Illustrative Language Understanding: Large-Scale Visual Grounding with Image Search

Jamie Ryan Kiros* **William Chan*** **Geoffrey E. Hinton**

Google Brain Toronto
{kiros, williamchan, geoffhinton}@google.com

Abstract

We introduce Picturebook, a large-scale lookup operation to ground language via 'snapshots' of our physical world accessed through image search. For each word in a vocabulary, we extract the top-k images from Google image search and feed the images through a convolutional network to extract a word embedding. We introduce a multimodal gating function to fuse our Picturebook embeddings with other word representations. We also introduce Inverse Picturebook, a mechanism to map a Picturebook embedding back into words. We experiment and report results across a wide range of tasks: word similarity, natural language inference, semantic relatedness, sentiment/topic classification, image-sentence ranking and machine translation. We also show that gate activations corresponding to Picturebook embeddings are highly correlated to human judgments of concreteness ratings.

1 Introduction

Constructing grounded representations of natural language is a promising step towards achieving human-like language learning. In recent years, a large amount of research has focused on integrating vision and language to obtain visually grounded word and sentence representations. One source of grounding, which has been utilized in existing work, is image search engines. Search engines allow us to obtain correspondences between language and images that are far less restricted than existing multimodal datasets which typically have restricted vocabularies. While true natural language understanding may require fully embodied cognition, search engines allow us to get a form of quasi-grounding from high-coverage 'snapshots' of our physical world provided by the interaction of millions of users.

One place to incorporate grounding is in the lookup table that maps tokens to vectors. The dominant approach to learning distributed word representations is through indexing a learned matrix. While immensely successful, this lookup operation is typically learned through co-occurrence objectives or a task-dependent reward signal. A very different way to obtain word embeddings is to aggregate features obtained by using the word as a query for an image search engine. This involves retrieving the top-k images from a search engine, running those through a convolutional network and aggregating the results. These word embeddings are grounded via the retrieved images. While several authors have considered this approach, it has been largely limited to a few thousand queries and only a small number of tasks.

In this paper we introduce Picturebook embeddings produced by image search using words as queries. Picturebook embeddings are obtained through a convolutional network trained with a semantic ranking objective on a proprietary image dataset with over 100+ million images (Wang et al., 2014). Using Google image search, a Picturebook embedding for a word is obtained by concatenating the k-feature vectors of our convolutional network on the top-k retrieved search results. The main contributions of our work are as follows:

- We obtain Picturebook embeddings for the 2.2 million words that occur in the Glove vocabulary (Pennington et al., 2014) [1], allowing each word to have a Glove embedding and a parallel grounded word representation. This collection of word representations that we visually

*Both authors contributed equally to this work.

[1] Common Crawl, 840B tokens

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 922–933
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

ground via image search is 2-3 orders of magnitude larger than prior work.

- We introduce a multimodal gating mechanism to selectively choose between Glove and Picturebook embeddings in a task-dependent way. We apply our approach to over a dozen datasets and several different tasks: word similarity, sentence relatedness, natural language inference, topic/sentiment classification, image sentence ranking and Machine Translation (MT).
- We introduce Inverse Picturebook to perform the inverse lookup operation. Given a Picturebook embedding, we find the closest words which would generate the embedding. This is useful for generative modelling tasks.
- We perform an extensive analysis of our gating mechanism, showing that the gate activations for Picturebook embeddings are highly correlated with human judgments of concreteness. We also show that Picturebook gate activations are negatively correlated with image dispersion (Kiela et al., 2014), indicating that our model selectively chooses between word embeddings based on their abstraction level.
- We highlight the importance of the convolutional network used to extract embeddings. In particular, networks trained with semantic labels result in better embeddings than those trained with visual labels, even when evaluating similarity on concrete words.

2 Related Work

The use of image search for obtaining word representations is not new. Table 1 illustrates existing methods that utilize image search and the tasks considered in their work. There has also been other work using other image sources such as ImageNet (Kiela and Bottou, 2014; Collell and Moens, 2016) over the WordNet synset vocabulary, and using Flickr photos and captions (Joulin et al., 2016). Our approach differs from the above methods in three main ways: a) we obtain search-grounded representations for over 2 million words as opposed to a few thousand, b) we apply our representations to a higher diversity of tasks than previously considered, and c) we introduce a multimodal gating mechanism that allows for a more flexible integration of features than mere concatenation.

Our work also relates to existing multimodal models combining different representations of the data (Hill and Korhonen, 2014). Various work has also fused text-based representations with image-based representations (Bruni et al., 2014; Lazaridou et al., 2015; Chrupala et al., 2015; Mao et al., 2016; Silberer et al., 2017; Kiela et al., 2017; Collell et al., 2017; Zablocki et al., 2018) and representations derived from a knowledge-graph (Thoma et al., 2017). More recently, gating-based approaches have been developed for fusing traditional word embeddings with visual representations. Arevalo et al. (2017) introduce a gating mechanism inspired by the LSTM while Kiela et al. (2018) describe an asymmetric gate that allows one modality to 'attend' to the other. The work that most closely matches ours is that of Wang et al. (2018) who also consider fusing Glove embeddings with visual features. However, their analysis is restricted to word similarity tasks and they require text-to-image regression to obtain visual embeddings for unseen words, due to the use of ImageNet. The use of image search allows us to obtain visual embeddings for a virtually unlimited vocabulary without needing a mapping function.

