Integrating Vision and Language Datasets to Measure Word Concreteness

Gitit Kehat and **James Pustejovsky**
Department of Computer Science
Brandeis University
Waltham, MA 02453 USA
{gititkeh,jamesp}@brandeis.edu

Abstract

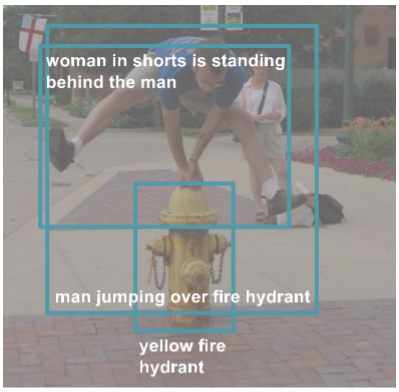
We present and take advantage of the inherent visualizability properties of words in visual corpora (the textual components of vision-language datasets) to compute concreteness scores for words. Our simple method does not require hand-annotated concreteness score lists for training, and yields state-of-the-art results when evaluated against concreteness scores lists and previously derived scores, as well as when used for metaphor detection.

1 Introduction

One of the most pervasive problems in cognitive science, linguistics, and AI has been establishing the semantic relationship between language and vision (Miller and Johnson-Laird, 1976; Winograd, 1972; Jackendoff, 1983; Waltz, 1993). In recent years, new datasets have emerged that enable researchers to approach this question from a new angle: that of determining both how linguistic expressions are grounded in visual images, and how features of visual images are expressible in language. To this end, large vision and language (VL) datasets have become increasingly popular, mostly used in combined VL tasks, such as visual captioning and question answering, image retrieval and more. However, *visual corpora*, the language corpora created in the service of image annotation, have properties that have yet to be exploited. Naturally, they tend to prefer concrete object labels and tangible event descriptions over abstract concepts and private or mental states (Dodge et al., 2012).

In this work, we provide further evidence that *visual corpora* are indeed less abstract than general corpora, and characterize this as a property of what we term a word's *visibility score*. We then show how this notion can be used to measure the concreteness of words, and demonstrate the usefulness of our calculated scores in solving the related problem of metaphor detection.

2 Related Work and Background



I love that crazy cat in the hat.

Figure 1: Visual Genome (top) with multiple captions, and SBU with a user-generated caption per image.

2.1 Abstractness and concreteness

A common notion for the concreteness of a word is to what extent the word represents things that can be perceived directly through the five senses (Brysbaert et al., 2014; Turney et al., 2011), such as *tiger* and *wet*. Accordingly, an abstract word

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 103–108,
Taipei, Taiwan, November 27 – December 1, 2017

represents a concept that is far from immediate perception, or alternatively, could be explained only by other words (as opposed to be demonstrated through image, taste, etc.), like *fun* and *truth*.

Concreteness scores are currently applied in tasks like concept visualization and image description generation, event detection in text and more. Previous methods for measuring words' concreteness used annotated datasets for training. The list by Turney et al. (2011) contains 114k pairs of words and concreteness scores, automatically generated by an algorithm trained on the MRC dataset (Coltheart, 1981). Köper and im Walde (2017) generated a huge concreteness scores list for 3M words, using 32K pairs from the list by Brysbaert et al. (2014) to train a neural network model with high correlation scores with the existed lists.

2.2 Vision and Language Datasets

VL datasets come in different formats, but they all match together visual and textual pieces of information. The visual pieces can be photos, clip-arts, paintings, etc., and the textual ones range from full texts and sentences to single words (see Figure 1). There are surprisingly few works that analyze *visual corpora* in terms of their linguistic properties. Dodge et al. (2012) found that Flickr captions have more references to physical objects. Ferraro et al. (2015) compared visual corpora using a set of linguistic criteria, including an abstract-to-concrete ratio to estimate the concreteness level of a corpus. We further discuss this task in Section 3.

