Cross-modal Knowledge Transfer: Improving the Word Embedding of Apple by Looking at Oranges

Fabian Both
Karlsruhe Institute of Technology
faban.both@student.kit.edu

Steffen Thoma
Karlsruhe Institute of Technology
steffen.thoma@.kit.edu

Achim Rettinger
Karlsruhe Institute of Technology
rettinger@.kit.edu

ABSTRACT

Capturing knowledge via learned latent vector representations of words, images and knowledge graph (KG) entities has shown state-of-the-art performance in computer vision, computational linguistics and KG tasks. Recent results demonstrate that the learning of such representations across modalities can be beneficial, since each modality captures complementary information. However, those approaches are limited to concepts with cross-modal alignments in the training data which are only available for just a few concepts. Especially for visual objects exist far fewer embeddings than for words or KG entities. We investigate whether a word embedding (e.g., for "apple") can still capture information from other modalities even if there is no matching concept within the other modalities (i.e., no images or KG entities of apples but of oranges as pictured in the title analogy). The empirical results of our knowledge transfer approach demonstrate that word embeddings do benefit from extrapolating information across modalities even for concepts that are not represented in the other modalities. Interestingly, this applies most to concrete concepts (e.g., dragonfly) while abstract concepts (e.g., animal) benefit most if aligned concepts are available in the other modalities.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**;

KEYWORDS

Multi-Modality, Knowledge Transfer, Word Similarity

ACM Reference Format:
Fabian Both, Steffen Thoma, and Achim Rettinger. 2017. Cross-modal Knowledge Transfer: Improving the Word Embedding of Apple by Looking at Oranges. In *K-CAP 2017: K-CAP 2017: Knowledge Capture Conference, December 4–6, 2017, Austin, TX, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3148011.3148026

1 INTRODUCTION

Unsupervised learning of latent vector representations (embeddings) for a set of concepts has become a key technique in different research communities to capture the raw information in a computable representation. In Computer Vision visual object features are learned from large image collections like ImageNet, in Computational Linguistics word embeddings are extracted from huge text corpora and in the Semantic Web community embeddings of entities are obtained from large knowledge graphs (KGs). The existing embedding approaches have become increasingly sophisticated and implementations have been extensively optimized and trained on huge datasets. However, those well performing models only integrate a subset of all available knowledge, e.g. only visual information is considered when training an image classifier. The reason is a limitation of available labeled or aligned training data.

Since the used raw encodings, methods and datasets are inherently different in each modality, the learned embeddings (fixed sized real-valued vectors) of such specialized models do capture different knowledge about the represented concepts. We pose the question if complementing knowledge can be transferred between concept embeddings from different modalities which were separately trained. A scalable solution to this problem would facilitate multi-modal enrichment procedures without adapting a domain specific model or training task. This obviously has great potential since a model could benefit from advances in the other fields, e.g., by increased availability of training data or improved learning methods.

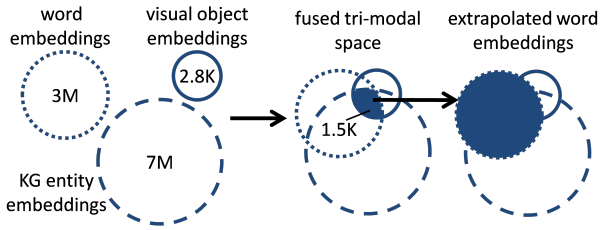


Figure 1: Schematic illustration of knowledge transfer from a fused tri-modal space (arrow on the right). The different circles depict the concept spaces (and quantity of represented concepts) of the single-modal embeddings.

For transferring knowledge between the embedding spaces, information about the same concept has to be aligned across modalities. The key limitation is that the intersection of concepts covered by all embedding spaces is very small within today's training data (1.5k for our experiments). While word embeddings are available for millions of words (in our data 3M, see Fig. 1), there exists only a small sample of a few thousand image embeddings that can be reliably mapped to word-level (2.8K). Thus, when trying to enhance a large embedding space like text with information from much smaller spaces like image embeddings the challenge becomes to extrapolate to concepts not covered in the smaller image space.

K-CAP 2017, December 4–6, 2017, Austin, TX, USA

ACM ISBN 978-1-4503-5553-7/17/12…$15.00
https://doi.org/10.1145/3148011.3148026

In this paper we aim to exploit existing embeddings from three modalities (images, text and KG) in order to augment the space of word embeddings with complementing knowledge from the other two modalities (see Fig. 1). To achieve this, we first construct a tri-modal concept space alike to [24] that is able to capture information of concepts that are available in all modalities (Sec. 2). Next, we propose two approaches, called Feature Mask and Feature Rebuilding, which transfer knowledge back into the single-modal embedding space of words and thus circumvent the limitation of the small multi-modal concept space in [24] (see Sec. 3).

In Sec. 4, we show that single-modal embeddings (e.g., the word embedding of apple in the 3M word space) can benefit from transferred cross-modal knowledge even for concepts that are not explicitly represented in the tri-modal space (since there were no training samples available in the other modalities).

