Efficient Multi-Modal Embeddings from Structured Data

Anita L. Verő
Department of Computer Science and Technology
University of Cambridge
alv34@cam.ac.uk

Ann Copestake
Department of Computer Science and Technology
University of Cambridge
aac10@cl.cam.ac.uk

Abstract
Multi-modal word semantics aims to enhance embeddings with perceptual input, assuming that human meaning representation is grounded in sensory experience. Most research focuses on evaluation involving direct visual input, however, visual grounding can contribute to linguistic applications as well. Another motivation for this paper is the growing need for more interpretable models and for evaluating model efficiency regarding size and performance. This work explores the impact of visual information for semantics when the evaluation involves no direct visual input, specifically semantic similarity and relatedness. We investigate a new embedding type in-between linguistic and visual modalities, based on the structured annotations of Visual Genome. We compare uni- and multi-modal models including structured, linguistic and image based representations. We measure the efficiency of each model with regard to data and model size, modality/data distribution and information gain. The analysis includes an interpretation of embedding structures. We found that this new embedding conveys complementary information for text based embeddings. It achieves comparable performance in an economic way, using orders of magnitude less resources than visual models.

1 Introduction
Multi-modal distributional semantics is inspired by the hypothesis that human semantic representation depends on sensorimotor experience (Symbol grounding problem (Harnad, 1990)), and the idea in cognitive science that meaning in the brain is represented both visually and verbally (Dual coding theory (Bucci, 1985)).

Most work focuses on evaluation tasks (and therefore on models which perform well on them) involving direct visual input, such as Visual Question Answering (Antol et al., 2015; Srivastava and Salakhutdinov, 2012; Kiros et al., 2014; Socher et al., 2014; Tsai et al., 2019; Lu et al., 2019; Su et al., 2019; Majumdar et al., 2020). In these cases the usefulness of visual input is not surprising. However, visual common sense can be useful in purely linguistic tasks as well (Bruni et al., 2014; Kiela and Bottou, 2014; Lazaridou et al., 2015). The analysis of the reasons has been addressed by fewer work (Kiela et al., 2014, 2016; Davis et al., 2019), however, more research is needed on the conditions and the type of information visual modalities contribute. Another motivation for this paper is the growing need for more interpretable models and to study model efficiency regarding size and performance (Bender et al., 2021).

This work explores the impact of visual information in meaning representations for semantic similarity and relatedness (Sec. 4). We investigate a new, visually structured, textual embedding (Sec. 3) in-between linguistic and visual modalities, based on the Visual Genome (VG) Scene-graphs as a structured data-source (Krishna et al., 2016).

Our main contribution is twofold. First, we measure the efficiency of each model with regard to data and model size and modality/data distribution (Sec. 5). Secondly, we analyse the interpretation of multi-modal models in terms of the Information...
Gain of modalities and embedding cluster structures (Sec. 6). Sec. 4 includes the description of evaluation. Experiments and discussion of the results are presented in Sec. 5 and 6. The code for all the experiments is openly available.

We found that our structured model conveys complementary information for text based embeddings. It achieves comparable or better performance in an economic way, by using orders of magnitude less resources than visual models. An important future application would be using automatically generated scenes graphs (Xu et al., 2020), mitigating the limitation of this approach, which is the manual labour required for creating VG. This would serve as a highly effective tool for low resource languages.

2 Related Work

Our VG Scene-graph based model (Sec. 3) is closely related to Kuzmenko and Herbelot, 2019; Herbelot, 2020. Even though our model is simpler (does not involve a truth theoretic model), our experiments support their claim about the efficiency of this data-source, while also showing its economic advantages in multi-modal models.

Our control method for data size and frequency ranges (Sec. 5) is inspired by the text-based case of Sahlgren and Lenci, 2016. Our results support theirs, while going far beyond in the amount of text data, besides adding other modalities.