3 The Concreteness Level of a Corpus

We demonstrate the differences in the concreteness level of several corpora using two concreteness ratings (or "concreteness scores") lists, each contains pairs of a word and a score, in some scale, and potentially additional meta-data regarding the annotation agreement. See Table 1 for examples.

The list of **40K concreteness ratings** by Brysbaert et al. (2014) contains ratings from 1.0 (abstract) to 5.0 (concrete) for almost 40K terms, 37K of them are unigrams[1], along with metadata like the standard deviation over the scores assigned to a term by the 30 annotators. The authors aimed to represent all English lemmas, for each they included several forms, each was scored separately (according to the definition in Section 2.1).

[1]The rest are bigrams, we worked with unigrams only.

	40K (1.04-5.0)	MRC (158-670)
turtle	5.0 (sd=0.0)	644
boat	4.93 (sd=0.37)	637
milk	4.92 (sd=0.39)	670
side	3.68 (sd=1.33)	394
symbol	3.11 (sd=1.37)	402
clean	3.07 (sd=1.41)	392
impossible	1.66 (sd=1.06)	198
immortality	1.52 (sd=0.87)	209
justification	1.52 (sd=0.83)	219

Table 1: Examples for words in the concreteness lists annotated as mostly concrete, in the middle, and mostly abstract.

The **MRC psycholinguistic database** (Coltheart, 1981) contains 4,295 words and concreteness scores (range from 158 to 670), given by human subjects through psychological experiments.

3.1 Descriptions of Corpora Studied

Brown corpus (Francis and Kucera, 1964). Following Ferraro et al. (2015), a representative of a non-visual "general"/"balanced" corpus.

Visual Genome (Krishna et al., 2016). The largest VL dataset to date, containing 5.4M region descriptions for more than 108K images, visual question answers and more, all created through crowd-sourcing. We used the set of all region descriptions (see Figure 1) as corpus.

SBU Captioned Photo Dataset (Ordonez et al., 2011). Another large scale dataset, containing user generated image descriptions for 1M images, created by quering Flickr. As a result, the captions are not necessarily full or accurate (see Figure 1).

Flickr 30K (Young et al., 2014). 5 captions per image for more than 31K real-world images from Flickr, created through crowd-sourcing.

Microsoft COCO (Lin et al., 2014). Includes object segmentation and 5 captions per image for more than 300K images from Flickr.

ImageNet (Deng et al., 2009). A dataset matching images and the corresponding WordNet synsets (Miller et al., 1990). We gathered all available annotated synsets as the ImageNet corpus.

We also created a set of non-visual Brown corpora by subtracting each of the corpora from the Brown corpus, to each we refer as $Brown_{NV} - VC$ in relation to some visual corpus VC.

3.2 Setup and Comparison Results

Given a corpus, we divided the words in each concreteness scores list into two non-overlapping sets (words contained in the corpus and words not

Corpus	C-list	Words-in	Words-out	Ave-in	Ave-out	Diff/Range%	Abs-ratio
Brown	40K	18191	18742	3.02	2.91	2.74%	15.24%
	MRC	3639	553	442.17	443.62	-0.28%	
Visual Genome	40K	14968	21965	3.5	2.61	22.49%	—
	MRC	3263	929	465.62	360.66	20.5%	
MSCOCO	40K	11786	25147	3.52	2.71	20.52%	12.96%
	MRC	2919	1273	469.21	380.79	17.26%	
Flickr 30k	40K	9874	27059	3.57	2.75	20.76%	14.98%
	MRC	2669	1523	471.45	391.38	15.63%	
ImageNet	40K	8397	28536	3.96	2.68	32.52%	—
	MRC	2365	1827	505.31	360.87	28.21%	
SBU	40K	20746	16187	3.3	2.55	18.85%	3.74%
	MRC	3789	403	452.67	345.41	20.94%	

Table 2: Corpus concreteness measuring using different concreteness score lists.