Method	tasks
(Bergsma and Durme, 2011)	bilingual lexicons
(Bergsma and Goebel, 2011)	lexical preference
(Kiela et al., 2014)	word similarity
(Kiela et al., 2015a)	lexical entailment detection
(Kiela et al., 2015b)	bilingual lexicons
(Shutova et al., 2016)	metaphor identification
(Bulat et al., 2015)	predicting property norms
(Kiela, 2016)	toolbox
(Vulic et al., 2016)	bilingual lexicons
(Kiela et al., 2016)	word similarity
(Anderson et al., 2017)	decoding brain activity
(Glavas et al., 2017)	semantic text similarity
(Bhaskar et al., 2017)	abstract vs concrete nouns
(Hartmann and Sogaard, 2017)	bilingual lexicons
(Bulat et al., 2017)	decoding brain activity

Table 1: Existing methods that use image search for grounding and their corresponding tasks.

3 Picturebook Embeddings

Our Picturebook embeddings ground language using the 'snapshots' returned by an image search engine. Given a word (or phrase), we image search for the top-k images and extract the images. We then pass each image through a CNN trained with a semantic ranking objective to extract its embedding. Our Picturebook embeddings reflect the search rankings by concatenating the individual embeddings in the order of the search results. We can perform all of these operations offline to construct a matrix E_p representing the Picturebook embeddings over a vocabulary.

3.1 Inducing Picturebook Embeddings

The convolutional network used to obtain Picturebook embeddings is based off of Wang et al. (2014). Let p_i, p_i^+, p_i^- denote a triplet of query, positive and negative images, respectively. We define the following hinge loss for a given triplet as follows:

$$l(p_i, p_i^+, p_i^-) = \max\{0, g + D(f(p_i), f(p_i^+)) - D(f(p_i), f(p_i^-))\} \tag{1}$$

where $f(p_i)$ represents the embedding of image p_i, $D(\cdot, \cdot)$ is the Euclidean distance and g is a margin (gap) hyperparameter. Suppose we have available pairwise relevance scores $r_{i,j} = r(p_i, p_j)$ indicating the similarity of images p_i and p_j. The objective function that is optimized is given by:

$$\begin{aligned} &\min \sum_i \xi_i + \lambda \|W\|_2^2 \\ &s.t. : l(p_i, p_i^+, p_i^-) \leq \xi_i \\ &\qquad \forall p_i, p_i^+, p_i^- \text{ such that } r(p_i, p_i^+) > r(p_i, p_i^-) \end{aligned} \tag{2}$$

where ξ_i are slack variables and W is a vector of the network's model parameters. The model is trained end-to-end using a proprietary dataset with 100+ million images. We refer the reader to Wang et al. (2014) for additional details of training, including the specifics of the architecture used.

After the model is trained, we can use the convolutional network as a feature extractor for images by computing an embedding vector $f(p)$ for an image p. Suppose we would like to obtain a Picturebook embedding for a given word w. We first perform an image search with query w to obtain a ranked list of images $p_1^w, \dots, p_k^w$. The Picturebook embedding for a word w is then represented as:

$$e_p(w) = [f(p_1^w); f(p_2^w); \dots; f(p_k^w)] \tag{3}$$

namely, the concatenation of the feature vectors in ranked order. In our model, each embedding results in a 64-dimensional vector with the final Picturebook embedding being $64 * k$ dimensions. Most of our experiments use $k = 10$ images resulting in a word embedding size of 640. To obtain the full collection of embeddings, we run the full Glove vocabulary (2.2M words) through image search to obtain a corresponding Picturebook embedding to each word in the Glove vocabulary.

3.2 Visual vs Semantic Similarity

The training procedure is heavily influenced by the choice of similarity function $r_{i,j}$. We consider two types of image similarity: visual and semantic. As an example, an image of a blue car would have high visual similarity to other blue cars but would have higher semantic similarity to cars of the same make, independent of color. In our experiments we consider two types of Picturebook embedding: one trained through optimizing for visual similarity and another for semantic similarity. As we will show in our experiments, the semantic Picturebook embeddings result in representations that are more useful for natural language processing tasks than the visual embeddings.

3.3 Multimodal Fusion Gating

Picturebook embeddings on their own are likely to be useful for representing concrete words but it is not clear whether they will be of benefit for abstract words. Consequently, we would like to fuse our Picturebook embeddings with other sources of information, for example Glove embeddings (Pennington et al., 2014) or randomly initialized embeddings that will be trained. Let $e_g = e_g(w)$ be our other embedding (i.e., Glove) for a word w and $e_p = e_p(w)$ be our Picturebook embedding. We fuse our embeddings using a multimodal gating mechanism:

$$g = \sigma(e_g, e_p) \tag{4}$$

$$e = g \odot \phi(e_g) + (1 - g) \odot \psi(e_p) \tag{5}$$

where σ is a 1 hidden layer DNN with ReLU activations and $\mathrm{sigmoid}$ outputs, ϕ and ψ are 1 hidden layer DNNs with ReLU activations and $\tanh$ outputs. The gating DNN σ allows the model to learn how visual a word is as a function of its input e_p and e_g. Similar gating mechanisms can be found in LSTMs (Hochreiter and Schmidhuber, 1997) and other multimodal models (Arevalo et al., 2017; Wang et al., 2018; Kiela et al., 2018). On some experiments we found it beneficial to include a skip connection from the hidden layer of σ. We chose this form of fusion over other approaches, such as CCA variants and metric learning methods, to allow for easier interpretability and analysis. We leave comparison of alternative fusion strategies for future work.