Since those findings are quite remarkable we conducted more experiments to shed some light on the obvious question: For which type of concepts does this extrapolation of knowledge work? The outcomes indicate that (i) homonyms benefit, since shared meanings get disambiguated as shown in the semantic segmentation plots in Sec. 4.2 and (ii) that the abstractness of concepts appears to be a good indicator whether concepts benefit (see Sec. 4.3).

In Sec. 4.4, we briefly summarize our findings before we discuss related work in Sec. 5 and conclude in Sec. 6.

2 CONSTRUCTING A TRI-MODAL CONCEPT SPACE

In the following, embedding models for the textual, KG, and visual domain are introduced. Subsequently, methods for alignment across modalities and fusion methods for these embeddings within a shared cross-modal space are alike to [24]. This is the foundation for our knowledge extrapolation approach which learns an embedding augmentation from a common multi-modal embedding space.

2.1 Single-modal Embeddings

For our work, we use the following well-known embedding approaches[1]:

- In the textual domain, the word2vec model proposed in [14] has been successfully applied to large scale text corpora and pushed state of the art for textual representations and NLP processing. In this work the word2vec model is used for latent vector representations of words and phrases which are constructed from their context.
- KG-concept embeddings are constructed with TransE [2] with type constraints [12]. Its objective function is based on a link prediction task and it has shown to be scalable to knowledge graphs with millions of vertices with good results.
- For visual representations, Inception-V3 [23] is used which was constructed to classify an image into one of multiple available categories. Therefore, convolutional filters are applied in multiple layers, abstracting local features. In this work the last neural layer before the softmax function is used as the latent representation for visual data.

[1]Please note, our transfer method is not limited to these three approaches but also applicable to any other embedding approach, like GLoVE [19] and HOLE [18].

2.2 Alignment to Multi-Modal Concept Space

To share the knowledge between the different embeddings, an alignment between them has to be established. Therefore we map all previously mentioned single-modal representations to a consolidated tri-modal shared concept space. In this work, we chose WordNet *lexemes* (words) as common concepts since numerous evaluation datasets for word similarity tasks are available.

- Textual representations of word2vec do not have to be adjusted since they are already on word-level.
- The alignment of TransE representations is performed by mapping KG-entities (addressable with a unique DBpedia URI) to words. Therefore each KG-concept is mapped to the most commonly used 'surface form' (word) for referring to the KG-concept.
- For aligning the visual embeddings to the word-level, single image representations from *ImageNet-1k* [22] are aggregated to WordNet synset representations via a featurewise max operator (alike to [11]). This is done for all images in a certain synset category. Additionally, 396 more abstract synset representations were built by combining the representation of its child nodes from the WordNet hierarchy. The alignment of these synset representations to *lexemes* can be performed with WordNet directly, since WordNet *lexemes* are assigned to at least one synset.

2.3 Modality Fusion in a Shared Space

To obtain a consolidated tri-modal shared space, that captures knowledge of all modalities in one representation, modality fusion is used. The aligned corpus of n different *lexemes* is represented in three matrices containing the latent vector representations of the three modalities: textual T, knowledge graph G, and visual V. After normalization of each concept vector to unit length, the representations are weighted individually, with weights w_T, w_G, and w_V.

$$M_{weighted} = \begin{bmatrix} w_T \cdot T \\ w_G \cdot G \\ w_V \cdot V \end{bmatrix}, \; w_T + w_G + w_V = 1$$

Besides simple concatenation (CONC) of concept vectors from different modalities, modalities can be fused using dimensionality reduction methods like PCA or SVD. These transformation methods can be computed on the stacked matrix $M_{weighted}$. Thus, the dimension of multi-modal concept vectors can be reduced and statistical smoothing effects can contribute to an overall performance improvement for those fused embeddings. See [25] for an evaluation of performance gains within the tri-modal concept space.

3 TRANSFERRING MULTI-MODAL KNOWLEDGE TO SINGLE MODALITIES

When fusing multi-modal information sources, the size of the common multi-modal space is a limiting factor for useful applications. In our case, the aligned tri-modal concept space has only 1523 concepts. Hereby the visual domain presents the major bottle neck

since humans have to annotate objects in images according to a classification task. In order to overcome the limitation of a small common concept space, we introduce our approach of transferring knowledge from the multi-modal space back into the single-modal space of the utilized models. Neither complex multi-modal learning from aligned corpora, nor extensive multi-modal training on hand-labeled samples is needed. Instead, we rearrange the embedding space of a single model to mimic the superior multi-modal embedding.

In the following, the transfer approaches are subdivided into a fitting and a transfer procedure. For fitting, all n samples in the shared concept space can be utilized. Afterwards, the transfer can be performed with any embedding of the fitted single modality.[2]

3.1 Feature Mask

In the Feature Mask approach, factors for correcting the features of a single-modal embedding are approximated from the common concept space. The features from the original embedding are multiplied with the learned feature mask in order to re-scale features (see Fig. 2a).

Fitting: An artificial similarity dataset is computed for all possible concept pairs in the common concept space. Thereby, the similarity is defined by the cosine vector similarity of the embedding pairs in the shared concept space. For the construction of fused multi-modal embeddings, any combination technique mentioned in section 2.3 can be utilized. The set of these similarity scores $sim \in \mathbb{R}^{\frac{n\cdot(n-1)}{2}}$ is then used as supervised training set.

