), r, g(e_t)) = \tanh\left(\left(g(e_h)^T W_r^{[1:k]}\right)^T g(e_t) + V_r \left[g(e_h) g(e_t)\right]^T + b_r\right), \quad (1)$$

where $W_r^{[1:k]} \in \mathbb{R}^{d \times d \times k}$, $V_r \in \mathbb{R}^{k \times 2d}$ and $b_r \in \mathbb{R}^k$ are embedding parameters, k is the number of slices as defined by Socher et al. (2013a), and $[g(e_h)g(e_t)]$ is the concatenation of the two vectors. Attributing scores to triples allows us to use hinge rank loss $\mathcal{L}_E(g, f)$ for learning the embedding as described in (Socher et al., 2013a). This margin-based loss function can then be written for the sample T' as:

$$\mathcal{L}_E(g, f) = \frac{1}{|T'|} \sum_{(e_h, r, e_t) \in T', (e_h, r, e_t') \notin T'} \left[\gamma + f(g(e_h), r, g(e_t)) - f(g(e_h), r, g(e_t'))\right]_+, \quad (2)$$

where γ is a margin parameter. (e_h, r, e_t') represents a false triple obtained by randomly replacing tail entity in (e_h, r, e_t) with a different one.

An essential property of a semantic space in the context of visual recognition is that the scoring function f should vary smoothly. This is particularly relevant in the neighborhood of entity points. We therefore modify $\mathcal{L}_E(g, f)$ by adding a noise term to the objective function:

$$\mathcal{L}_S(g, f) = \alpha \mathcal{L}_E(g, f) + (1 - \alpha)\mathcal{L}_E(\hat{g}, f), \quad (3)$$

where $\alpha > 0$ is a trade-off parameter. We use $\hat{g}$ to denote the operation of adding a normally distributed perturbation to the entity embeddings: $\hat{g}(e) = g(e) + \mathcal{N}(\mathbf{0}, s\mathbf{I})$. We will see that this is of particular importance when we use the embedding space for multi-modal transfer since the embedding vectors from, say, images will be near, but not exactly in the same place as corresponding entity vectors.

Image embedding We model the image embedding $h : I \to V$ using a convolutional neural network (Fukushima, 1980). For this task, we have access to a set D of labeled images (x, e), where the labels $e \in E$ are entity types from the knowledge graph. Let g^* be the learned entity embedding g obtained by minimizing the objective in Equation 3. In order to embed images into the previously obtained representation space, we propose to use a least-squares objective mapping images onto their corresponding entities:

$$\mathcal{L}_I(h) = \frac{1}{|D|} \sum_{(x,e) \in D} ||h(x) - g^*(e)||_2^2. \quad (4)$$

To predict knowledge graph links for images, we first compute $h(x), x \in I$. Once mapped into the mutual embedding space, the representation of the image $\mathbf{v}$ can be used to evaluate the score function f and predict true triples based on its value. Image embeddings can thus be used individually to determine the properties of each particular image (Figure 2). Additionally, more general properties for a certain object or concept can be inferred by averaging over the set of image embedding vectors for the same object class. For the task studied in this paper, we will use the latter approach in order to determine the representation and properties of a certain new entity based on its image instances. Here, we consider the case where the system has access to sets of images belonging to the same class. However, on a real dataset of unlabeled images about unknown concepts, a natural extension would be to combine our method with algorithms that can determine if two images depict the same concept (Wang et al., 2014).

4 Experiments

Dataset We create a new set of true triples based on the WordNet knowledge graph by reasoning over the links and expanding transitive properties up to depth four. We do this to ensure that the semantic embedding accounts for transitivity. We select only triples where both head and tail are nouns and select the relationships (*hypernym, hyponym, part meronym, part holonym, member meronym, member holonym*). This process roughly yields 10^6 triples, hence we call this dataset WordNet 1 million (WN1M).