3.4 Contextual Gating

The gating described above is non-contextual, in the sense that each embedding computes a gate value independent of the context the words occur in. In some cases it may be beneficial to use contextual gates that are aware of the sentence that words appear in to decide how to weight Glove and Picturebook embeddings. For contextual gates, we use the same approach as above except we replace the controller $\sigma(e_g, e_p)$ with inputs that have been fed through a bidirectional-LSTM, e.g. $\sigma(\text{BiLSTM}(e_g),\text{BiLSTM}(e_p))$. We experiment with contextual gating for all experiments that use a bidirectional-LSTM encoder.

3.5 Inverse Picturebook

Picturebook embeddings can be seen as a form of implicit image search: given a word (or phrase), image search the word query and concatenate the embeddings of the images produced by a CNN. Up until now, we have only discussed scenarios where we have a word and we want to perform this implicit search operation. In generative modelling problems (i.e., MT), we want to perform the opposite operation. Given a Picturebook embedding, we want to find the closest word or phrase aligned to the representation. For example, given the word 'bicycle' in English and its Picturebook embedding, we want to find the closest French word that would generate this representation (i.e., 'vélo'). We want to perform this inverse image search operation given its Picturebook embedding.

We introduce a differentiable mechanism which allows us to align words across source and target languages in the Picturebook embedding domain. Let h be our internal representation of our model (i.e., seq2seq decoder state), and e_i be the i-th word embedding from our Picturebook embedding matrix E_p:

$$p(y_i|h) = \frac{\exp(\langle h, e_i \rangle)}{\sum_j \exp(\langle h, e_j \rangle)} \quad (6)$$

Given a representation h, Equation 6 simply finds the most similar word in the embedding space. This can be easily implemented by setting the output softmax matrix as the transpose of the Picturebook embedding matrix E_p. In practice, we find adding additional parameters helps with learning:

$$p(y_i|h) = \frac{\exp(\langle h, e_i + e'_i \rangle + b_i)}{\sum_j \exp(\langle h, e_j + e'_j \rangle + b_j)} \quad (7)$$

where e'_i is a trainable weight vector per word and b_i is a trainable bias per word. A similar technique to tie the softmax matrix as the transpose of the embedding matrix can be found in language modelling (Press and Wolf, 2017; Inan et al., 2017).

4 Experiments

To evaluate the effectiveness of our embeddings, we perform both quantitative and qualitative evaluation across a wide range of natural language processing tasks. Hyperparameter details of each experiment are included in the appendix. Since the use of Picturebook embeddings adds extra parameters to our models, we include a baseline for each experiment (either based on Glove or learned embeddings) that we extensively tune. In most experiments, we end up with baselines that are stronger than what has previously been reported.

4.1 Nearest neighbours

In order to get a sense of the representations our model learns, we first compute nearest neighbour results of several words, shown in Table 2. These results can be interpreted as follows: the words that appear as neighbours are those which have semantically similar images to that of the query. Often this captures visual similarity as well. Some words capture multimodality, such as 'deep' referring both to deep sea as well as to AI. Searching for cities returns cities which have visually similar characteristics. Words like 'sun' also return the corresponding word in different languages, such as 'Sol' in Spanish and 'Soleil' in French. Finally, it's worth highlighting that the most frequent association of a word may not be what is represented in image search results. For example, the word 'is' returns words related to terrorists and ISIS and 'it' returns words related to scary and clowns due to the 2017 film of the same name. We also report nearest neighbour examples across languages in Appendix A.1.

4.2 Word similarity

Our first quantitative experiment aims to determine how well Picturebook embeddings capture word similarity. We use the SimLex-999 dataset (Hill et al., 2015) and report results across 9 categories: all (the whole evaluation), adjectives, nouns, verbs, concreteness quartiles and the hardest 333 pairs. For the concreteness quartiles, the first quartile corresponds to the most abstract words, while the last corresponds to the most concrete words. The hardest pairs are those for which similarity is difficult to distinguish from relatedness. This is an interesting category since image-based word embeddings are perhaps less likely to confuse similarity with relatedness than distributional-based methods. For Glove, scores

language	deep	network	Melbourne	association	sun	life	not
interdisciplinary	deepest	internet	Austin	inclusion	prominence	praising	Nosign
languages	deep-sea	cyberspace	Raleigh	committees	Sol	rejoicing	prohibited
literacy	manta	networks	Cincinnati	social	Soleil	freedom	Forbidden
sociology	depths	blueprints	Yokohama	groupe	Sole	glorifying	no
multilingual	Jarvis	connectivity	Cleveland	members	Venere	worshipping	no-fly
inclusion	cyber	interconnections	Tampa	participation	Marte	healed	forbid
communications	AI	blueprint	Pittsburgh	personnel	eclipses	praise	10
linguistics	hackers	AI	Boston	involvement	Venus	healing	prohibiting
values	restarting	interconnected	Rochester	staffing	eclipse	trust	forbidden
user-generated	diver	tech	Frankfurt	meetings	fireballs	happiness	Stop

Table 2: Nearest neighbours of words. Results are retrieved over the 100K most frequent words.