Consider an embedding $e_{single} = (f_1, ..., f_m)$ from a single modality, m denotes the number of features $f_i \in \mathbb{R}$ for this modality. For each concept pair in the common concept space, the two single-modal embeddings $[e_{single_1}, e_{single_2}]$ are known. A difference vector d between these vector pairs is computed with

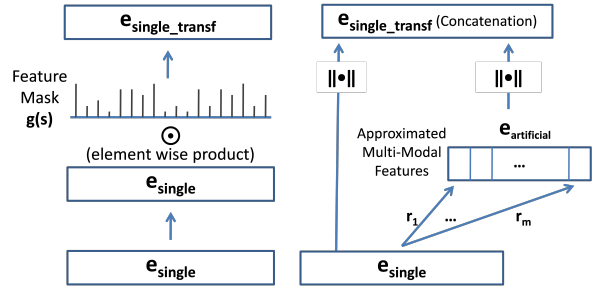
$$d = \frac{e_{single_1} \odot e_{single_2}}{\|e_{single_1}\| \cdot \|e_{single_2}\|} \ (\odot \text{ is element wise multiplication})$$

resulting in $\frac{n\cdot(n-1)}{2}$ difference vectors. These are stacked in the matrix $D \in \mathbb{R}^{\frac{n\cdot(n-1)}{2} \times m}$ where d_i refers to the i-th column of matrix D. With these difference vectors, a scaling factor s_i can be estimated for each feature f_i. A high scaling factor indicates that feature i is useful to differentiate concepts according to the similarity scores in sim. A vector of such scaling factors $s \in \mathbb{R}^m$ can be approximated via pair wise correlation so that $s_i = corr(d_i, sim)$. Another approach is using the coefficients of a multivariate linear regression of the form $sim \sim D \cdot s$.

Transfer: A transformation for all features of a single-modal embedding is computed by using the scaling factor with an activation function g:

$$e_{single\_transf} = e_{single} \odot g(s)$$

A standard activation function is the sigmoid function which also works for negative values s_i. We tested many alternatives and the sigmoid function turned out to be the most reliable and yielding the best results. Note that the value range of s should be adjusted according to the activation function.



(a) Feature mask multiplication. (b) Multi-modal feature rebuilding.

Figure 2: Information transfer approaches for a single modal embedding.

3.2 Feature Rebuilding

Let's consider the simplistic example in which fruits would be only differentiated with one specific feature. If the single-modal embedding would assign feature values from high to low to 'apple', 'coconut' and 'pear', the entity 'apple' could, by simply re-scaling of the fruit feature, never become more similar to 'pear' with regard to cosine similarity than to 'coconut' ('coconut' would stay between 'apple' and 'pear'). In such cases, reordering is needed which can be achieved by adding additional features to a single-modal embedding with our Feature Rebuilding approach. These additional features approximate features from the shared concept space but are only constructed with features of the single-modal embedding (see Fig. 2b).

Fitting: Starting point is a common concept space with k feature dimensions which is obtained by any combination technique from section 2.3. The transfer function is then learned with input from the single-modal embedding e_{single} with m features. Therefore all features from the common concept space are rebuilt with the input of the single embedding. These reconstructed features are further referred to as artificial features. For each artificial feature $i \in [1, 2, ..., k]$, a reconstruction function $r_i : \mathbb{R}^m \rightarrow \mathbb{R}$ is learned. These functions r_i can be fitted via any regression method. In this work we use linear regression and a neural network.

Transfer: For a given embedding e_{single}, all artificial features are computed subsequently and are combined to an artificial embedding:

$$e_{artificial} = (r_1(e_{single}), ..., r_k(e_{single}))$$

The transformed embedding is then obtained by concatenating the normalized original embedding and the normalized artificial embedding[3]:

$$e_{single\_transf} = \left(\frac{e_{single}}{\|e_{single}\|}, \frac{e_{artificial}}{\|e_{artificial}\|}\right)$$

[2]Besides the presented transfer approaches, we also tested information transfer via fitted PCA and SVD transformations on the multi-modal concept space. These methods, however, yielded poor results for information abstraction to concepts outside of the shared space.

[3]Note that normalization is important since differences in the vector dimensions and value ranges would not be meaningful otherwise. See [25] for more details on the effect of normalization.

4 EVALUATION

In the following, we evaluate the fused multi-modal embeddings on standard word similarity evaluation datasets and compare them to the single modal embeddings of word2vec, TransE, and Inception-V3. Thereafter, knowledge transfer methods are applied to improve the word2vec embeddings for words outside of the shared concept space. For this work, pre-trained models are utilized in case of Inception-V3[4] (trained on 1.3 million images) and word2vec[5] (trained on 100 billion words). Since no KG entity representations for a complete KG were readily available, we trained the KG representations by running TransE on the DBpedia KG[6], resulting in over 7 million KG-concept representations.