3 Embedding Types

We define three different types of embeddings, which vary in training data type and learning algorithm. Linguistic Embeddings ($E_L$) are vector spaces which are learnt by an algorithm trained on large text data. Visual Embeddings ($E_V$) consist of vectors which have been trained on images associated to words. The learning algorithm is typically a Convolutional Neural Network (LeCun et al., 1989). In this paper we present results using Skip-Gram with Negative Sampling (SGNS) (Mikolov et al., 2013a,b) for $E_L$, due to its simplicity and efficiency, trained on a 2020 English Wikipedia dump. We ran a feedforward step of ResNet-152 (He et al., 2016) on Google Images which has been found the best performing in Kiela et al., 2016.

Structured Embeddings ($E_S$) are in-between visual and linguistic, trained on visually structured text. We exploit the dense VG image annotations represented as scene-graphs of objects, attributes, and pairwise relationships. We introduce a new method for generating contexts from VG Scene-graphs, which is the input for an SGNS. Context words are extracted by taking $r$ steps around a target node (word) in a breath first search manner (Fig. 1), where radius $r$ is a parameter. In this work we present results using $r = 3$, which we found best performing.

4 Evaluation

We evaluate on two standard semantic similarity and relatedness judgement datasets: MEN (Bruni et al., 2014) and SimLex-999 (Hill et al., 2015). Model performance is assessed through the Spearman $\rho_s$ rank correlation between the embedding similarity scores for a given pair of words, and human judgements.

For creating multi-modal representations we apply mid-level fusion, concatenating the L2-normalized representations (Bruni et al., 2014; Kiela et al., 2016), since it allows us to study the information preserved in the individual modalities.

5 Size and Performance

We perform two experiments. Firstly, we compare our models while restricting the training data size of $E_L$. The corpus is sampled randomly to subsets with increasing number of tokens. In a second experiment, we sub-sample the corpus by token frequency ranges into three equally large parts; HIGH, MEDIUM and LOW range. We sample from thevocabularies of $E_L$, $E_V$ and $E_S$ by the different frequency ranges coming from $E_L$’s training corpus. We present these results in App. B for we did not find any pattern.

Fig. 2 shows the effect of $E_L$ corpus size on the performance of uni-modal $E_L$ and the combined $E_L + E_S$ and $E_L + E_V$ on the embeddings’ common coverage subsets of MEN (Fig. 2a) and SimLex (Fig. 2b). $E_S$ and $E_V$ are constant since only $E_L$’s training data is varied. (Results on the full datasets are in App. B.) Axis $x$ represents the size of the training corpus (in the number of tokens). Error bars indicate variance after three runs.

Illustration: https://visualgenome.org/
Figure 2: Effect of $E_L$ training corpus (token) quantity on performance on the common coverage subsets of evaluation pairs (73% on MEN, 56% on SimLex).

| Model   | SGNS | SGNS | ResNet-152 |
|---------|------|------|------------|
| Tr. Data | Wikipedia | VG | ImageNet + |
| Size    | ~13G | ~9M | ~1.28M + |
| tokens  | tokens | 15,770 images (jpg) |
| Storage size | 14GB | ~1.8GB | ~140GB |
| Compr. size | ~5GB | ~0.2GB | ~140GB |

Table 1: Training data sizes.

of random down-sampling of the data. Tbl. 1 gives an account of the amount of training data each model requires. The last line shows the size after compression by Lempel-Ziv coding (LZ77). Since ImageNet images are already in jpg format LZ77 was not able to achieve any further compression.

The first striking result is that $E_S$ alone, with ~9M tokens, outperforms $E_L$, with ~1G tokens, on both evaluation tasks. Secondly, when combined with linguistic data, $E_S$ greatly outperforms $E_V$ on MEN and underperforms it on SimLex, however, their difference becomes marginal as text data increases. Importantly, $E_S$ achieves this result with orders of magnitude less data than required by $E_V$ (Tbl. 1). Moreover, ResNet-152 with ~6.8G parameters outputs a 1.7 times bigger model (4.8MB) than SGNS, used for $E_L$ and $E_S$ (2.8MB), consisting of 151,200 parameters.4

$E_S$ performs similarly to the FastText VG description model of Herbelot, 2020 on SimLex. The increase of $E_L$ performance is in line with Sahlgren and Lenci, 2016 until 2G tokens (they stopped at 1G), after which it plateaus.

Overall, we conclude that our structured visuo-linguistic embedding contributes to a linguistic model in a much more economic way than the

4On the common subset of their vocabularies of 1203 words.