Model	all	adjs	nouns	verbs	conc-q1	conc-q2	conc-q3	conc-q4	hard
Glove	40.8	**62.2**	42.8	19.6	**43.3**	41.6	42.3	40.2	27.2
Picturebook	37.3	11.7	48.2	17.3	14.4	27.5	46.2	**60.7**	28.8
Glove + Picturebook	**45.5**	46.2	**52.1**	**22.8**	36.7	**41.7**	**50.4**	57.3	**32.5**
Picturebook (Visual)	31.3	11.1	38.8	20.4	13.9	26.1	38.7	47.7	23.9
Picturebook (Semantic)	37.3	11.7	48.2	17.3	14.4	27.5	46.2	60.7	28.8
Picturebook (1)	24.5	2.6	33.5	12.1	4.7	17.8	32.8	47.8	13.6
Picturebook (2)	28.4	6.5	38.9	9.0	5.0	21.3	34.3	55.1	15.7
Picturebook (3)	30.3	11.9	41.9	3.1	2.6	24.3	37.5	58.3	18.4
Picturebook (5)	34.4	6.8	44.5	18.0	9.0	27.9	42.8	58.3	25.9
Picturebook (10)	37.3	11.7	48.2	17.3	14.4	27.5	46.2	60.7	28.8

Table 3: SimLex-999 results (Spearman's ρ). Best results overall are bolded. Best results per section are underlined. Bracketed numbers signify the number of images used. Some rows are copied across sections for ease of reading.

are computed via cosine similarity. For computing a score between 2 word pairs with Picturebook, we set $s(w^{(1)}, w^{(2)}) = -min_{i,j}\ d(e_i^{(1)}, e_j^{(2)})$. [2] That is, the score is minus the smallest cosine distance between all pairs of images of the two words. Note that this reduces to negative cosine distance when using only 1 image per word. We also report results combining Glove and Picturebook by summing their two independent similarity scores. By default, we use 10 images for each embedding using the semantic convolutional network.

Table 3 displays our results, from which several observations can be made. First, we observe that combining Glove and Picturebook leads to improved similarity across most categories. For adjectives and the most abstract category, Glove performs significantly better, while for the most concrete category Picturebook is significantly better. This result confirms that Glove and Picturebook capture very different properties of words. Next we observe that the performance of Picturebook gets progressively better across each concreteness quartile rating, with a 20 point improvement over Glove for the most concrete category. For the hardest subset of words, Picturebook performs slightly better than Glove while Glove performs better across all pairs. We also compare to a convolutional network trained with visual similarity. We observe a performance difference between our visual and semantic embeddings: on all categories except verbs, the semantic embeddings outperform visual ones, even on the most concrete categories. This indicates the importance of the type of similarity used for training the model. Finally we note that adding more images nearly consistently improves similarity scores across categories. Kiela et al. (2016) showed that after 10-20 images, performance tends to saturate. All subsequent experiments use 10 images with semantic Picturebook.

4.3 Sentential Inference and Relatedness

We next consider experiments on 3 pairwise prediction datasets: SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2017) and SICK (Marelli et al., 2014). The first two are natural language inference tasks and the third is a sentence semantic relatedness task. We explore the use of two types of sentential encoders: Bag-of-Words (BoW) and BiLSTM-Max (Conneau et al., 2017a).

[2] We found scoring all pairs of images to outperform scoring only the corresponding equally ranked image.

Model	SNLI		MultiNLI		SICK Relatedness		
	dev	test	dev-mat	dev-mis	test-p	test-s	test-mse
Glove (bow)	85.2	84.2	70.5	69.9	86.8	79.8	25.2
Picturebook (bow)	84.0	83.8	67.9	67.1	85.8	79.3	27.0
Glove + Picturebook (bow)	86.2	85.2	71.3	70.9	**87.2**	**80.9**	**24.4**
BiLSTM-Max (Conneau et al., 2017a)	85.0	84.5					
Glove	86.8	86.3	74.1	**74.5**			
Picturebook	85.2	85.1	70.7	70.3			
Glove + Picturebook	86.7	86.1	73.7	73.7			
Glove + Picturebook + Contextual Gating	**86.9**	**86.5**	**74.2**	74.4			

Table 4: Classification accuracies are reported for SNLI and MulitNLI. For SICK we report Pearson, Spearman and MSE. Higher is better for all metrics except MSE. Best results overall per column are bolded. Best results per section are underlined.

Three sets of features are used: Glove only, Picturebook only and Glove + Picturebook. For the latter, we use multimodal gating for all encoders and contextual gating in the BiLSTM-Max model. For SICK, we follow previous work and report average results across 5 runs (Tai et al., 2015). Due to the small size of the dataset, we only experiment with BoW on SICK. The full details of hyperparameters are discussed in Appendix B.