4.1 Concept Similarity

For evaluation, we considered the extended list of [4] for word pair similarities and relatedness tasks. Therefore, we construct for each word similarity dataset, the relevant subset, for which concept representations are available in all three modalities. Due to the low sample size of some subsets, only *MEN* [3], *WS-353* [5], *SimLex-999* [10] and *MTurk-771* [9] can be used for validation of our tri-modal embeddings. These subsets are also available online[7].

For the fusion methods presented in Sec. 2.3, we use 100 dimensions as we discovered that 100 dimensions are sufficient to encode the relevant information of all aligned representations for the word similarity task. Further, the weighting of the modalities when combining them was optimized over all evaluation subsets. This is an intermediate step to learn the weighting proportions also for the following knowledge transfer. The resulting optimal weight triples are reported in Table 1, exhibiting a similar modality proportion for all combination approaches.

Table 1: Weights for combination methods

	W_{Text}	W_{KG}	W_{Visual}
CONC	0.25	0.15	0.60
SVD	0.25	0.10	0.65
PCA	0.30	0.05	0.65

In Table 2, the performance of the weighted multi-modal combination methods is reported. Weighted combination methods substantially outperform single-modal embeddings on *MEN* and *SIMLEX-999*. Overall, the weighted fusion of modalities successfully improves single-modal embeddings in the tri-modal concept space.

Next, we evaluate our augmented word2vec embeddings as described in Sec. 3 for out-of-training data instances that are not covered by multi-modal concepts. This includes the list of [4]: *MC-30* [15], *MTurk-287* [20], *RG-65* [21], *RW-STANFORD* [13], *VERB-143* [1] and *YP-130* [28], which were not used for validation of the fused tri-modal concepts. Therefore, we subdivide these datasets into 'in-sample' word pairs, which are covered with concepts in the tri-modal shared space, and 'out-of-sample' word pairs as the set of remaining word pairs.

[4]https://storage.googleapis.com/download.tensorflow.org/models/inception_dec_2015.zip
[5]https://code.google.com/archive/p/word2vec
[6]http://wiki.DBpedia.org/services-resources/datasets/data-set-39
[7]https://people.aifb.kit.edu/sto/Transfer

Table 2: Spearman rank correlation score on subsets of the evaluation datasets. Combined embeddings were normalized and weighted.

	MEN	WS-353	SimLex-999	MTurk-771
Inception-V3	0.619	0.526	0.522	0.308
word2vec	0.740	0.707	0.423	0.594
TransE	0.423	0.425	0.246	0.275
CONC	0.806	0.726	0.586	0.589
SVD	**0.847**	0.687	**0.616**	**0.618**
PCA	0.836	**0.760**	0.586	0.568

Feature Mask. The artificial similarity set of the feature mask method is created via concatenation of embeddings from all three modalities and fusion with PCA[8]. The single-modal embeddings are normalized and weighted with the weight triple ($w_{text}, w_{KG}, w_{visual}$) = (0.3, 0.05, 0.65) which is derived from in-sample experiments (see Table 1). With the 1523 concepts from the shared tri-modal space, cosine similarities were computed for over 2.3 million different concept pairs. The scaling factors were approximated with feature wise pearson correlation (Mask_P) and multivariate linear regression (Mask_R), as introduced in Sec. 3.1. The obtained scaling factors s were adjusted with $s' = \frac{2 \cdot (s - median(s))}{max(s) - min(s)}$.

The performance of the Mask_P and the Mask_R knowledge transfer approach with a sigmoid activation function are reported in Table 3. On the in-sample subsets, Mask_R outperforms Mask_P on all evaluation sets, also improving the original word2vec embeddings on all subsets except for *SIMLEX-999*. However, both approaches exhibit a relative small performance increase on most of the full and out-of-sample evaluation sets compared to the word2vec embedding. While the knowledge transfer via Mask_R works for entities within the common concept space, neither Mask_R nor Mask_P extrapolate information well to unseen concepts.

Feature Rebuilding. The multi-modal representations for the feature rebuilding approach are constructed via concatenation of embeddings from Inception-V3 and TransE, which are normalized, weighted with $w_{visual} = 0.95$ and $w_{KG} = 0.05$ and transformed to a 100 dimensional embedding via PCA[9]. Including TransE with a weight of $w_{KG} = 0.05$ improved results compared to visual features only. Excluding the *textual* domain when learning the transfer function for the word2vec embeddings showed superior results. The textual information is incorporated by concatenating word2vec with the artificial embeddings after the reconstruction.

For feature approximation of this 100-dimensional multi-modal embedding space, two different approaches are examined: Rebuild_L, for which a linear regression was applied to each artificial feature and Rebuild_N, for which a neural network with 10 hidden units was trained for each artificial feature. The results for Rebuild_L and Rebuild_N are reported in Table 3.

Both approaches successfully improve word2vec embeddings on all in-sample subsets, thus effectively encoding knowledge from

[8]Note that PCA produces linearly decorrelated features which facilitates feature wise approximations.
[9]We performed a grid search with step size 0.05 on in-sample evaluation sets.