Figure 3: Estimated Mutual Informations: $I(E_L,E_V)$ (red) and $I(E_L,E_S)$ (blue) for different corpus sizes and frequency ranges. Estimated by $I_{HSIC}$ algorithm with $\sigma =$ median.

6 Interpretation

We aim to peek under the hood of performance results and understand why $E_S$ contributes remarkably despite its small size.

6.1 Information Gain

In this experiment we aim to measure the information gain $E_S$ and $E_V$ each contribute when combined with $E_L$. By treating the embedding spaces as samples from multivariate distributions we formulate the question in the following way: Are two distributions of semantic spaces from different modalities independent from each other?

We measure the Mutual Informations: $I(E_L,E_S)$ and $I(E_L,E_V)$. Each embedding space represents samples from a multivariate distribution. Since $I$ is a special case of divergence, divergence estimators can be employed to estimate it, such as methods based on k-Nearest Neighbor distances ($I_{KNN}$) (Wang et al., 2009). A more robust approach in high dimensions is to introduce non-linearity using a kernel, when calculating the distances, such as the Hilbert-Schmidt Independence Criterion (HSIC) kernel algorithm (Gretton et al., 2005). We applied an open source Python implementation of these algorithms from the Information Theoretical Estimators Toolbox5 (Szabó, 2014).

In Fig 3 axis y shows $I(E_L,E_V)$ (red) and $I(E_L,E_S)$ (blue), where $I$ is the estimated Shannon mutual information using the $I_{HSIC}$ kernel method. Fig. 3a presents results across $E_L$ data sizes, Fig. 3b includes token frequency ranges (as in Sec. 5). With the exception of the 8M $E_L$ condition, $I(E_L,E_V)$ tend to be greater than $I(E_L,E_S)$, which suggests that $E_S$ is “more inde-
pendent” from the linguistic model $E_L$ than the image based $E_V$. This is surprising, as intuitively the text based $E_S$ is more linguistic than $E_V$, but suggests that $E_S$ provides proportionally more complementary information to $E_L$ than $E_V$.

To test robustness we ran $I_{HSIC}$ for a range of embedding dimensions, which we projected using PCA (Wold et al., 1987), and tried different $\sigma$ settings. These and the $I_{KNN}$ results support the main findings and are shown in App. C.

6.2 Qualitative Interpretation of Structure

In order to grasp how the concept structures of embeddings differ from each other we investigate their cluster structure. We perform K-means (MacQueen et al., 1967) and Agglomerative (Ward Jr, 1963) clustering on them. To measure the rate of clusterization, when the labels are not known, we use three standard metrics Davies–Bouldin Index (DBI), Calinski-Harabaz Index (CHI) and Silhouette Coefficient (SC).6 We also present T-SNE visualisations (Maaten and Hinton, 2008; Wattenberg et al., 2016) of the embeddings.7

Fig. 4a - 4c report on clustering metrics for increasing number of clusters. Fig. 4d - 4f show T-SNE plots of 20 cluster labels obtained by K-means clustering. On all three metrics, all models outperform the random embedding baseline of the same vocabulary8 and dimensionality. $E_V$ and $E_L + E_V$ consistently achieve the best scores, however, $E_S$ produces comparable DBI. On SC $E_S$ substantially outperforms the others, up until $\sim$20 clusters, when it drops dramatically. This suggests that $E_S$ has a much more cohesive structure of $\sim$20 clusters, but becomes incohesive if we try and break it into more. This may be an artefact of VG, as Krishna et al., 2016 reports an average of 17 description clusters on an image.9

On the 2D T-SNE projections, $E_V$ appears to have the most defined clusters. We labelled the 20 cluster case of all embeddings by hand for both K-means and Agglomerative clustering as control. We also generated labels based on the most common WordNet synsets (Miller, 1995) occurring in the cluster. We found that $E_S$ clusters represent more concrete concepts and are more visual than $E_L$ but less than $E_V$. Example clusters and a detailed analysis can be found in App. D.

7 Conclusion and Future Work

We found that our VG Scene-graph based model achieves comparable or better performance in an economic way, by using orders of magnitude less resources than visual models. When combined, it enriches our linguistic model with more divergent information than the image based one. Its clusters represent more concrete concepts, in-between visual and linguistic domains.

