Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean

Felix Hill
Computer Laboratory
University of Cambridge
felix.hill@cl.cam.ac.uk

Anna Korhonen
Computer Laboratory
University of Cambridge
anna.korhonen@cl.cam.ac.uk

Abstract

Models that acquire semantic representations from both linguistic and perceptual input are of interest to researchers in NLP because of the obvious parallels with human language learning. Performance advantages of the multi-modal approach over language-only models have been clearly established when models are required to learn concrete noun concepts. However, such concepts are comparatively rare in everyday language. In this work, we present a new means of extending the scope of multi-modal models to more commonly-occurring abstract lexical concepts via an approach that learns multi-modal embeddings. Our architecture outperforms previous approaches in combining input from distinct modalities, and propagates perceptual information on concrete concepts to abstract concepts more effectively than alternatives. We discuss the implications of our results both for optimizing the performance of multi-modal models and for theories of abstract conceptual representation.

1 Introduction

Multi-modal models that learn semantic representations from both language and information about the perceptible properties of concepts were originally motivated by parallels with human word learning (Andrews et al., 2009) and evidence that many concepts are grounded in perception (Barsalou and Wiemer-Hastings, 2005). The perceptual information in such models is generally mined directly from images (Feng and Lapata, 2010; Bruni et al., 2012) or from data collected in psychological studies (Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013).

By exploiting the additional information encoded in perceptual input, multi-modal models can outperform language-only models on a range of semantic NLP tasks, including modelling similarity (Bruni et al., 2014; Kiela et al., 2014) and free association (Silberer and Lapata, 2012), predicting compositionality (Roller and Schulte im Walde, 2013) and concept categorization (Silberer and Lapata, 2014). However, to date, these previous approaches to multi-modal concept learning focus on concrete words such as *cat* or *dog*, rather than abstract concepts, such as *curiosity* or *loyalty*. However, differences between abstract and concrete processing and representation (Paivio, 1991; Hill et al., 2013; Kiela et al., 2014) suggest that conclusions about concrete concept learning may not necessarily hold in the general case. In this paper, we therefore focus on multi-modal models for learning both abstract and concrete concepts.

Although concrete concepts might seem more basic or fundamental, the vast majority of open-class, meaning-bearing words in everyday language are in fact abstract. 72% of the noun or verb tokens in the British National Corpus (Leech et al., 1994) are rated by human judges[1] as more abstract than the noun *war*, for instance, a concept many would already consider to be quite abstract. Moreover, abstract concepts by definition encode higher-level (more general) principles than concrete concepts, which typically reside naturally in a single semantic category or domain (Crutch and Warrington, 2005). It is therefore likely that abstract representations may prove highly applicable for multi-task, multi-domain or transfer learning models, which aim to acquire 'general-purpose' conceptual knowledge without reference to a specific objective or task (Collobert and Weston, 2008; Mesnil et al., 2012).

In a recent paper, Hill et al. (2014) investigate whether the multi-modal models cited above are

[1] Contributors to the USF dataset (Nelson et al., 2004).

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 255–265, October 25-29, 2014, Doha, Qatar. ©2014 Association for Computational Linguistics

effective for learning concepts other than concrete nouns. They observe that representations of certain abstract concepts can indeed be enhanced in multi-modal models by combining perceptual and linguistic input with an *information propagation* step. Hill et al. (2014) propose ridge regression as an alternative to the nearest-neighbour averaging proposed by Johns and Jones (2012) for such propagation, and show that it is more robust to changes in the type of concept to be learned. However, both methods are somewhat inelegant, in that they learn separate linguistic and 'pseudo-perceptual' representations, which must be combined via a separate *information combination* step. Moreover, for the majority of abstract concepts, the best performing multi-modal model employing these techniques remains less effective than conventional text-only representation learning model.

Motivated by these observations, we introduce an architecture for learning both abstract and concrete representations that generalizes the skipgram model of Mikolov et al. (2013) from text-based to multi-modal learning. Aspects of the model design are influenced by considering the process of human language learning. The model moderates the training input to include more perceptual information about commonly-occurring concrete concepts and less information about rarer concepts. Moreover, it integrates the processes of combining perceptual and linguistic input and propagating information from concrete to abstract concepts into a single representation update process based on back-propagation.

We train our model on running-text language and two sources of perceptual descriptors for concrete nouns: the ESPGame dataset of annotated images (Von Ahn and Dabbish, 2004) and the CSLB set of concept property norms (Devereux et al., 2013). We find that our model combines information from the different modalities more effectively than previous methods, resulting in an improved ability to model the USF free association gold standard (Nelson et al., 2004) for concrete nouns. In addition, the architecture propagates the extra-linguistic input for concrete nouns to improve representations of abstract concepts more effectively than alternative methods. While this propagation can effectively extend the advantage of the multi-modal approach to many more concepts than simple concrete nouns, we observe that the benefit of adding perceptual input appears to decrease as target concepts become more abstract. Indeed, for the most abstract concepts of all, language-only models still provide the most effective learning mechanism.