Table 3: Spearman rank correlation of knowledge transfer approaches on word similarity evaluation sets. Thereby, *all* refers to the complete evaluation sets, *in* to in-sample and *out* to out-of-sample subsets. For *RW-STANFORD* only the covered subset of 1863/2034 word pairs in word2vec were evaluated. For *VERB-143* and *YP-130* no in-sample word pairs exist.

		MEN	WS-353	SimLex-999	MTurk-771	MC-30	MTurk-287	RG-65	RW-STANFORD	VERB-143	YP-130
word2vec	all	0.762	0.700	0.442	0.671	0.788	0.687	0.750	0.529	0.474	0.559
	out	0.762	0.685	0.442	0.676	0.791	**0.677**	0.763	0.530	-	-
	in	0.740	0.707	0.423	0.594	-	-	-	-	-	-
$Mask_P$	all	0.768	0.699	0.448	0.678	0.809	**0.688**	0.776	0.523	0.467	0.546
	out	0.768	0.686	0.448	0.683	0.815	**0.677**	0.786	0.525	-	-
	in	0.743	0.702	0.409	0.598	-	-	-	-	-	-
$Mask_R$	all	0.769	0.702	0.444	0.674	0.805	0.682	0.756	0.527	**0.490**	0.555
	out	0.767	0.689	0.444	0.676	0.790	0.672	0.759	0.528	-	-
	in	0.770	0.744	0.420	0.646	-	-	-	-	-	-
$Rebuild_L$	all	0.779	0.697	0.464	0.680	0.816	0.641	0.766	**0.530**	0.424	**0.580**
	out	0.772	0.678	0.458	0.678	0.780	0.628	0.743	**0.532**	-	-
	in	**0.838**	**0.770**	0.581	0.691	-	-	-	-	-	-
$Rebuild_N$	all	**0.799**	**0.703**	**0.467**	**0.692**	**0.854**	0.648	**0.812**	0.517	0.381	0.560
	out	**0.794**	**0.686**	**0.461**	**0.690**	**0.855**	0.638	**0.800**	0.518	-	-
	in	0.824	0.736	**0.588**	**0.711**	-	-	-	-	-	-

the multi-modal concept space. When applying the transfer function to unknown data to overcome the limitation of a small multi-modal concept space, the $Rebuild_N$ method extrapolates better than $Rebuild_L$. Also $Rebuild_N$ dominates the linear regression approach for the full evaluation sets except for *RW-STANFORD*, *VERB-143*, and *YP-130*. Furthermore, $Rebuild_N$ improves the performance of the initial word2vec embeddings on all evaluation sets but *MTurk-287*, *RW-STANFORD*, and *VERB-143*. Also, this enhancement of word2vec embeddings is not caused by a local improvement of concept embeddings covered in the tri-modal space, since we can observe a consistent performance improvement on the out-of-sample subsets. Interestingly, $Rebuild_N$ shows a large performance dip of the initial word2vec embeddings on *VERB-143*, which indicates that the representation of verbs in word2vec might be systematically different from representations of the WordNet nouns for which visual features are available in the tri-modal shared space. This adds up with the effect that a verb maintains a dependency relation with its syntactic arguments (subject and object) and can therefore not as easily be modeled as a noun [8]. Also for *MTurk-287* and *RW-STANFORD*, the knowledge transfer fails due to a higher abstraction complexity. MTurk-287 is a collection of entities from DBpedia with distant relations while *RW-STANFORD* contains rare and complex words exclusively.

4.2 Semantic Segmentation

To illustrate the improvement of out-of-sample embeddings we provide an entity segmentation plot in Fig. 3 with two DBpedia entity types: *land vehicles* and *birds*. For that we calculated the first two PCA components of the respective embeddings of the two types.

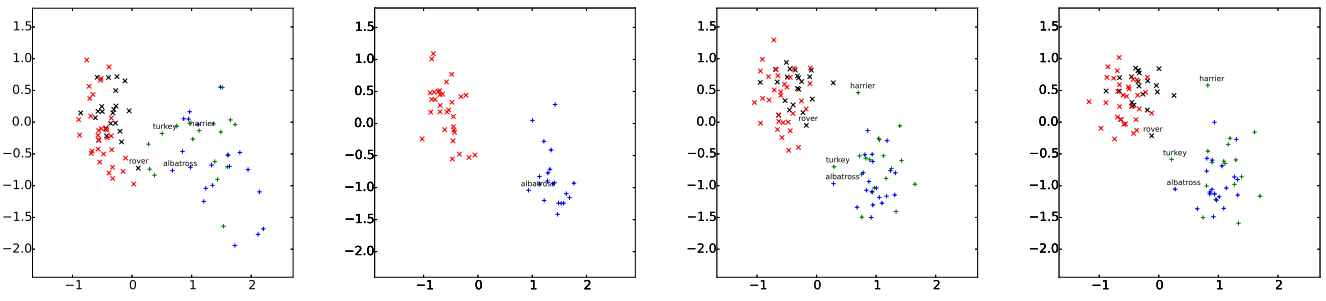
In the pure textual embedding space of word2vec (Fig. 3a), the different types of *land vehicles* and *birds* are not as well separated as in the other three plots. Fig. 3b shows a better separability in the tri-modal space which can be only shown for entities which have a representation in all three modalities text, image, and KG (red 'x' and blue '+'). Fig. 3c and Fig. 3d show that the better separability of the tri-modal space can be transferred to the out-of-sample entities of the types *land vehicles* and *birds* which are shown with black 'x' and green '+'. For $Mask_P$ and $Mask_R$ we omit the corresponding plots since they were inferior when extrapolating to out-of-sample instances.