Corpus	D/R% 40K	D/R% MRC
$Brown_{NV} - VG$	-14.47%	-20.35%
$Brown_{NV} - MSCOCO$	-11.37%	-17.13%
$Brown_{NV} - Flickr30k$	-10.30%	-15.76%
$Brown_{NV} - ImageNet$	-12.15%	-26.21%
$Brown_{NV} - SBU$	-15.13%	-22.28%

Table 3: The Diff/Range% of the non-visual Brown corpora.

contained in the corpus), and calculated the average concreteness score of each set, as well as the difference of the two averages normalized by the score range of the list ('Diff/Range%') (see Table 2). We can see the clear differences between the concreteness level of the Brown corpus (negligible Diff/Range%) and the rest of the visual corpora (15.0% - 32%), which show nicely that the Brown corpus is indeed "balanced" in terms of concreteness. The 'Abs-ratio' column refers to previous results by Ferraro et al. (2015), who calculated an abstract-to-concrete ratio (Abs-ratio) with a fixed common-abstract-terms list, where corpus words in the list were considered as "abstract" and the rest as "concrete". The results were highly dependent on corpus size (with more words outside the fixed list ("concrete") as vocabulary grows). Accordingly, the Abs-ratios of the Brown corpus and most of the visual corpora were very similar, and large corpora such as the SBU got significantly lower ratios.

Table 3 shows the same calculations on the non-visual Brown corpora. The large negative ratios ((-26)% - (-10)%) show that these corpora are less concrete than the original Brown corpus, and are much more abstract than all the visual corpora.

4 Predicting Concreteness Scores

The leading principal here is that words contained in visual corpora tend to have significantly higher concreteness scores, and words in non-visual corpora tend to have significantly lower scores. We do not use concreteness scores lists for training, but only a visual corpus and a generic corpus to build *visibility scores* for each word, from which a concreteness score is estimated.

4.1 Visibility Scores

The *concreteness score* of a word w consists of the *concrete visibility score* and the *abstract visibility score*, both are normalized sums computed in the same manner (with only the reference corpus different, a visual for the concrete case and a non-visual for the abstract case). Each term $nei(w)$ in the set of n-best nearest neighbors of w (extracted from a model of 300-dimensional vectors for 3M terms from the Google News dataset [2]) contained in the reference corpus contributes its cosine similarity to w ($in(w) = 1.0$ if w is in the reference corpus, o/w 0.0), then the sum is normalized by the sum of all similarities:

$$ConVisEmbScore(w) = \frac{in(w) + \sum_{nei(w) \in VisCor} Sim(w, nei)}{in(w) + \sum_{nei(w)} Sim(w, nei)}, \quad (1)$$

$$AbsVisEmbScore(w) = \frac{in(w) + \sum_{nei(w) \in Brown-VisCor} Sim(w, nei)}{in(w) + \sum_{nei(w)} Sim(w, nei)}, \quad (2)$$

[2] available at https://code.google.com/archive/p/word2vec

max-sd	num-neigh	Spearman	Pearson	MSE
max(sd) =1.89	Turney	0.74	0.74	0.58
	100	0.72	0.72	0.61
	50	0.71	0.70	0.72
	10	0.63	0.62	1.11
med(sd) =1.22	Turney	0.79	0.81	0.89
	100	0.78	0.81	0.69
	50	0.77	0.79	0.78
	10	0.71	0.73	1.13
mean(sd) =1.16	Turney	0.80	0.82	0.99
	100	**0.79**	**0.82**	0.69
	50	0.78	0.81	0.78
	10	0.72	0.75	1.12

Table 4: Predicting concreteness scores.

The overall concreteness score for w is then:

$$ConcretenessScore(w) = ConVisEmbScore(w) - AbsVisEmbScore(w) \quad (3)$$

(2) and (1) range from 0.0 (non of the neighbors is in the reference corpus) to 1.0 (all of them are). Hence, (3) ranges between -1.0 and 1.0, where a higher score means more concrete word.