Finally, we investigate the optimum quantity and type of perceptual input for such models. Between the most concrete concepts, which can be effectively represented directly in the perceptual modality, and the most abstract concepts, which cannot, we identify a set of concepts that cannot be represented effectively directly in the perceptual modality, but still benefit from perceptual input propagated in the model via concrete concepts.

The motivation in designing our model and experiments is both practical and theoretical. Taken together, the empirical observations we present are potentially important for optimizing the learning of representations of concrete and abstract concepts in multi-modal models. In addition, they offer a degree of insight into the poorly understood issue of how abstract concepts may be encoded in human memory.

2 Model Design

Before describing how our multi-modal architecture encodes and integrates perceptual information, we first describe the underlying corpus-based representation learning model.

Language-only Model Our multi-modal architecture builds on the continuous log-linear skipgram language model proposed by Mikolov et al. (2013). This model learns lexical representations in a similar way to neural-probabilistic language models (NPLM) but without a non-linear hidden layer, a simplification that facilitates the efficient learning of large vocabularies of dense representations, generally referred to as *embeddings* (Turian et al., 2010). Embeddings learned by the model achieve state-of-the-art performance on several evaluations including sentence completion and analogy modelling (Mikolov et al., 2013).

For each word type w in the vocabulary V, the model learns both a 'target-embedding' $r_w \in \mathbb{R}^d$ and a 'context-embedding' $\hat{r}_w \in \mathbb{R}^d$ such that, given a target word, its ability to predict nearby context words is maximized. The probability of seeing context word c given target w is defined as:

$$p(c|w) = \frac{e^{\hat{r}_c \cdot r_w}}{\sum_{v \in V} e^{\hat{r}_v \cdot r_w}}$$

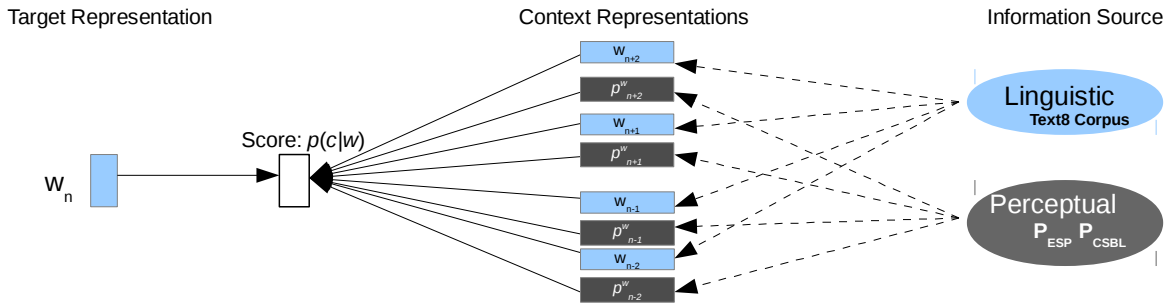

Figure 1: Our multi-modal model architecture. Light boxes are elements of the original Mikolov et al. (2013) model. For target words w_n in the domain of $\mathbf{P}$ (concrete concepts), the model updates its representations based on corpus context words $w_{n\pm i}$, then on words $p^w_{n\pm i}$ in perceptual pseudo-sentences. For w_n not in the domain of $\mathbf{P}$ (abstract concepts), updates are based solely on the $w_{n\pm i}$.

The model learns from a set of target-word, context-word pairs, extracted from a corpus of sentences as follows. In a given sentence S (of length N), for each position $n \leq N$, each word w_n is treated in turn as a target word. An integer $t(n)$ is then sampled from a uniform distribution on $\{1, \ldots k\}$, where $k > 0$ is a predefined maximum context-window parameter. The pair tokens $\{(w_n, w_{n+j}) : -t(n) \leq j \leq t(n), w_i \in S\}$ are then appended to the training data. Thus, target/context training pairs are such that (i) only words within a k-window of the target are selected as context words for that target, and (ii) words closer to the target are more likely to be selected than those further away.

The training objective is then to maximize the sum of the log probabilities T across of all such examples from S and across all sentences in the corpus, where T is defined as follows:

$$T = \frac{1}{N} \sum_{n=1}^{N} \sum_{-t(n) \leq j \leq t(n), j \neq 0} log(p(w_{n+j}|w_n))$$

The model free parameters (target-embeddings and context-embeddings of dimension d for each word in the corpus with frequency above a certain threshold f) are updated according to stochastic gradient descent and backpropation, with learning rate controlled by Adagrad (Duchi et al., 2011). For efficiency, the output layer is encoded as a hierarchical softmax function based on a binary Huffman tree (Morin and Bengio, 2005).

As with other distributional architectures, the model captures conceptual semantics by exploiting the fact that words appearing in similar linguistic contexts are likely to have similar meanings. Informally, the model adjusts its embeddings to increase the 'probability' of seeing the language in the training corpus. Since this probability increases with the $p(c|w)$, and the $p(c|w)$ increase with the dot product $\hat{r}_c \cdot r_w$, the updates have the effect of moving each target-embedding incrementally 'closer' to the context-embeddings of its collocates. In the target-embedding space, this results in embeddings of concept words that regularly occur in similar contexts moving closer together.