Table 4 displays our results. For BoW models, adding Picturebook embeddings to Glove results in significant gains across all three tasks. For BiLSTM-Max, our contextual gating sets a new state-of-the-art on SNLI sentence encoding methods (methods without interaction layers), outperforming the recently proposed methods of Im and Cho (2017); Shen et al. (2018). It is worth noting the effect that different encoders have when using our embeddings. While non-contextual gating is sufficient to improve bag-of-words methods, with BiLSTM-Max it slightly hurts performance over the Glove baseline. Adding contextual gating was necessary to improve over the Glove baseline on SNLI. Finally we note the strength of our own Glove baseline over the reported results of Conneau et al. (2017a), from which we improve on their accuracy from 85.0 to 86.8 on the development set. [3]

4.4 Sentiment and Topic Classification

Our next set of experiments aims to determine how well Picturebook embeddings do on tasks that are primarily non-visual, such as topic and sentiment classification. We experiment with 7 datasets provided by Zhang et al. (2015) and compare bag-of-words models against n-gram baselines provided by the authors as well as fastText (Joulin et al., 2017). Hyperparameter details are reported in Appendix B.

Our experimental results are provided in Table 5. Perhaps unsurprisingly, adding Picturebook to Glove matches or only slightly improves on 5 out of 7 tasks and obtains a lower result on AG News and Yahoo. Our results show that Picturebook embeddings, while minimally aiding in performance, can perform reasonably well on their own - outperforming the n-gram baselines of (Zhang et al., 2015) on 5 out of 7 tasks and the unigram fastText baseline on all 7 tasks. This result shows that our embeddings are able to work as a general text embedding, though they typically lag behind Glove. We note that the best performing methods on these tasks are based on convolutional neural networks (Conneau et al., 2017b).

4.5 Image-Sentence Ranking

We next consider experiments that map images and sentences into a common vector space for retrieval. Here, we utilize VSE++ (Faghri et al., 2017) as our base model and evaluate on the COCO dataset (Lin et al., 2014). VSE++ improves over the original CNN-LSTM embedding method of Kiros et al. (2015a) by using hard negatives instead of summing over contrastive examples. We re-implement their model with 2 modifications: 1) we replace the unidirectional LSTM encoder with a BiLSTM-Max sentence encoder and 2) we use Inception-V3 (Szegedy et al., 2016) as our CNN instead of ResNet 152 (He et al., 2016). As in previous work, we report the mean Recall@K (R@K) and the median rank over 1000 images and 5000 sentences. Full details of the hyperparameters are in Appendix B.

Table 6 displays our results on this task.

[3] All reported results on SNLI are available at `https://nlp.stanford.edu/projects/snli/`

Model	AG	DBP	Yelp P.	Yelp F.	Yah. A.	Amz. F.	Amz. P.
BoW (Zhang et al., 2015)	88.8	96.6	92.2	58.0	68.9	54.6	90.4
ngrams (Zhang et al., 2015)	92.0	98.6	95.6	56.3	68.5	54.3	92.0
ngrams TFIDF (Zhang et al., 2015)	92.4	**98.7**	95.4	54.8	68.5	52.4	91.5
fastText (Joulin et al., 2017)	91.5	98.1	93.8	60.4	72.0	55.8	91.2
fastText-bigram (Joulin et al., 2017)	92.5	98.6	**95.7**	**63.9**	72.3	**60.2**	**94.6**
Glove (bow)	**94.0**	98.6	94.4	61.7	**74.1**	58.5	93.2
Picturebook (bow)	92.8	98.5	94.4	61.6	73.3	57.8	92.9
Glove + Picturebook (bow)	93.9	98.6	94.5	61.9	73.8	58.7	93.2

Table 5: Test accuracy [%] on topic and sentiment classification datasets. Best results per dataset are bolded, best results per section are underlined. We compare directly against other bag of ngram baselines.

	Image Annotation				Image Search			
Model	R@1	R@5	R@10	Med *r*	R@1	R@5	R@10	Med *r*
VSE++ (Faghri et al., 2017)	**64.6**		95.7	1	52.0		92.0	1
Glove	**64.6**	88.9	95.5	1	53.7	86.5	94.4	1
Picturebook	62.4	90.2	95.3	1	54.2	86.4	94.3	1
Glove + Picturebook	61.8	89.2	95.0	1	54.1	86.7	**94.7**	1
Glove + Picturebook + Contextual Gating	63.4	**90.3**	**96.5**	1	**55.2**	**87.2**	94.4	1

Table 6: COCO test-set results for image-sentence retrieval experiments. Our models use VSE++. R@K is Recall@K (high is good). Med *r* is the median rank (low is good).

Our Glove baseline was able to match or outperform the reported results in Faghri et al. (2017) with the exception of Recall@10 for image annotation, where it performs slightly worse. Glove+Picturebook improves over the Glove baseline for image search but falls short on image annotation. However, using contextual gating results in improvements over the baseline on all metrics except R@1 for image annotation. Our reported results have been recently outperformed by Gu et al. (2018); Huang et al. (2018b); Lee et al. (2018), which are more sophisticated methods that incorporate generative modelling, reinforcement learning and attention.

4.6 Machine Translation

We experiment with the Multi30k (Elliott et al., 2016, 2017) dataset for MT. We compare our Picturebook models with other text-only non-ensembled models on the Flickr Test2016, Flickr Test2017 and MSCOCO test sets from Caglayan et al. (2017), the winner of the WMT 17 Multimodal Machine Translation competition (Elliott et al., 2017). We use the standard seq2seq (Sutskever et al., 2015) with content-based attention (Bahdanau et al., 2015) model and we describe our hyperparmeters in Appendix B.