Notice that no corpus-frequencies were taken into account in the above sums. This is because VL datasets are often human-focused with unrealistic high-weight for terms describing people. In addition, words in the corpora were only lowercaesd but not stemmed since we noticed it cut off too much information, leading to poorer results (due to the loss of potential discriminating concreteness features that are characteristic of many derivational suffixes). For example, 40K's scores for several unstemmed forms of the stem *woman*: woman (4.46), womanhood (2.55), womanishness (1.79), womanize (2.82), womanlike (3.14).

4.2 Results and Discussion

To best demonstrate the strength of our method, we present the results gathered by using a unified visual corpus that is both large enough to be used as a reference corpus, and has higher Diff/Range ratios. This unified corpus, which we call the ***Big Visual Corpus (BVC)*** consists of the Visual Genome, MSCOCO, Flickr30K, and ImageNet, and contains over 98K lowercaesd (but otherwise non-normalized) terms. Its Diff/Range%, according to the 40K and MRC lists are 25.5% and 24.53%, respectively. The generic corpus used is the Brown corpus, and respectively, the non-visual reference corpus is $Brown_{NV} - BVC$.

We follow the simple practice from Köper and im Walde (2017) and map all scores into the same interval using the following continuous function:

$$f(w) = \frac{(b-a)(x-min)}{max-min} + a, \quad (4)$$

where $[min, max]$ is the original interval and $[a, b]$ the new interval. In our case, $a = 1.04$ (the minimum unigram score in the 40K list) and $b = 5.0$. We then compute the correlation between our scores and the 40K list's scores and compare them to the correlations of the previously calculated scores by Turney et al. (2011) (see Table 4). We parameterize over both the number of neighbors (up to 100) taken into account in (1) and (2) and the maximal standard deviation (sd) of words in the 40K list we consider in computing the correlations. Using the mean sd as a threshold shows better correlations with the subset considered. Also, considering more neighbors improves all evaluation metrics.

5 Metaphor Detection

We utilize our concreteness scores to solve the task of Metaphor Detection, where a set of literal and non-literal samples is given, and the goal is to classify each into the correct class. We follow Black's (1979) observation that a metaphor is essentially an interaction between two terms, creating an implication-complex to resolve two incompatible meanings. Operationally, we follow Turney et al. and their adoption of Lakoff and Johnson's (1980) notion that metaphor is a way to move knowledge from a concrete domain to an abstract one. Hence, there should be a correlation between the "degree of abstractness in a word's context [...] with the likelihood that the word is used metaphorically." (Turney et al., 2011). We show our results on two annotated datasets:

5.1 The TSV Dataset

This dataset by Tsvetkov et al. (2014) includes several sets with instances annotated as "metaphorical" or "literal" by 5 annotators, from which we experimented with two sets. The first set, which we call TSV-AN, contains 200 adjective-noun (AN) pairs, 100 instances per class. For example, "clean conscience" is annotated as metaphorical, and "clean air" as literal. The second set, which we call TSV-SVO, contains subject-verb-object (SVO) triples or pairs (when missing 'S'/'O'), 111 for each class.

We build a logistic regression model using 10-fold cross-validation for each of the TSV-AN and

Dataset	Features	Precision	Recall	F1
TSV-AN	Linguistic	0.73	0.80	0.76
	Visual	0.60	0.91	0.73
	Multimodal	0.67	0.96	0.79
	Vis-Emb.	0.84	0.72	0.77
TSV-SVO	**Vis-Emb.**	0.83	0.80	0.81

Table 5: Results of our method (Vis-Emb.) on the dataset by Tsvetkov et al. compared to previous results by Shutova et al.

TSV-SVO sets. The feature vector for each phrase in the sets is simple, consists of our assigned concreteness score for each word in the phrase. For the second set, we divide each triple into two pairs to get 150 "literal" 'SV'/'VO' pairs and 165 "metaphorical" ones. We flipped the feature vector of the 'VO' pairs to represent scores in the form of 'OV', so that the nouns and verb would appear at consistent places in the vector. As a reference to our results, we bring previous results by Shutova et al. (2016), who used linguistic embedding model, visual embedding model, and a multimodal model that mixed the two (see Table 5).