Multi-modal Extension We extend the Mikolov et al. (2013) architecture via a simple means of introducing perceptual information that aligns with human language learning. Based on the assumption that frequency in domain-general linguistic corpora correlates with the likelihood of 'experiencing' a concept in the world (Bybee and Hopper, 2001; Chater and Manning, 2006), perceptual information is introduced to the model whenever designated concrete concepts are encountered in the running-text linguistic input. This has the effect of introducing more perceptual input for commonly experienced concrete concepts and less input for rarer concrete concepts.

To implement this process, perceptual information is extracted from external sources and encoded in an associative array $\mathbf{P}$, which maps (typically concrete) words w to bags of perceptual features $\mathbf{b}(w)$. The construction of this array depends on the perceptual information source; the process for our chosen sources is detailed in Section 2.1.

Training our model begins as before on running-text. When a sentence S_m containing a word w in the domain of $\mathbf{P}$ is encountered, the model finishes training on S_m and begins learning from a perceptual pseudo-sentence $\hat{S_m}(w)$. $\hat{S_m}(w)$ is constructed by alternating the token w with a fea-

$\hat{S}(crocodile)$ = **Crocodile** legs **crocodile** teeth **crocodile** teeth **crocodile** scales **crocodile** green **crocodile**.

$\hat{S}(screwdriver)$ = **Screwdriver** handle **screwdriver** flat **screwdriver** long **screwdriver** handle **screwdriver** head.

Figure 2: Example pseudo-sentences generated by our model.

ture sampled at random from $\mathbf{b}(w)$ until $\hat{S_m}(w)$ is the same length as S_m (see Figure 2). Because we want the ensuing perceptual learning process to focus on how w relates to its perceptual properties (rather than how those properties relate to each other), we insert multiple instances of w into $\hat{S_m}(w)$. This ensures that the majority of training cases derived from $\hat{S_m}(w)$ are instances of (w, feature) rather than (feature, feature) pairs. Once training on $\hat{S_m}(w)$ is complete, the model reverts to the next 'genuine' (linguistic) sentence S_{m+1}, and the process continues. Thus, when a concrete concept is encountered in the corpus, its embedding is first updated based on language (moved incrementally closer to concepts appearing in similar linguistic contexts), and then on perception (moved incrementally closer to concepts with the same or similar perceptual features).

For greater flexibility, we introduce a parameter α reflecting the raw quantity of perceptual information relative to linguistic input. When $\alpha = 2$, two pseudo-sentences are generated and inserted for every corpus occurrence of a token from the domain of $\mathbf{P}$. For non-integral α, the number of sentences inserted is $\lfloor \alpha \rfloor$, and a further sentence is added with probability $\alpha - \lfloor \alpha \rfloor$.

In all experiments reported in the following sections we set the window size parameter $k = 5$ and the minimum frequency parameter $f = 3$, which guarantees that the model learns embeddings for all concepts in our evaluation sets. While the model learns both target and context-embeddings for each word in the vocabulary, we conduct our experiments with the target embeddings only. We set the dimension parameter $d = 300$ as this produces high quality embeddings in the language-only case (Mikolov et al., 2013).

2.1 Information Sources

We construct the associative array of perceptual information $\mathbf{P}$ from two sources typical of those used for multi-modal semantic models.

ESPGame Dataset The ESP-Game dataset (ESP) (Von Ahn and Dabbish, 2004) consists of 100,000 images, each annotated with a list of lexical concepts that appear in that image.

For any concept w identified in an ESP image, we construct a corresponding bag of features $\mathbf{b}(w)$. For each ESP image I that contains w, we append the other concept tokens identified in I to $\mathbf{b}(w)$. Thus, the more frequently a concept co-occurs with w in images, the more its corresponding lexical token occurs in $\mathbf{b}(w)$. The array $\mathbf{P_{ESP}}$ in this case then consists of the $(w, \mathbf{b}(w))$ pairs.

CSLB Property Norms The Centre for Speech, Language and the Brain norms (CSLB) (Devereux et al., 2013) is a recently-released dataset containing semantic properties for 638 concrete concepts produced by human annotators. The CSLB dataset was compiled in the same way as the McRae et al. (2005) property norms used widely in multi-modal models (Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013); we use CSLB because it contains more concepts. For each concept, the proportion of the 30 annotators that produced a given feature can also be employed as a measure of the strength of that feature.

When encoding the CSLB data in $\mathbf{P}$, we first map properties to lexical forms (e.g. *is_green* becomes *green*). By directly identifying perceptual features and linguistic forms in this way, we treat features observed in the perceptual data as (sub)concepts to be acquired via the same multi-modal input streams and stored in the same domain-general memory as the evaluation concepts. This design decision in fact corresponds to a view of cognition that is sometimes disputed (Fodor, 1983). In future studies we hope to compare the present approach to architectures with domain-specific conceptual memories.