The triples are split into three disjoint sets: (a) A training set containing 1 million triples based on entities $e \in E'$, (b) A standard test set containing 20,000 triples based on entities that also occur in the training set ($e \in E'$), (c) A hard test set containing all the triples associated with a selected set of 50 entities $e \in E \setminus E'$ that we held out from sets 1 and 2 and for which images are available.

To train the image embedding, we use the ILSVRC-2012 image dataset (Russakovsky et al., 2015), which we split into four distinct sets: (a) a training set containing images from 750 classes, (b) a validation set (VAL) with distinct images from the same 750 classes, (c) a zero-shot test set (ZS) of 200 image classes that are held out entirely during the training of the image model, but are present in the knowledge graph, (d) an open world test set (OW) of 50 classes that are left out from the image embedding training corresponding and correspond to the 50 entities left out from the knowledge graph embedding training. The classes for each of the previous categories were picked randomly and fixed for all the experiments. Following the protocol of Xian et al. (2017), we also evaluate our method on the 1K most populated classes of ImageNet that are not part of ILSVRC-2012. These 1K entities are present in the knowledge graph, therefore this set falls in the ZS category.

Knowledge embedding We train two knowledge graph embedding functions, one based on the original NTL architecture (Socher et al., 2013a), and one based on our adapted smooth NTL method (SNTL, Equation 3). For both embedding algorithms we choose the dimensionality of the embedding space $d = 60$, with $k = 6$ slices in the layer. A corrupted triple is generated for each true triple by replacing the tail entity with a random one. The Gaussian noise parameter is set to $s = 0.1$, slightly larger than the error obtained from the image embedding models (Equation 4) in each dimension.

Image embedding For the image model, we use the VGG16 architecture (Simonyan and Zisserman, 2015). Following Frome et al. (2013), we use convolution-filter weights which were pre-trained for a classification task. The final layer has a number of units equal to the dimensionality of the knowledge graph embedding. The output is normalized to have unit L_2 norm. We train two separate image embedding models, one with the NTL entity vectors as targets which we call Image NLT (INTL) and one with the smooth SNTL entity vectors as targets (SINTL).

4.1 Performance Evaluation

Knowledge base extension Three metrics that are commonly used to quantify the precision of a link prediction algorithm are: (a) the mean rank of true triples (μ_r) in a list of possible triples sorted by the scoring function, (b) the fraction of the top n triples that is correct (precision, or $p@n$), and (c) the mean reciprocal rank (MRR).

We compare the results of the base (INTL) and smooth (SINTL) models for image link prediction on three different datasets defined in the protocol. To compute metrics on link prediction, we first

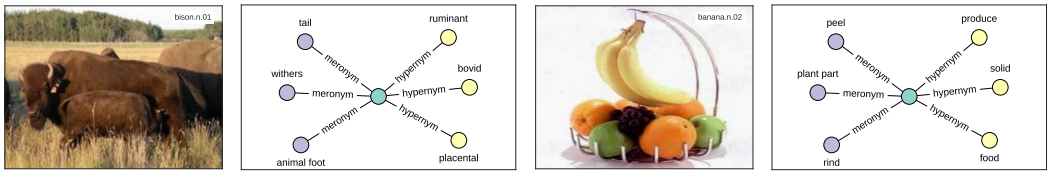

Figure 2: Examples of knowledge graph outputs. Top three hypernyms (A is a kind of B) and top three meronyms (A has part B) are shown. These results are obtained in the open-world setup.

Table 1: Property prediction performance per class.

Dataset		ZS			OW			1K	
Metric	$p@3$	μ_r	MRR	$p@3$	μ_r	MRR	$p@3$	μ_r	MRR
Random	0.001	0.500	0.001	0.001	0.500	0.001	0.001	0.500	0.001
INTL	0.080	0.060	0.158	0.083	0.125	0.167	0.032	0.191	0.067
SINTL	**0.229**	**0.043**	**0.337**	**0.222**	**0.080**	**0.318**	**0.083**	**0.171**	**0.144**

compute the vectors in the semantic space for each image belonging to a particular class, then average those vectors and use the mean in the scoring function to rank triples.