Investigating transformers, Bayesian MI estimators and other evaluations could be potential extensions of these studies. Applying automatically generated scenes graphs would mitigate the main limitation of this approach, which is the manual labour required for creating VG. This would serve

---

6 We used their Scikit-learn implementation https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

7 We applied the Tensorboard implementation: https://www.tensorflow.org/tensorboard. We present results for perplexity = 30 as we did not find much difference between the perplexity settings in their suggested range of 5 – 50.

8 The common subset of all embedding vocabularies of 1203 words.

9 Although they clustered averaged pre-trained word representations of region descriptions, not scene graphs, therefore, their results are not directly comparable to ours.
as a highly effective tool with important applications for low resource languages.

References
Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, NY, USA.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. J. Artif. Intell. Res.(JAIR), 49(2014):1–47.

Wilma Bucci. 1985. Dual coding: A cognitive model for psychoanalytic research. Journal of the American Psychoanalytic Association, 33(3):571–607.

Christopher Davis, Luana Bulat, Anita Lilla Verő, and Ekaterina Shutova. 2019. Deconstructing multimodality: visual properties and visual context in human semantic processing. In Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019), pages 118–124.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In International conference on algorithmic learning theory, pages 63–77. Springer.

Stevan Harnad. 1990. The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1-3):335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Aurélie Herbelot. 2020. Re-solve it: simulating the acquisition of core semantic competences from small data. In Proceedings of the 24th Conference on Computational Natural Language Learning, pages 344–354.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. Association for Computational Linguistics.

Douwe Kiela and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In Proceedings of EMNLP, pages 36–45.

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 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), volume 2, pages 835–841.

Douwe Kiela, Anita L. Verő, and Stephen Clark. 2016. Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-16).

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In International Conference on Machine Learning, pages 595–603.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In Proceedings of NIPS, pages 1106–1114.

Elizaveta Kuzmenko and Aurélie Herbelot. 2019. Distributional semantics in the real world: building word vector representations from a truth-theoretic model. In Proceedings of the 13th International Conference on Computational Semantics-Short Papers, pages 16–23.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 153–163.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4):541–551.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. Journal of machine learning research, 9(Nov):2579–2605.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA.
Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. 2020. Improving vision-and-language navigation with image-text pairs from the web. In European Conference on Computer Vision, pages 259–274. Springer.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems, pages 3111–3119.

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In EMNLP 2016.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (ICLR), 2015.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics, 2:207–218.

Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In Advances in neural information processing systems, pages 2222–2230.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. In International Conference on Learning Representations.

Zoltán Szabó. 2014. Information theoretical estimators toolbox. The Journal of Machine Learning Research, 15(1):283–287.

Yao-Hung Hubert Tsai, Shaqiee Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Annual Meeting of the Association for Computational Linguistics.
A Comparison of Various Models

Linguistic models are three FastText (Mikolov et al., 2018): wiki-news-300d-1M, wiki-news-300d-1M-subword, crawl-300d-2M and a Skip-Gram model (SGNS) trained on a 2013 Wikipedia dump.10

Visual models compared: AlexNet (Krizhevsky et al., 2012), ResNet-152 architecture (He et al., 2016), VGG (Simonyan and Zisserman, 2014), trained on Google Images, and two models trained on VG internal object bounding box images or on the whole images, similarly to Davis et al., 2019.

Results are shown in Tbl. 2-5.

B Size, Frequency and Performance

Fig. 5 presents performance results controlled for $E_L$ data size and token frequency ranges on the full evaluation datasets.