For each concept w in CSLB, we then construct a feature bag $\mathbf{b}(w)$ by appending lexical forms to $\mathbf{b}(w)$ such that the count of each feature word is equal to the strength of that feature for w. Thus, when features are sampled from $\mathbf{b}(w)$ to create pseudo-sentences (as detailed previously) the probability of a feature word occurring in a sentence reflects feature strength. The array $\mathbf{P_{CSLB}}$ then consists of all $(w, \mathbf{b}(w))$ pairs.

Linguistic Input The linguistic input to all models is the 400m word Text8 Corpus[2] of

[2]From http://mattmahoney.net/dc/textdata.html

ESPGame		CSLB	
Image 1	Image 2	Crocodile	Screwdriver
red	wreck	has 4 legs (7)	has handle (28)
chihuaua	cyan	has tail (18)	has head (5)
eyes	man	has jaw (7)	is long (9)
little	crash	has scales (8)	is plastic (18)
ear	accident	has teeth (20)	is metal (28)
nose	street	is green (10)	
small		is large (10)	

Table 1: Concepts identified in images in the ESP Game (left) and features produced for concepts by human annotators in the CSLB dataset (with feature strength, max=30).

Concept 1	Concept 2	Assoc.
abdomen (6.83)	stomach (6.04)	0.566
throw (4.05)	ball (6.08)	0.234
hope (1.18)	glory (3.53)	0.192
egg (5.79)	milk (6.66)	0.012

Table 2: Example concept pairs (with mean concreteness rating) and free-association scores from the USF dataset.

Wikipedia text, split into sentences and with punctuation removed.

2.2 Evaluation

We evaluate the quality of representations by how well they reflect *free association* scores, an empirical measure of cognitive conceptual proximity. The University of South Florida Norms (USF) (Nelson et al., 2004) contain free association scores for over 40,000 concept pairs, and have been widely used in NLP to evaluate semantic representations (Andrews et al., 2009; Feng and Lapata, 2010; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013). Each concept that we extract from the USF database has also been rated for conceptual concreteness on a Likert scale of 1-7 by at least 10 human annotators. Following previous studies (Huang et al., 2012; Silberer and Lapata, 2012), we measure the (Spearman ρ) correlation between association scores and the cosine similarity of vector representations.

We create separate abstract and concrete concept lists by ranking the USF concepts according to concreteness and sampling at random from the first and fourth quartiles respectively. We also introduce a complementary noun/verb dichotomy, on the intuition that information propagation may occur differently from noun to noun or from noun to verb (because of their distinct structural relationships in sentences). POS-tags are not assigned as part of the USF data, so we draw the noun/verb distinction based on the majority POS-tag of USF concepts in the lemmatized British National Corpus (Leech et al., 1994). The abstract/concrete and noun/verb dichotomies yield four distinct concept lists. For consistency, the concrete noun list is filtered so that each concrete noun concept w has a perceptual representation **b**(w) in both $\mathbf{P_{ESP}}$ and $\mathbf{P_{CSLB}}$. For the four resulting concept lists C (concrete/abstract, noun/verb), a corresponding set of evaluation pairs $\{(w_1, w_2) \in USF : w_1, w_2 \in C\}$ is extracted (see Table 3 for details).

Concept Type	List	Pairs	Examples
concrete nouns	541	1418	*yacht, cup*
abstract nouns	100	295	*fear, respect*
all nouns	666	1815	*fear, cup*
concrete verbs	50	66	*kiss, launch*
abstract verbs	50	127	*differ, obey*
all verbs	100	221	*kiss, obey*

Table 3: Details the subsets of USF data used in our evaluations, downloadable from our website.

3 Results and Discussion

Our experiments were designed to answer four questions, outlined in the following subsections: (1) Which model architectures perform best at *combining* information pertinent to multiple modalities when such information exists explicitly (as common for concrete concepts)? (2) Which model architectures best propagate perceptual information to concepts for which it does not exist explicitly (as is common for abstract concepts)? (3) Is it preferable to include all of the perceptual input that can be obtained from a given source, or to filter this input stream in some way? (4) How much perceptual vs. linguistic input is optimal for learning various concept types?

3.1 Combining information sources

To evaluate our approach as a method of information combination we compared its performance on the concrete noun evaluation set against three alternative methods. The first alternative is simple concatenation of these perceptual vectors with linguistic vectors embeddings learned

by the Mikolov et al. (2013) model on the Text8 Corpus. In the second alternative (proposed for multi-modal models by Silberer and Lapata (2012)), *canonical correlation analysis* (CCA) (Hardoon et al., 2004) was applied to the vectors of both modalities. CCA yields reduced-dimensionality representations that preserve underlying inter-modal correlations, which are then concatenated. The final alternative, proposed by Bruni et al. (2014) involves applying Singular Value Decomposition (SVD) to the matrix of concatenated multi-modal representations, yielding smoothed representations.[3]

When implementing the concatenation, CCA and SVD methods, we first encoded the perceptual input directly into sparse feature vectors, with coordinates for each of the 2726 features in CSLB and for each of the 100,000 images in ESP. This sparse encoding matches the approach taken by Silberer and Lapata (2012), for CCA and concatenation, and by Hill et al. (2014) for the ridge regression method of propagation (see below).