Table 1 shows that our SINTL model significantly outperforms the baseline INTL model on all three datasets. The open-world (OW) case, where classes were also omitted from the knowledge graph performs only slightly worse than the zero-shot (ZS) dataset. Performance on the more challenging 1K dataset is lower.

The ability of our model to predict correct relationships in the *zero-shot* setting demonstrates the capability of our model to generalize to object classes that were not present in image embedding training. This finding proves that incorporating knowledge graphs improves visual recognition capabilities beyond what is possible without side information.

Our model can also make correct predictions in the *open world* setting where object classes are neither in the knowledge graph nor in the image training set. This means we can use images of novel object classes to add new nodes and relations to the knowledge graph. Figure 2 shows selected examples of images with their predicted links.

Smoothness of the semantic space As shown by the numerical results, the smoothness of f greatly improves visual recognition. We now propose to analyze this property of the semantic embedding by estimating its Lipschitz continuity constant ℓ: the lower the value, the smoother the function. This measure quantifies the largest variation of a function under a small change in its input. In practice, we are unable to compute this theoretical metric for our model. Instead, we propose to estimate ℓ as an average value over the training set using the difference in loss function between each triple and its noisy version. We find that our modified loss function yields $\ell = 0.174$ in contrast with $\ell = 0.645$ for the NTL model, proving that our semantic embedding is smooth.

5 Conclusion and Outlook

We presented a mutual semantic embedding space built on images with knowledge base extension capabilities. We showed that using knowledge graphs improves visual recognition by enabling new tasks like predicting properties of objects in open-world images, and extending knowledge graphs. Furthermore, the output of our method can be used in reasoning systems to facilitate automated decision-making based on images.

One significant feature of our method is that it treats on equal footing both visual recognition (for known and unknown classes), as well as knowledge graph extension. Because of this, we speculate that an evolution of this approach could be the foundation of a learning system that updates its knowledge based on visual stimuli and automatically improves its recognition capabilities. Moreover, other data modes could be integrated to the system which has the knowledge graph as its central representation of knowledge.

References

A. Bordes, X. Glorot, J. Weston, and Y. Bengio. Joint learning of words and meaning representations for open-text semantic parsing. In *International Conference on Artificial Intelligence and Statistics*, 2012.

A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2787–2795. 2013.

A. Bordes, X. Glorot, J. Weston, and Y. Bengio. A semantic matching energy function for learning with multi-relational data - application to word-sense disambiguation. *Machine Learning*, 94(2):233–259, 2014.

A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3, 2010.

X. Chen, A. Shrivastava, and A. Gupta. Neil: Extracting visual knowledge from web data. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

S. Huang, M. Elhoseiny, A. Elgammal, and D. Yang. Learning Hypergraph-regularized Attribute Predictors. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2015.

R. Jenatton, N. Le Roux, A. Bordes, and G. R. Obozinski. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3167–3175. 2012.

C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2009.

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data. In *International Conference on Machine Learning (ICML)*, pages 809–816, 2011.

M. Nickel, V. Tresp, and H.-P. Kriegel. Factorizing yago: Scalable machine learning for linked data. In *International Conference on World Wide Web*, pages 271–280, 2012.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Z. Shi, T. M. Hospedales, and T. Xiang. Weakly supervised object-attribute prediction and localisation. In *European Conference on Computer Vision (ECCV)*, 2014.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

R. Socher, D. Chen, C. D. Manning, and A. Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems (NIPS)*, pages 926–934. 2013a.

R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems (NIPS)*, pages 935–943, 2013b.

I. Sutskever, J. B. Tenenbaum, and R. R. Salakhutdinov. Modelling relational data using bayesian clustered tensor factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1821–1828. 2009.

R. Vedantam, X. Lin, T. Batra, C. Lawrence Zitnick, and D. Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE international conference on computer vision*, pages 2542–2550, 2015.

J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning-the good, the bad and the ugly. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

R. Xie, Z. Liu, T.-s. Chua, H. Luan, and M. Sun. Image-embodied knowledge representation learning. *arXiv preprint arXiv:1609.07028*, 2016.