Table 7 summarizes our English → German results and Table 8 summarizes our English → French results. We find our models to perform better in BLEU than METEOR relatively compared to (Caglayan et al., 2017). We believe this is due to the fact we did not use Byte Pair Encoding (BPE) (Sennrich et al., 2016), and METEOR captures word stemming (Denkowski and Lavie, 2014). This is also highlighted where our French models perform better than our German models relatively, due to the compounding nature of German words. Since seq2seq MT models are typically trained without Glove embeddings, we also did not use Glove embeddings for this task, but rather we combine randomly initialized learnable embeddings with the fixed Picturebook embeddings. We find the gating mechanism not to help much with the MT task since the trainable embeddings are free to change their norm magnitudes. We did not experiment with regularizing the norm of the embeddings. On the English → German tasks, we find our Picturebook model to perform on average 0.8 BLEU or 0.7 METEOR over our baseline. On the German task, compared to the previously best published results (Caglayan et al., 2017) we do better in BLEU but slightly worse in METEOR. We suspect this is due to the fact that we did not use BPE. On the English → French task, the Picturebook models do on average 1.2 BLEU better or 1.0 METEOR over our baseline.

We also report results for the IWSLT 2014 German-English task (Cettolo et al., 2014) in Table 9. Compared to our baseline, we report a gain of 0.3 and 1.1 BLEU for German → English and English → German respectively. We

Model	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
BPE (Caglayan et al., 2017)	38.1	**57.3**	30.8	**51.6**	26.4	**46.8**
Baseline	38.9	56.5	32.6	50.7	26.8	45.4
Picturebook	39.6	56.9	31.8	50.1	27.7	45.8
Picturebook + Inverse Picturebook	**40.2**	57.2	32.3	50.7	27.8	46.3
Picturebook + Inverse Picturebook + Gating	40.0	**57.3**	**33.0**	51.1	**27.9**	46.5

Table 7: Machine Translation results on the Multi30k English → German task. We note that our models do not use BPE, and we perform better in BLEU relative to METEOR.

Model	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
BPE (Caglayan et al., 2017)	52.5	69.6	50.4	67.5	41.2	61.3
Baseline	60.7	74.1	52.3	67.4	42.8	60.6
Picturebook	61.0	74.2	52.4	67.5	43.1	61.0
Picturebook + Inverse Picturebook	61.8	75.0	52.6	67.7	42.8	61.2
Picturebook + Inverse Picturebook + Gating	**62.1**	**75.2**	**53.6**	**68.4**	**43.8**	**61.6**

Table 8: Machine Translation results on the Multi30k English → French task.

report new state-of-the-art results for the English → German task at 25.4 BLEU, while our German → English model achieves 29.6 BLEU which is slightly behind the recently proposed Neural Phrase-based Machine Translation (NPMT) model at 29.9 (Huang et al., 2018a). We note that the NPMT is not a seq2seq model and can be augmented with our Picturebook embeddings. We also note that our models may not be directly comparable to previously published seq2seq models from (Wiseman and Rush, 2016; Bahdanau et al., 2017) since we used a deeper encoder and decoder.

4.7 Limitations

We explored the use of Picturebook for larger machine translation tasks, including the popular WMT14 benchmarks. For these tasks, we found that models that incorporate Picturebook led to faster convergence. However, we were not able to improve upon BLEU scores from equivalent models that do not use Picturebook. This indicates that while our embeddings are useful for smaller MT experiments, further research is needed on how to best incorporate grounded representations in larger translation tasks.

4.8 Gate Analysis

In this section we perform an extensive analysis of the gating mechanism for models trained across datasets used in our experiments. In our first experiment, we aim to determine how well gate activations correlate to a) human judgments of concreteness and b) image dispersion (Kiela et al., 2014). For concreteness ratings, we use the dataset of Brysbaert et al. (2013) which provides ratings for 40,000 English lemmas. Image dispersion is the average distance between all pairs of images returned from a search query. It was shown in Kiela et al. (2014) that abstract words tend to have higher dispersion ratings, due to having much higher variety in the types of images returned from a query. On the other hand, low dispersion ratings were more associated with concrete words. For each word, we compute the mean gate activation value for Picturebook embeddings. [4] For concreteness ratings, we take the intersection of words that have ratings with the dataset vocabulary. We then compute the Spearman correlation of mean gate activations with a) concreteness ratings and b) image dispersion scores.

Table 10 illustrates the result of this analysis. We observe that gates have high correlations with concreteness ratings and strong negative correlations with image dispersion scores. Moreover, this result holds true across all datasets, even those that are not inherently visual. These results provide evidence that our gating mechanism actively prefers Glove embeddings for abstract words and Picturebook embeddings for concrete words. Appendix A contains examples of words that most strongly activate Glove and Picturebook gates.