We compare these alternatives to our proposed model with $\alpha = 1$. In The CSLB and ESP models, all training pseudo-sentences are generated from the arrays $\mathbf{P_{CSLB}}$ and $\mathbf{P_{ESP}}$ respectively. In the models classed as *CSLB&ESP*, a random choice between $\mathbf{P_{CSLB}}$ and $\mathbf{P_{ESP}}$ is made every time perceptual input is included (so that the overall quantity of perceptual information is the same).

As shown in Figure 2 (left side), the embeddings learned by our model achieve a higher correlation with the USF data than simple concatenation, CCA and SVD regardless of perceptual input source. With the optimal perceptual source (ESP only), for instance, the correlation is 11% higher that the next best alternative method, CCA.

One possible factor behind this improvement is that, in our model, the learned representations fully integrate the two modalities, whereas for both CCA and the concatenation method each representation feature (whether of reduced dimension or not) corresponds to a particular modality. This deeper integration may help our architecture to overcome the challenges inherent in information combination such as inter-modality differences in information content and representation sparsity. It is also important to note that Bruni et al. (2014) applied their SVD method with comparatively dense perceptual representations extracted from images, whereas our dataset-based perceptual vectors were sparsely-encoded.

[3]CCA was implemented using the *CCA* package in R. SVD was implemented using *SVDLIBC* (`http://tedlab.mit.edu/~dr/SVDLIBC/`), with truncation factor $k = 1024$ as per (Bruni et al., 2014).

3.2 Propagating input to abstract concepts

To test the process of information propagation in our model, we evaluated the learned embeddings of more abstract concepts. We compared our approach with two recently-proposed alternative methods for inferring perceptual features when explicit perceptual information is unavailable.

Johns and Jones In the method of Johns and Jones (2012), pseudo-perceptual representations for target concepts without a perceptual representations (uni-modal concepts) are inferred as a weighted average of the perceptual representations of concepts that do have such a representation (bi-modal concepts).

In the first step of their two-step method, for each uni-modal concept $\mathbf{k}$, a quasi-perceptual representation is computed as an average of the perceptual representations of bi-modal concepts, weighted by the proximity between each of these concepts and $\mathbf{k}$

$$\mathbf{k}^p = \sum_{\mathbf{c} \in \bar{C}} S(\mathbf{k}^l, \mathbf{c}^l)^\lambda \cdot \mathbf{c}^p$$

where $\bar{C}$ is the set of bi-modal concepts, $\mathbf{c}^p$ and $\mathbf{k}^p$ are the perceptual representations for $\mathbf{c}$ and $\mathbf{k}$ respectively, and $\mathbf{c}^l$ and $\mathbf{k}^l$ the linguistic representations. The exponent parameter λ reflects the learning rate.

In step two, the initial quasi-perceptual representations are inferred for a second time, but with the weighted average calculated over the perceptual or initial quasi-perceptual representations of all other words, not just those that were originally bi-modal. As with Johns and Jones (2012), we set the learning rate parameter λ to be 3 in the first step and 13 in the second.

Ridge Regression An alternative, proposed for the present purpose by Hill et al. (2014), uses *ridge regression* (Myers, 1990). Ridge regression is a variant of least squares regression in which a regularization term is added to the training objective to favor solutions with certain properties.

For bi-modal concepts of dimension n_p, we apply ridge regression to learn n_p linear functions

$f_i : \mathbb{R}^{n_l} \rightarrow \mathbb{R}$ that map the linguistic representations (of dimension n_l) to a particular perceptual feature i. These functions are then applied together to map the linguistic representations of uni-modal concepts to full quasi-perceptual representations.

Following Hill et al. (2014), we take the Euclidian l_2 norm of the inferred parameter vector as the regularization term. This ensures that the regression favors lower coefficients and a smoother solution function, which should provide better generalization performance than simple linear regression. The objective for learning the f_i is then to minimize

$$\|\mathbf{a}X - Y_i\|_2^2 + \|\mathbf{a}\|_2^2$$

where $\mathbf{a}$ is the vector of regression coefficients, X is a matrix of linguistic representations and Y_i a vector of the perceptual feature i for the set of bi-modal concepts.

Comparisons We applied the Johns and Jones method and ridge regression starting from linguistic embeddings acquired by the Mikolov et al. (2013) model on the Text8 Corpus, and concatenated the resulting pseudo-perceptual and linguistic representations. As with the implementation of our model, the perceptual input for these alternative models was limited to concrete nouns (i.e. concrete nouns were the only bi-modal concepts in the models).

Figure 3 (right side) shows the propagation performance of the three models. While the correlations overall may seem somewhat low, this is a consequence of the difficulty of modelling the USF data. In fact, the performance of both the language-only model and our multi-modal extension across the concept types (from 0.18 to 0.36) is equal to or higher than previous models evaluated on the same data (Feng and Lapata, 2010; Silberer and Lapata, 2012; Silberer et al., 2013).