Table 2: Spearman correlation on the full MEN dataset. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

| Models                        | Spearman | P-value | Coverage |
|-------------------------------|----------|---------|----------|
| wikinews                      | 0.7938   | 0       | 3000     |
| wikinews_sub                  | 0.8032   | 0       | 3000     |
| crawl                         | **0.8455** | 0       | 3000     |
| SGNS13                        | 0.6826   | 0       | 3000     |
| Google AlexNet                | 0.5025   | 0       | 3000     |
| Google VGG                    | 0.5118   | 0       | 3000     |
| VG-internal                   | 0.3737   | 0       | 2784     |
| VG-whole                      | 0.409    | 0       | 2784     |
| Google ResNet-152             | 0.4688   | 0       | 3000     |
| VG SceneGraph                 | 0.4238   | 0       | 2574     |
| wikinews+Google AlexNet       | 0.5033   | 0       | 3000     |
| wikinews+VG SceneGraph        | 0.6489   | 0       | 2574     |
| wikinews+Google VGG           | 0.5129   | 0       | 3000     |
| wikinews+VG-internal          | 0.3754   | 0       | 2784     |
| wikinews+VG-whole             | 0.4114   | 0       | 2784     |
| wikinews+VG-whole+VG-internal | 0.4768   | 0       | 3000     |
| wikinews+VG-whole+VG-whole    | 0.4117   | 0       | 2784     |
| wikinews+Google ResNet-152    | 0.5026   | 0       | 3000     |
| wikinews+VG SceneGraph        | 0.4477   | 0       | 2574     |
| wikinews+VG-internal          | 0.5118   | 0       | 3000     |
| wikinews+VG-whole             | 0.3738   | 0       | 2784     |
| wikinews+VG-whole+VG-internal | 0.4091   | 0       | 2784     |
| wikinews+VG-whole+VG-whole    | 0.4091   | 0       | 2784     |
| crawl+Google AlexNet          | 0.5068   | 0       | 3000     |
| crawl+VG SceneGraph           | 0.8114   | 0       | 2574     |
| crawl+Google VGG              | 0.5176   | 0       | 3000     |
| crawl+VG-internal             | 0.3825   | 0       | 2784     |
| crawl+VG-whole                | 0.4213   | 0       | 2784     |
| crawl+Google ResNet-152       | 0.5059   | 0       | 3000     |
| SGNS13+Google AlexNet         | 0.5046   | 0       | 3000     |
| SGNS13+VG SceneGraph          | **0.6951** | 0       | 2574     |
| SGNS13+Google VGG             | 0.5145   | 0       | 3000     |
| SGNS13+VG-internal            | 0.3774   | 0       | 2784     |
| SGNS13+VG-whole               | 0.4148   | 0       | 2784     |
| SGNS13+Google ResNet-152      | 0.4846   | 0       | 3000     |

Table 3: Spearman correlation on the full SimLex dataset. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

| Models                        | Spearman | P-value | Coverage |
|-------------------------------|----------|---------|----------|
| wikinews                      | 0.45     | 0       | 999      |
| wikinews_sub                  | 0.4409   | 0       | 999      |
| crawl                         | **0.5031** | 0       | 999      |
| SGNS13                        | 0.31     | 0       | 999      |
| Google AlexNet                | 0.3401   | 0       | 999      |
| Google VGG                    | 0.342    | 0       | 999      |
| VG-internal                   | 0.307    | 0.0016  | 103      |
| VG-whole                      | 0.1853   | 0.061   | 103      |
| Google ResNet-152             | 0.3465   | 0       | 999      |
| VG SceneGraph                 | 0.2587   | 0       | 593      |
| wikinews+Google AlexNet       | 0.3406   | 0       | 999      |
| wikinews+VG SceneGraph        | 0.3451   | 0       | 999      |
| wikinews+Google VGG           | 0.3426   | 0       | 999      |
| wikinews+VG-internal          | 0.3072   | 0.0016  | 103      |
| wikinews+VG-whole             | 0.1846   | 0.062   | 103      |
| wikinews+Google ResNet-152    | 0.3401   | 0       | 999      |
| wikinews+VG SceneGraph        | 0.2951   | 0       | 593      |
| wikinews+VG-internal          | 0.3516   | 0       | 999      |
| wikinews+VG-whole             | 0.2951   | 0       | 593      |
| wikinews+VG-internal          | 0.307    | 0.0016  | 103      |
| wikinews+VG-whole             | 0.1848   | 0.0616  | 103      |
| wikinews+Google ResNet-152    | 0.3471   | 0       | 999      |
| crawl+Google AlexNet          | 0.3427   | 0       | 999      |
| crawl+VG SceneGraph           | 0.4438   | 0       | 993      |
| crawl+Google VGG              | 0.3448   | 0       | 999      |
| crawl+VG-internal             | 0.3122   | 0.0013  | 103      |
| crawl+VG-whole                | 0.1891   | 0.0557  | 103      |
| crawl+Google ResNet-152       | 0.3687   | 0       | 999      |
| SGNS13+Google AlexNet         | **0.341** | 0       | 999      |
| SGNS13+VG SceneGraph          | 0.2914   | 0       | 999      |
| SGNS13+Google VGG             | **0.3428** | 0       | 999      |
| SGNS13+VG-internal            | 0.3076   | 0.0016  | 103      |
| SGNS13+VG-whole               | 0.1832   | 0.064   | 103      |
| SGNS13+Google ResNet-152      | **0.3539** | 0       | 999      |