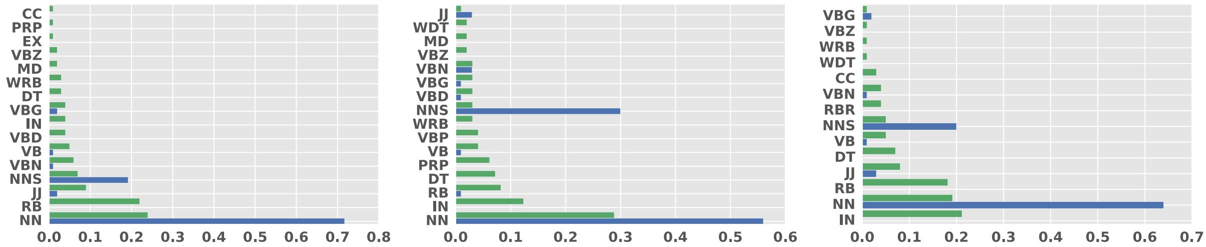
[4] We only consider non-contextualized gates.

Model	DE → EN BLEU	EN → DE BLEU
MIXER (Ranzato et al., 2016)	21.8	
Beam Search Optimization (Wiseman and Rush, 2016)	25.5	
Actor-Critic + Log Likelihood (Bahdanau et al., 2017)	28.5	
Neural Phrase-based Machine Translation (Huang et al., 2018a)	**29.9**	25.1
Baseline	29.3	24.3
Picturebook	29.6	**25.4**

Table 9: Machine Translation results on the IWSLT 2014 German-English task.

Rank	SNLI		MultiNLI		COCO		AG-News		DBpedia		Yelp		Amazon	
	ccorr	disp	ccorr	disp	ccorr	disp	ccorr	disp	ccorr	disp	ccorr	disp	ccorr	disp
top-1%	73	-41	39	-27	53	-22	60	-16	56	-30	47	-28	32	-17
top-10%	54	-39	48	-34	34	-23	52	-24	54	-32	49	-26	50	-30
all	35	-30	30	-27	21	-16	36	-17	39	-30	24	-20	33	-31

Table 10: Correlations (rounded, x100) of mean Picturebook gate activations to human judgements of concreteness ratings (ccorr) and image dispersion (disp) within the specified most frequent words.



(a) SNLI

(b) MultiNLI

(c) AG-News

Figure 1: POS analysis. Top bar for each tag is Glove, bottom is Picturebook. Tags are sorted by Glove frequencies. Results taken over the top 100 mean activation values within the 10K most frequent words.

Finally we analyze the parts-of-speech (POS) of the highest activated words. These results are shown in Figure 1. The highest scoring Picturebook words are almost all singular and plural nouns (NN / NNS). We also observe tags which are exclusively Glove oriented, namely adverbs (RB), prepositions (IN) and determiners (DT).

5 Conclusion

Traditionally, word representations have been built on co-occurrences of neighbouring words; and such representations only make use of the statistics of the text distribution. Picturebook embeddings offer an alternative approach to constructing word representations grounded in image search engines. In this work we demonstrated that Picturebook complements traditional embeddings on a wide variety of tasks. Through the use of multimodal gating, our models lead to interpretable weightings of abstract vs concrete words. In future work, we would like to explore other aspects of search engines for language grounding as well as the effect these embeddings may have on learning generic sentence representations (Kiros et al., 2015b; Hill et al., 2016; Conneau et al., 2017a; Logeswaran and Lee, 2018). Recently, contextualized word representations have shown promising improvements when combined with existing embeddings (Melamud et al., 2016; Peters et al., 2017; McCann et al., 2017; Peters et al., 2018). We expect that integrating Picturebook with these embeddings to lead to further performance improvements as well.

Acknowledgments

The authors would like to thank Chuck Rosenberg, Tom Duerig, Neil Alldrin, Zhen Li, Filipe Gonçalves, Mia Chen, Zhifeng Chen, Samy Bengio, Yu Zhang, Kevin Swersky, Felix Hill and the ACL anonymous reviewers for their valuable advice and feedback.

References

Andrew Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually Grounded and Textual Semantic Models Differentially Decode Brain Activity Associated with Concrete and Abstract Nouns. In *ACL*.

John Arevalo, Thamar Solorio, Manuel Montes y Gomez, and Fabio A. Gonzalez. 2017. Gated Multimodal Units for Information Fusion. In *arXiv:1702.01992*.

Jimmy Ba, Jamie Kiros, and Geoffrey Hinton. 2016. Layer Normalization. In *arXiv:1607.06450*.

Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An Actor-Critic Algorithm for Sequence Prediction. In *ICLR*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Shane Bergsma and Benjamin Van Durme. 2011. Learning Bilingual Lexicons using the Visual Similarity of Labeled Web Images. In *IJCAI*.

Shane Bergsma and Randy Goebel. 2011. Using Visual Information to Predict Lexical Preference. In *RANLP*.

Sai Abishek Bhaskar, Maximilian Koper, Sabine Schulte Im Walde, and Diego Frassinelli. 2017. Exploring Multi-Modal Text+Image Models to Distinguish between Abstract and Concrete Nouns. In *IWCS Workshop on Foundations of Situated and Multimodal Communication*.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher Manning. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research* 49(1).

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3).

Luana Bulat, Stephen Clark, and Ekaterina Shutova. 2017. Speaking, Seeing, Understanding: Correlating semantic models with conceptual representation in the brain. In *EMNLP*.