When we look at the outliers of the clusters, we see that 'harrier' gets closer to the land vehicles and that 'turkey' and 'albatross' get farther away from the bird cluster. This phenomenon is caused by the ambiguity of the words, i.e. harrier, turkey, and albatross are also respective names for a military jet, a country, and a plane. So it is only natural that harrier as an air vehicle gets closer to the other vehicles and turkey and albatross get farther away from the bird cluster.

In summary, this indicates that the tri-modal fused embeddings exhibit a better entity clustering and separation than word2vec alone. This structural improvement is successfully transferred to uni-modal concepts with the $Rebuild_N$ approach, so concept representations are, in this example, indeed augmented in a more intuitive way.

4.3 Concept Abstractness

Further we investigated which word pairs benefit most from the transfer. The performance is measured by the rank difference of the predicted rank and the true rank (similarity) of a given word pair in an evaluation set. A relative performance improvement compared to word2vec is achieved, if the predicted rank of a word pair comes closer to the human provided gold standard ranking. Examples for out-of-sample word pairs with high performance gains relative to word2vec can be seen in Table 4 while word pairs which do not benefit or even decrease in performance are shown in Table 5. We



(a) word2vec (b) tri-modal space (c) Rebuild$_L$ (d) Rebuild$_N$

Figure 3: First two PCA components for various *land vehicles* (red and black 'x') and *birds* (blue and green '+'). The red 'x' and blue '+' are represented in the tri-modal space while the black 'x' and green '+' do not have a tri-modal representation. The segmentation of these two different concept types is improved in the tri-modal space (b) and the improvements are successfully transferred back by both *Rebuild*$_L$ (c) and *Rebuild*$_N$ (d).

investigated the best and worst changes due to the transfer and found an indication that the more concrete a word pair the better the transfer result.

To make "concreteness" quantifiable we calculated a proxy of abstractness based on the WordNet hierarchy. For a given word we compute the depth of the subtree of hyponyms for all word senses that are listed for this particular word in the WordNet hierarchy. We weight these depth scores by their WordNet sense number (depths of frequently encountered word senses are more important), resulting in an average abstractness score for each word. A high score means that a word aggregates many more specific meanings and is thus a proxy for the abstraction level of the word. Our investigations regarding the word type i.e. verb, adjective, noun and so on were inconclusive.

Table 4: Examples for word-pairs with performance gains.

data set	word pair	abstractness
MEN	(bloom,rose)	1.527
	(dragonfly,underwater)	1.000
	(guy,rusty)	1.337
RW-STANFORD	(angrier,huffy)	1.000
	(kingship,rank)	1,234
	(princedom,rank)	1,234
VERB-143	(strike,says)	1.525
	(affected,apply)	1.513

Since the abstractness seems to be a good indicator for the transfer success, we analyzed how the different modalities deal with different levels of abstractness. To illustrate the effect, we calculated the mean abstractness score for the 10 most and 10 least improved word pairs relative to word2vec in each evaluation dataset[10]. In Table 6, you can see the abstractness scores within the tri-modal space and in Table 7 for the textual space for which no tri-modal representations were available.

[10]We average abstractness scores of both words in a word pair.

Table 5: Examples for word-pairs with performance decreases.

data set	word pair	abstractness
MEN	(animal,zoo)	2.792
	(cute,mammal)	3.888
	(nest,reptile)	3.638
RW-STANFORD	(regionalisms,address)	2.128
	(membership,relationship)	2.056
	(brandish,expose)	1.707
VERB-143	(happens,produce)	3.273
	(providing,showing)	2.667

You can observe that augmented word embeddings can deal very well with word pairs of high abstractness within the tri-modal space. Especially representations of abstract concepts benefit from this more holistic modelling through complementary information sources caused by the ability of the transfer function to incorporate multi-modal information. Since the textual word2vec embeddings struggle with those abstract concepts, the performance gain is most prominent in these cases. The visual space is good in capturing the similarity of abstract concepts which is presumably partly caused by the use of the WordNet hierarchy of the *ImageNet-1k* dataset while the KG results are inconclusive.