For learning representations of concrete verbs, our approach achieves a 69% increase in performance over the next best alternative. The performance of the model on abstract verbs is marginally inferior to Johns and Jones' method. Nevertheless, the clear advantage for concrete verbs makes our model the best choice for learning representations of verbs in general, as shown by performance on the set *all verbs*, which also includes mixed abstract-concrete pairs.

Our model is also marginally inferior to alternative approaches in learning representations of abstract nouns. However, in this case, no method improves on the linguistic-only baseline. It is possible that perceptual information is simply so removed from the core semantics of these concepts that they are best acquired via the linguistic medium alone, regardless of learning mechanism. The moderately inferior performance of our method in such cases is likely caused by its greater inherent inter-modal dependence compared with methods that simply concatenate uni-modal representations. When the perceptual signal is of low quality, this greater inter-modal dependence allows the linguistic signal to be obscured.

The trade-off, however, is generally higher-quality representations when the perceptual signal is stronger, exemplified by the fact that our proposed approach outperforms alternatives on pairs generated from both abstract and concrete nouns (*all nouns*). Indeed, the low performance of the Johns and Jones method on *all nouns* is striking given that: (a) It performs best on *abstract nouns* ($\rho = .282$), and (b) For concrete nouns it reverts to simple concatenation, which also performs comparatively well ($\rho = .249$). The poor performance of the Jobns and Jones method on *all nouns* must therefore derive its comparisons of mixed abstract-concrete or concrete-abstract pairs. This suggests that the pseudo-perceptual representations inferred by this method for abstract concepts method may not be compatible with the directly-encoded perceptual representations of concrete concepts, rendering the comparison computation between items of differing concreteness inaccurate.

3.3 Direct representation vs. propagation

Although property norm datasets such as the CSLB data typically consist of perceptual feature information for concrete nouns only, image-based datasets such as ESP do contain information on more abstract concepts, which was omitted from the previous experiments. Indeed, image banks such as Google Images contain millions of photographs portraying quite abstract concepts, such as *love* or *war*. On the other hand, encodings or descriptions of abstract concepts are generally more subjective and less reliable than those of concrete concepts (Wiemer-Hastings and Xu, 2005). We therefore investigated whether or not it is preferable to include this additional information as model input or to restrict perceptual input

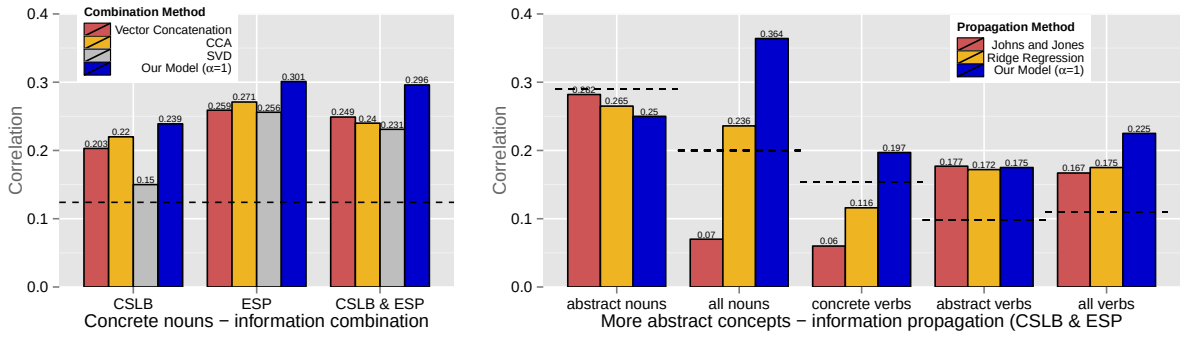

Figure 3: The proposed approach compared with other methods of information combination (left) and propagation. Dashed lines indicate language-only model baseline. For brevity we include both perceptual input sources ESP and CSLB when comparing means of propagation; results with individual information sources were similar.

to concrete nouns as previously.

Of our evaluation sets, it was possible to construct from ESP (and add to $\mathbf{P_{ESP}}$) representations for all of the concrete verbs, and for approximately half of the abstract verbs and abstract nouns. Figure 4 (top), shows the performance of a our model trained on all available perceptual input versus the model in which the perceptual input was restricted to concrete nouns.

The results reflect a clear manifestation of the abstract/concrete distinction. Concrete verbs behave similarly to concrete nouns, in that they can be effectively represented directly from perceptual information sources. The information encoded in these representations is beneficial to the model and increases performance. In contrast, constructing 'perceptual' representations of abstract verbs and abstract nouns directly from perceptual information sources is clearly counter-productive (to the extent that performance also degrades on the combined sets *all nouns* and *all verbs*). It appears in these cases that the perceptual input acts to obscure or contradict the otherwise useful signal inferred from the corpus.