10https://fasttext.cc/docs/en/english-vectors.html
C Information Gain

Results for different projected dimensions $d$ and $\sigma$ settings are shown in Fig. 6 and 7.

D Cluster Structure

Example clusters on the embeddings’ common vocabulary can be found in Tbl. 6-8. Tbl. 9 includes an analysis on Jaccard similarities among different clusterings using K-means and Agglomerative clustering. All clusters with automatic WordNet, as well as the authors’ annotation will be publicly available online.

Main observations:

- Wikipedia based $E_L$ has more clusters with abstract topics, such as verbs, activities and communication.
- $E_S$ has more concrete clusters e.g., train, vehicles, building structures, containers or furnishing.

Table 4: Spearman correlation on the common subset of the MEN dataset. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

| Models           | Spearman | P-value | Coverage |
|------------------|----------|---------|----------|
| wikinews         | 0.7989   | 0       | 2481     |
| wikinews_sub     | 0.8035   | 0       | 2481     |
| SGNS13           | 0.673    | 0       | 2481     |
| Google AlexNet   | 0.4751   | 0       | 2481     |
| Google VGG       | 0.5146   | 0       | 2481     |
| VG-internal      | 0.3751   | 0       | 2481     |
| VG-whole         | 0.4057   | 0       | 2481     |
| Google ResNet-152| 0.4742   | 0       | 2481     |
| VG-SceneGraph    | 0.4421   | 0       | 2481     |
| wikinews+Google AlexNet | 0.1559 | 0 | 2481 |
| wikinews+VG SceneGraph | 0.6561 | 0 | 2481 |
| wikinews+Google VGG | 0.5158 | 0 | 2481 |
| wikinews+VG-internal | 0.3769 | 0 | 2481 |
| wikinews+VG-whole | 0.4083 | 0 | 2481 |
| wikinews+Google ResNet-152 | 0.482 | 0 | 2481 |
| wikinews_sub+VG SceneGraph | 0.5151 | 0 | 2481 |
| wikinews_sub+Google AlexNet | 0.4622 | 0 | 2481 |
| wikinews_sub+Google VGG | 0.5147 | 0 | 2481 |
| wikinews_sub+VG-internal | 0.3752 | 0 | 2481 |
| wikinews_sub+VG-whole | 0.4058 | 0 | 2481 |
| wikinews_sub+Google ResNet-152 | 0.4747 | 0 | 2481 |
| crawl+Google AlexNet | 0.5191 | 0 | 2481 |
| crawl+VG SceneGraph | 0.8144 | 0 | 2481 |
| crawl+Google VGG | 0.5204 | 0 | 2481 |
| crawl+VG-internal | 0.3841 | 0 | 2481 |
| crawl+VG-whole | 0.4187 | 0 | 2481 |
| crawl+Google ResNet-152 | 0.5015 | 0 | 2481 |
| SGNS13+Google AlexNet | 0.5149 | 0 | 2481 |
| SGNS13+VG SceneGraph | 0.609 | 0 | 2481 |
| SGNS13+Google VGG | 0.5172 | 0 | 2481 |
| SGNS13+VG-internal | 0.3789 | 0 | 2481 |
| SGNS13+VG-whole | 0.412 | 0 | 2481 |
| SGNS13+Google ResNet-152 | 0.4887 | 0 | 2481 |

Table 5: Spearman correlation on the common subset of the SimLex dataset. The table sections contain linguistic, visual and multi-modal embeddings in this order. Red colour signifies the best performance, blue means that the multi-modal embedding outperformed the corresponding uni-modal ones.