Luana Bulat, Douwe Kiela, and Stephen Clark. 2015. Vision and Feature Norms: Improving automatic feature norm learning through cross-modal maps. In *NAACL*.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes Garcia-Martinez, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC Submissions for WMT17 Multimodal Translation Task. In *Conference on Machine Translation*.

Mauro Cettolo, Jan Niehues, Sebastian Stuker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *IWSLT*.

Grzegorz Chrupala, Akos Kadar, and Afra Alishah. 2015. Learning language through pictures. In *EMNLP*.

Guillem Collell and Marie-Francine Moens. 2016. Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations. In *COLING*.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined Visual Representations as Multimodal Embeddings. In *AAAI*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017a. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *EMNLP*.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017b. Very deep convolutional networks for text classification. In *EACL*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *EACL: Workshop on Statistical Machine Translation*.

Desmond Elliott, Stella Frank, Loic Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Conference on Machine Translation*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Description. In *ACL: Workshop on Vision and Language*.

Fartash Faghri, David Fleet, Jamie Kiros, and Sanja Fidler. 2017. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *arXiv:1707.05612*.

Goran Glavas, Ivan Vulic, and Simone Paolo Ponzetto. 2017. If Sentences Could See: Investigating Visual Information for Semantic Textual Similarity. In *IWCS*.

Jiuxiang Gu, Jianfei Cai, Shafiq Joty, Li Niu, and Gang Wang. 2018. Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval with Generative Models. In *CVPR*.

Mareike Hartmann and Anders Sogaard. 2017. Limitations of Cross-Lingual Learning from Image Search. In *arXiv:1709.05914*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *NAACL*.

Felix Hill and Anna Korhonen. 2014. Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean. In *EMNLP*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics* 41(4).

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9(8).

Po-Sen Huang, Chong Wang, Sitao Huang, Dengyong Zhou, and Li Deng. 2018a. Towards Neural Phrase-based Machine Translation. In *ICLR*.

Yan Huang, Qi Wu, and Liang Wang. 2018b. Learning Semantic Concepts and Order for Image and Sentence Matching. In *CVPR*.

Jinbae Im and Sungzoon Cho. 2017. Distance-based self-attention network for natural language inference. In *arXiv:1712.02047*.

Hakan Inan, Khashayar Khosravi, and Richard Socher. 2017. Tying Word Vectors and Word Classifiers: A Loss Framework for Languag Modeling. In *ICLR*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *EACL*.

Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. 2016. Learning Visual Features from Large Weakly Supervised Data. In *ECCV*.

Douwe Kiela. 2016. MMFeat: A Toolkit for Extracting Multi-Modal Features. In *ACL: System Demonstrations*.

Douwe Kiela and Leon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *EMNLP*.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2017. Learning Visually Grounded Sentence Representations. In *arXiv:1707.06320*.

Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2018. Efficient Large-Scale Multi-Modal Classification. In *AAAI*.

Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More. In *ACL*.

Douwe Kiela, Laura Rimell, Ivan Vulic, and Stephen Clark. 2015a. Exploiting Image Generality for Lexical Entailment Detection. In *ACL*.

Douwe Kiela, Anita Vero, and Stephen Clark. 2016. Comparing data sources and architectures for deep visual representation learning in semantics. In *EMNLP*.

Douwe Kiela, Ivan Vulic, and Stephen Clark. 2015b. Visual Bilingual Lexicon Induction with Transferred ConvNet Features. In *EMNLP*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2015a. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. In *arXiv:1411.2539*.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015b. Skip-thought vectors. In *NIPS*.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *ACL*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *arXiv:1803.08024*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *ICLR*.

Junhua Mao, Jiajing Xu, Yushi Jing, and Alan Yuille. 2016. Training and Evaluating Multimodal Word Embeddings with Large-scale Web Annotated Images. In *NIPS*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *LREC*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *NIPS*.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *CoNLL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *ICLR Workshop*.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *arXiv:1802.05365*.

Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In *EACL*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks. In *ICLR*.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. 2016. Recurrent Dropout without Memory Loss. In *COLING*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Sen Wang, and Chengqi Zhang. 2018. Reinforced Self-Attention Network: a Hybrid of Hard and Soft Attention for Sequence Modeling. In *arXiv:1801.10296*.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black Holes and White Rabbits: Metaphor Identification with Visual Features. In *NAACL*.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2017. Visually Grounded Meaning Representations. *PAMI* .

Ilya Sutskever, Oriol Vinyals, and Quoc Le. 2015. Sequence to Sequence Learning with Neural Networks. In *NIPS*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR*.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *ACL*.

Steffen Thoma, Achim Rettinger, and Fabian Both. 2017. Towards Holistic Concept Representations: Embedding Relational Knowledge, Visual Attributes, and Distributional Word Semantics. In *ISWC*.

Ivan Vulic, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016. Multi-Modal Representations for Improved Bilingual Lexicon Learning. In *ACL*.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning Fine-grained Image Similarity with Deep Ranking. In *CVPR*.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2018. Learning Multimodal Word Representation via Dynamic Fusion Methods. In *AAAI*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2017. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *arXiv:1704.05426*.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-Sequence Learning as Beam-Search Optimization. In *EMNLP*.

Éloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. 2018. Learning Multi-Modal Word Representation Grounded in Visual Context. In *AAAI*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *NIPS*.