As shown in the previous section, the inclusion of any form of perceptual input inhibits the learning of abstract nouns. However, this is not the case for abstract verbs. Our model learns higher quality representations of abstract verbs if perceptual input is restricted to concrete nouns than if no perceptual input is included at all *and* when perceptual input is included for both concrete nouns and abstract verbs. This supports the idea of a gradual scale of concreteness: The most concrete concepts can be effectively represented directly in the perceptual modality; somewhat more abstract concepts cannot be represented directly in the perceptual modality, but have representations that are improved by propagating perceptual input from concrete concepts via language; and the most abstract concepts are best acquired via language alone.

3.4 Source and quantity of perceptual input

For different concept types, we tested the effect of varying the proportion of perceptual to linguistic input (the parameter α). Perceptual input was restricted to concrete nouns as in Sections 3.1-3.2.

As shown in Figure 4, performance on concrete nouns improves (albeit to a decreasing degree) as α increases. When learning concrete noun representations, linguistic input is apparently redundant if perceptual input is of sufficient quality and quantity. For the other concept types, in each case there is an optimal value for α in the range .5–2, above which perceptual input obscures the linguistic signal and performance degrades. The proximity of these optima to 1 suggests that for optimal learning, when a concrete concept is experienced approximately equal weight should be given to available perceptual and linguistic information.

4 Conclusions

Motivated by the notable prevalence of abstract concepts in everyday language, and their likely importance to flexible, general-purpose representation learning, we have investigated how abstract and concrete representations can be acquired by multi-modal models. In doing so, we presented a simple and easy-to-implement architecture for acquiring semantic representations of both types of

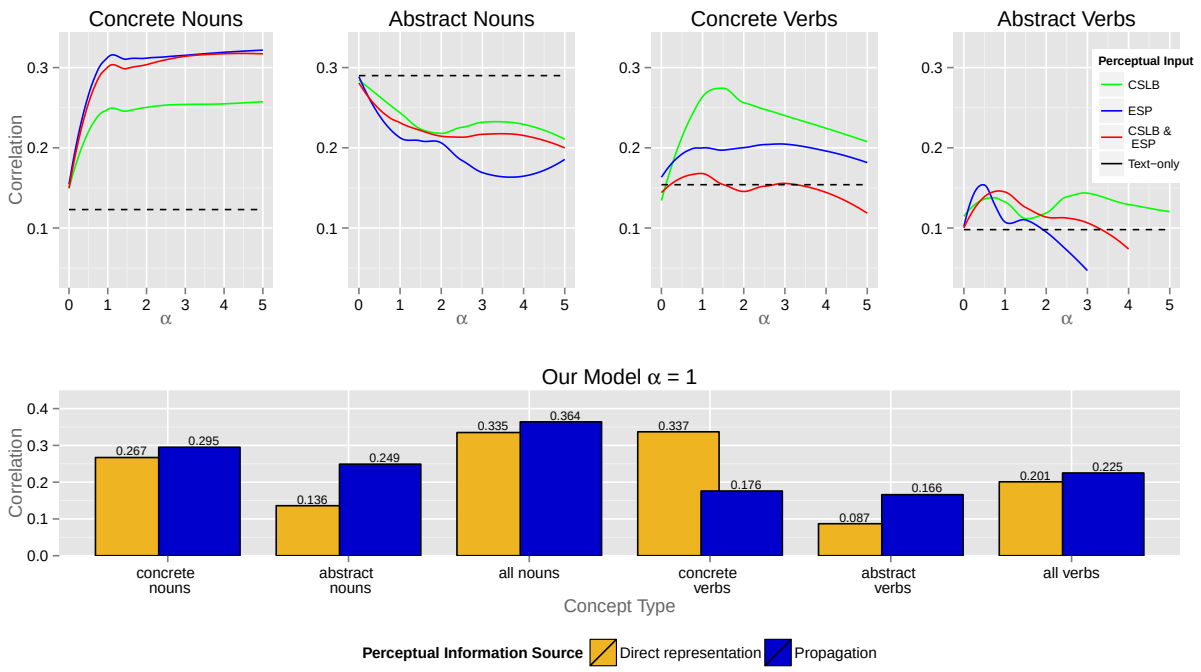

Figure 4: Top: Comparing the strategy of directly representing abstract concepts from perceptual information where available (yellow bars) vs. propagating via concrete concepts. Bottom: The effect of increasing α on correlation with USF pairs (Spearman ρ) for each concept type. Horizontal dashed lines indicate language-only model baseline.

concept from linguistic and perceptual input.

While neuro-probabilistic language models have been applied to the problem of multi-modal representation learning previously (Srivastava and Salakhutdinov, 2012; Wu et al., 2013; Silberer and Lapata, 2014) our model and experiments develop this work in several important ways. First, we address the problem of learning abstract concepts. By isolating concepts of different concreteness and part-of-speech in our evaluation sets, and separating the processes of information combination and propagation, we demonstrate that the multi-modal approach is indeed effective for some, but perhaps not all, abstract concepts. In addition, our model introduces a clear parallel with human language learning. Perceptual input is introduced precisely when concrete concepts are 'experienced' by the model in the corpus text, much like a language learner experiencing concrete entities via sensory perception.