When evaluating augmented word embeddings of concepts outside of the tri-modal concept space, an interesting shift can be observed as depicted in Table 7. Overall, concrete out-of-sample word pairs are improved the most. Even for evaluation sets with similar concepts to the transfer training space (e.g. out-of-sample word pairs of MEN), relations between concrete words are improved the most by knowledge extrapolation. This makes sense since representations of abstract concepts are presumably incomplete in word2vec. Thus, the learned transfer function is confronted with inaccurate concept representations which impose additional noise apart from approximation errors outside of the tri-modal space. While multi-modal embeddings are well structured with respect

Table 6: Mean abstractness scores of word pairs in the shared tri-modal space of respective evaluation datasets. For text, visual and KG embeddings the 10 best and 10 worst performing word pairs are averaged. In case of $Rebuild_L$ and $Rebuild_N$, the relative improvement through multi-modal information is captured by averaging the abstractness scores of the 10 most and 10 least improved word pairs. Ranking improvement is measured relative to word2vec, i.e. whether the ranking came closer to the human provided gold standard ranking.

	text		visual		KG		$Rebuild_L$		$Rebuild_N$	
	top	bottom	top	bottom	top	bottom	top	bottom	top	bottom
MEN	1.871	**2.659**	**2.312**	1.790	**2.560**	2.153	**2.552**	2.106	**2.694**	2.222
WS-353	2.236	**2.259**	**2.507**	2.287	2.019	**2.907**	2.415	**2.424**	**2.508**	2.243
SimLex-999	1.619	**2.028**	1.603	**2.006**	1.863	**2.242**	**2.166**	1.773	**2.140**	1.846
MTurk-771	1.825	**2.392**	**2.232**	2.197	**2.222**	2.108	**2.199**	1.990	**2.505**	1.979

Table 7: Mean abstractness score for the 10 most and 10 least improved out-of-sample word pairs of the respective evaluation datasets. Ranking improvement is measured relative to word2vec.

	$Rebuild_L$		$Rebuild_N$	
	top	bottom	top	bottom
MEN	1.665	**1.977**	1.809	**2.401**
WS-353	**2.009**	1.992	1.869	**1.932**
SimLex-999	**1.995**	1.559	**2.377**	1.781
MTurk-771	1.943	**2.103**	2.040	**2.162**
MC-30	1.813	**1.978**	1.820	**1.992**
MTurk-287	**1.774**	1.571	**1.995**	1.620
RG-65	1.834	**1.914**	1.851	**1.896**
RW-STANDFORD	1.578	**2.132**	1.446	**1.717**
VERB-143	1.634	**1.960**	1.722	**1.960**
YP-130	**2.090**	2.043	**2.224**	2.010

to abstract concepts, this information is difficult to transfer to the word2vec embedding and restructuring the word2vec space is most challenging in these scenarios with high abstractness scores.

4.4 Summary of Key Findings

Our key findings provide interesting insights for knowledge representations of concepts in general and for concept embeddings in particular:

Concept Similarity: Concept representations fused across three modalities come closer to the human notion of similarity than single-modal embeddings. Less self-evident is that word embeddings without matching concepts in other modalities can be improved by transferring abstract knowledge from the other modalities.

Semantic Segmentation: The tri-modal space improves the semantic segmentation of concepts, specifically for homonyms. Again, not self-evident is that this also extrapolates to the transfer space.

Concept Abstractness: Abstract words improve most by combining embeddings from different modalities. When transferring multi-modal knowledge to out-of-sample embeddings of word2vec, more concrete words benefit.

5 RELATED WORK

In recent years two lines of research in representation learning for exploiting information across modalities have emerged: On the one hand, the fusion of embeddings is achieved after independent training of each modality. On the other hand, embeddings for each modality or the fused embedding space are jointly optimized. Our approach is of the first category and – to the best of our knowledge – the first approach that can exploit available embeddings of more than two modalities and extrapolates cross-modal knowledge from concepts not covered by all modalities.

Examples of the first category include [11] which construct bi-modal concept representations by concatenating independently trained visual and textual representations. Instead of visual representations, Goikoetxea et al. [7] use textual embeddings trained with a text corpus and embeddings of a hierarchical structure learned from WordNet. These embeddings are then concatenated and transformed with various statistical methods (e.g. PCA). Obviously, the fusion approaches above do not include visual, textual and structural knowledge at the same time. An approach that does combine three sources, i.e., multiple languages and images is [16]. However, since text is represented as polylingual topics, the task is reduced to learning bi-modal embeddings. Bruni et al. [3] extend a word representation with knowledge from images with the same tag. While tag based image datasets cover a larger number of concepts than ImageNet, these also introduce noise in the training and alignment steps and still cannot cover enough concepts so that our knowledge transfer approach would become obsolete.

An example for a joint-optimization approach from the second category is [27], in which word embeddings are enriched with information from a knowledge graph. Another approach for joint learning of word embeddings and KG embeddings which are represented in the same vector space is investigated in [26]. They link word embeddings and knowledge graph embeddings during training with an alignment function. But none of the joint learning approaches mentioned so far is able to deal with instances with missing information in some modalities. A joint-optimization approach that is related to our approach in this aspect is [6], since they address (bi-modal) information transfer for instances that are not covered by all modalities. They use a deep visual-semantic embedding model which learns image and tag embeddings simultaneously. They are able to improve visual embeddings of unseen image categories through information encoded in the co-occurrence of words

in a text corpus. In [17], the combination of embeddings of audio and video sources is examined. They apply an encoder and decoder system for joint learning of these embeddings which is also able to handle missing data from one of the modalities during training. In contrast to us, these joint-optimization approaches are dealing with missing information of bi-modal embeddings through a jointly learned common embedding space, instead of infusing knowledge after training. This imposes restrictions on the learning objective and embedding dimension. In addition, all joint learning approaches so far are restricted to two modalities and do not exploit available pre-trained embeddings. Another drawback of joint-optimization approaches in real-world application is that the embeddings learned from an aligned corpus of unstructured content are harder to interpret by humans, since they cannot be related to concepts with explicit semantics. For instance, entity segmentation plots like Fig. 3 cannot be directly constructed.