5.2 The TroFi dataset

The dataset by Birke and Sarkar (2006) contains annotated "literal" and "non-literal" sentences from the Wall Street Journal for 50 verbs. We follow the exact same algorithm used in Turney et al. (2011) on a subset of 25 verbs, while replacing their concreteness scores with ours.

We build a 5-dimensional feature vector for each sentence, composed of the average concreteness score of words with each of the following part-of-speech tags: noun, proper noun, verb, adjective, adverb. When there are no words with a specific POS tag in the sentence, the value 0.0 is assigned to the corresponding place in the vector. The feature vectors are then used in a logistic regression classifier to build a separate model for each verb using 10-fold cross-validation. Table 6 shows our results along with the previous results by Turney et al. (and their probability-matching case).

6 Conclusion and Future Work

Even without the matching images, the captions in vision and language datasets contain useful information regarding the visibility of words appearing in them. In addition, the connection between the visibility of a word and its concreteness level is well known from psychological experiments.

Features	Accuracy	F1-score
Vis-Emb.	0.713	0.657
Turney et al.	0.734	0.639
Probability Matching	0.605	0.500

Table 6: Classifying the sentences related to 25 verbs in the TroFi dataset. Accuracy and F1-score are macro-averaged.

We exploited these properties in crafting visibility scores, based only on the occurrences of a word's neighbors (in the semantic space) in the visual corpora, and calculated a concreteness score out of them for the word. We then experimented within the related task of metaphor detection, and showed comparable results to previous works. Our method and algorithm, though relatively simple and intuitive, give surprisingly good (comparable) results, while not requiring any multimodal processing at all.

Acknowledgments

We would like to thank the three anonymous reviewers for their suggestions and comments.
This work was supported by Contract W911NF-15-C-0238 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO).

References

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy*. http://aclweb.org/anthology/E/E06/E06-1042.pdf.

Max Black. 1979. More about metaphor.[in] a. ortony (ed.), metaphor and thought.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46(3):904–911.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33(4):497–505. https://doi.org/10.1080/14640748108400805.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25*

June 2009, Miami, Florida, USA. pages 248–255. https://doi.org/10.1109/CVPRW.2009.5206848.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, III Hal Daumé, Alexander C. Berg, and Tamara L. Berg. 2012. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL HLT '12, pages 762–772. http://dl.acm.org/citation.cfm?id=2382029.2382153.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao (Kenneth) Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 207–213. http://aclweb.org/anthology/D/D15/D15-1021.pdf.

W Nelson Francis and Henry Kucera. 1964. Brown corpus. *Department of Linguistics, Brown University, Providence, Rhode Island* 1. http://icame.uib.no/brown/bcm.html.

Ray Jackendoff. 1983. *Semantics and cognition*. MIT Press.

Maximilian Köper and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. *SENSE 2017* page 24.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *CoRR* abs/1602.07332. http://arxiv.org/abs/1602.07332.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago press.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. pages 740–755. https://doi.org/10.1007/978-3-319-10602-1$_4$8.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography* 3(4):235–244.

George A Miller and Philip N Johnson-Laird. 1976. *Language and perception.*. Belknap Press.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*. pages 1143–1151. http://papers.nips.cc/paper/4470-im2text-describing-images-using-1-million-captioned-photographs.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. pages 160–170. http://aclweb.org/anthology/N/N16/N16-1020.pdf.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. pages 248–258. http://aclweb.org/anthology/P/P14/P14-1024.pdf.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 680–690. http://www.aclweb.org/anthology/D11-1063.

David L Waltz. 1993. Relating images, concepts, and words. In *Intelligent Systems*, Springer, pages 21–38.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology* 3(1):1–191.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* 2:67–78. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229.