Taken together, our findings indicate the utility of distinguishing three concept types when learning representations in the multi-modal setting.

Type I Concepts that can be effectively represented directly in the perceptual modality. For such concepts, generally concrete nouns or concrete verbs, our proposed approach provides a simple means of combining perceptual and linguistic input. The resulting multi-modal representations are of higher quality than those learned via other approaches, resulting in a performance improvement of over 10% in modelling free association.

Type II Concepts, including abstract verbs, that cannot be effectively represented directly in the perceptual modality, but whose representations can be improved by joint learning from linguistic input and perceptual information about related concepts. Our model can effectively propagate perceptual input (exploiting the relations inferred from the linguistic input) from Type I concepts to enhance the representations of Type II concepts above the language-only baseline. Because of the frequency of abstract concepts, such propagation extends the benefit of the multi-modal approach to a far wider range of language than models based solely in the concrete domain.

Type III Concepts that are more effectively learned via language-only models than multi-modal models, such as abstract nouns. Neither

our proposed approach nor alternative propagation methods achieve an improvement in representation quality for these concepts over the language-only baseline. Of course, it is an empirical question whether a multi-modal approach could ever enhance the representation learning of these concepts, one with potential implications for cognitive theories of grounding (a topic of much debate in psychology (Grafton, 2009; Barsalou, 2010)).

Additionally, we investigated the optimum type and quantity of perceptual input for learning concepts of different types. We showed that too much perceptual input can result in degraded representations. For concepts of type I and II, the optimal quantity resulted from setting $\alpha = 1$; i.e. whenever a concrete concept was encountered, the model learned from an equal number of language-based and perception-based examples. While we make no formal claims here, such observations may ultimately provide insight into human language learning and semantic memory.

In future we will address the question of whether Type III concepts can ever be enhanced via multi-modal learning, and investigate multi-modal models that optimally learn concepts of each type. This may involve filtering the perceptual input stream for concepts according to concreteness, and possibly more elaborate model architectures that facilitate distinct representational frameworks for abstract and concrete concepts.

Acknowledgements

Thanks to the Royal Society and St John's College for supporting this research, and to Yoshua Bengio and Diarmuid Ó Séaghdha for helpful discussions.

References

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463.

Lawrence W Barsalou and Katja Wiemer-Hastings. 2005. Situating abstract concepts. *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought*, pages 129–163.

Lawrence W Barsalou. 2010. Grounded cognition: past, present, and future. *Topics in Cognitive Science*, 2(4):716–724.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Joan L Bybee and Paul J Hopper. 2001. *Frequency and the Emergence of Linguistic Structure*, volume 45. John Benjamins Publishing.

Nick Chater and Christopher D Manning. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167. ACM.

Sebastian J Crutch and Elizabeth K Warrington. 2005. Abstract and concrete concepts have structurally different representational frameworks. *Brain*, 128(3):615–627.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2013. The centre for speech, language and the brain (cslb) concept property norms. *Behavior Research Methods*, pages 1–9.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Yansong Feng and Mirella Lapata. 2010. Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 91–99. Association for Computational Linguistics.

Jerry A Fodor. 1983. *The modularity of mind: An essay on faculty psychology*. MIT press.

Scott T Grafton. 2009. Embodied cognition and the simulation of action to understand others. *Annals of the New York Academy of Sciences*, 1156(1):97–117.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Felix Hill, Anna Korhonen, and Christian Bentz. 2013. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for abstract and concrete concept semantics. *Transactions of the Association for Computational Linguistics*.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Brendan T Johns and Michael N Jones. 2012. Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, 4(1):103–120.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the annual meeting of the Association for Computational Linguistics*. ACL.

Geoffrey Leech, Roger Garside, and Michael Bryant. 1994. Claws4: the tagging of the British National Corpus. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 622–628. Association for Computational Linguistics.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian J Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, et al. 2012. Unsupervised and transfer learning challenge: a deep learning approach. *Journal of Machine Learning Research-Proceedings Track*, 27:97–110.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference of Learning Representations*, Scottsdale, Arizona, USA.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *Proceedings of the international Workshop on Artificial Intelligence and Statistics*, pages 246–252.

Raymond H Myers. 1990. *Classical and Modern Regression with Applications*, volume 2. Duxbury Press Belmont, CA.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255.

Stephen Roller and Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, Seattle, Washington, USA, October. Association for Computational Linguistics.

Carina Silberer and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433. Association for Computational Linguistics.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, August*.

Nitish Srivastava and Ruslan Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *NIPS*, pages 2231–2239.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394. Association for Computational Linguistics.

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 319–326. ACM.

Katja Wiemer-Hastings and Xu Xu. 2005. Content differences for abstract and concrete concepts. *Cognitive Science*, 29(5):719–736.

Pengcheng Wu, Steven CH Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. 2013. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 153–162. ACM.