6 CONCLUSIONS AND FUTURE WORK

This work contributes to the area of knowledge transfer between multi-modal concept representations. Our approach is able to fuse unstructured information of concepts from text and images with structured information from KGs in a meaningful way. This aims towards making computational knowledge representations get closer to a human like perception of concepts.

In order to overcome the restricting number of concepts available in certain modalities (like the visual modality), a novel transfer approach for multi-modal information is presented which extrapolates information from a small common concept space of only 1523 concepts to a large space of another modality (in our case to 3 million word embeddings).

Interestingly, word embeddings do benefit from extrapolating information across modalities even for concepts that are not represented in the other modalities. Our studies indicate that this applies most to concrete concepts and homonyms.

We are confident that our findings will spawn more research in cross-modal knowledge capture. Both, in order to extend the quality, expressiveness and coverage of cross-modal representations but also to obtain a fine-grained understanding of which knowledge benefits how and why.

In the (near) future, we would like to broaden our empirical evaluation to tasks in other modalities, like link prediction in knowledge graphs. Also, extending the number of concepts in the tri-modal space could highly improve the transfer results. Including visual representations of image tags might help to validate that.

REFERENCES

[1] Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An Unsupervised Model for Instance Level Subcategorization Acquisition.. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar*. 278–289.

[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*. 2787–2795.

[3] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research (JAIR)* 49 (2014), 1–47.

[4] Manaal Faruqui and Chris Dyer. 2014. Community Evaluation and Exchange of Word Vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 19–24.

[5] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing Search in Context: The Concept Revisited. In *Proceedings of the 10th international conference on World Wide Web*. 406–414.

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Advances in neural information processing systems*. 2121–2129.

[7] Josu Goikoetxea, Eneko Agirre, and Aitor Soroa. 2016. Single or Multiple? Combining Word Representations Independently Learned from Text and WordNet. In *Thirtieth AAAI Conference on Artificial Intelligence*. 2608–2614.

[8] Yu Gong, Kaiqi Zhao, and Kenny Qili Zhu. 2016. Representing Verbs as Argument Concepts.. In *AAAI*. 2615–2621.

[9] Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-Scale Learning of Word Relatedness with Constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1406–1414.

[10] Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics* 41, 4 (2015), 665–695.

[11] Douwe Kiela and Léon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics.. In *Empirical Methods in Natural Language Processing (EMNLP)*. 36–45.

[12] Denis Krompaß, Stephan Baier, and Volker Tresp. 2015. Type-Constrained Representation Learning in Knowledge Graphs. In *The Semantic Web-ISWC 2015*. Springer, 640–655.

[13] Thang Luong, Richard Socher, and Christopher Manning. 2013. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 104–113.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*. 3111–3119.

[15] George A Miller and Walter G Charles. 1991. Contextual Correlates of Semantic Similarity. *Language and cognitive processes* 6, 1 (1991), 1–28.

[16] Aditya Mogadala and Achim Rettinger. 2015. Multi-modal Correlated Centroid Space for Multi-lingual Cross-Modal Retrieval. In *Advances in Information Retrieval: 37th European Conference on IR Research, ECIR 2015*. Springer International Publishing, Cham, 68–79. https://doi.org/10.1007/978-3-319-16354-3_9

[17] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal Deep Learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.

[18] Maximilian Nickel, Lorenzo Rosasco, and Tomaso A. Poggio. 2016. Holographic Embeddings of Knowledge Graphs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, Dale Schuurmans and Michael P. Wellman (Eds.). AAAI Press, 1955–1961. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12484

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[20] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A Word at a Time: Computing Word Relatedness using Temporal Semantic Analysis. In *Proc. of the 20th international conference on World wide web*. 337–346.

[21] Herbert Rubenstein and John B Goodenough. 1965. Contextual Correlates of Synonymy. *Commun. ACM* 8, 10 (1965), 627–633.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2818–2826. http://dx.doi.org/10.1109/CVPR.2016.308

[24] Steffen Thoma, Achim Rettinger, and Fabian Both. 2017. Knowledge Fusion via Embeddings from Text, Knowledge Graphs, and Images. *arXiv preprint arXiv:1704.06084* (2017).

[25] Steffen Thoma, Achim Rettinger, and Fabian Both. 2017. Towards Holistic Concept Representations: Embedding Relational Knowledge, Visual Attributes, and Distributional Word Semantics. *The Semantic Web - ISWC 2017*.

[26] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph and Text Jointly Embedding. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1591–1601.

[27] Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. 2014. RC-NET: A General Framework for Incorporating Knowledge into Word Representations. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 1219–1228.

[28] Dongqiang Yang and David M. W. Powers. 2006. Verb Similarity on the Taxonomy of WordNet. In *Proceedings of the Third International WordNet Conference — GWC 2006*. Masaryk University, 121–128.