- Image based $E_V$ includes more clusters related to the outdoors, such as “travel”, “transportation”, “landscape” and “vacation” and on appearance, such as “colours & materials”.
- Concepts that all three embeddings capture: “food”, “colours”, “plants”, “animals” and “body parts”.

| Models           | Spearman | P-value | Coverage |
|------------------|----------|---------|----------|
| wikinews         | 0.2818   | 0.0039  | 103   |
| wikinews_sub     | 0.2454   | 0.0125  | 103   |
| crawl            | 0.3748   | 0.0001  | 103   |
| SGNS13           | 0.1146   | 0.2491  | 103   |
| Google AlexNet   | 0.5539   | 0       | 103   |
| Google VGG       | 0.5261   | 0       | 103   |
| VG-internal      | 0.307    | 0.0016  | 103   |
| VG-whole         | 0.1853   | 0.061   | 103   |
| Google ResNet-152| 0.5015   | 0       | 103   |
| VG-SceneGraph    | 0.2992   | 0.0021  | 103   |
| wikinews+Google AlexNet | 0.5539 | 0 | 103 |
| wikinews+VG SceneGraph | 0.2919 | 0.0028 | 103 |
| wikinews+Google VGG | 0.5261 | 0 | 103 |
| wikinews+VG-internal | 0.3072 | 0.0016 | 103 |
| wikinews+VG-whole | 0.1846 | 0.062 | 103 |
| wikinews+Google ResNet-152 | 0.503 | 0 | 103 |
| wikinews_sub+VG SceneGraph | 0.2979 | 0.0022 | 103 |
| wikinews_sub+Google AlexNet | 0.5539 | 0 | 103 |
| wikinews_sub+VG-internal | 0.3072 | 0.0016 | 103 |
| wikinews_sub+VG-whole | 0.1846 | 0.061 | 103 |
| wikinews_sub+Google ResNet-152 | 0.5006 | 0 | 103 |
| crawl+Google AlexNet | 0.5518 | 0 | 103 |
| crawl+VG SceneGraph | 0.341 | 0.0004 | 103 |
| crawl+Google VGG | 0.5247 | 0 | 103 |
| crawl+VG-internal | 0.3122 | 0.0013 | 103 |
| crawl+VG-whole | 0.1891 | 0.0557 | 103 |
| crawl+Google ResNet-152 | 0.4893 | 0 | 103 |
| SGNS13+Google AlexNet | 0.5517 | 0 | 103 |
| SGNS13+VG SceneGraph | 0.25 | 0.0109 | 103 |
| SGNS13+Google VGG | 0.5264 | 0 | 103 |
| SGNS13+VG-internal | 0.3076 | 0.0016 | 103 |
| SGNS13+VG-whole | 0.1832 | 0.064 | 103 |
| SGNS13+Google ResNet-152 | 0.4922 | 0 | 103 |
| WordNet lb. | Own lb. | Members |
|------------|---------|---------|
| artifact   | train   | railway, railroad, subway, curve, tunnel, run, shelter, train, station, tram, highway, track, rail, way, engine, stop, gate, bridge, smoke |
| line       | plants  | bloom, foliage, grave, dead, vine, blossom, ivy, pod, cactus, tree, moss, root, leave, limb, forest, bush, plant, lily, branch, weed, leaf, vein, sunshine, log, fence, flower, sunlight, wood, palm, bench, sun |
| whole      | rooms   | classroom, hallway, hall, closet, bedroom, room, bathroom, garage, office, cafe, museum, doorway, kitchen, shop, restaurant, store, mannequin, stadium, market, ceiling, corner |
| plant      | birds   | hummingbird, gull, peacock, hawk, pelican, crow, parrot, seagull, wing, swan, pigeon, owl, goose, flamingo, nest, eagle, tail, bird, silhouette, duck, chest, body, ledge, giraffe, zebra |
| organ      | vertebrates | animals |
| vascular   | food    | drizzle, nuts, herb, beef, flour, season, cereal, cherry, breakfast, sugar, steak, bacon, burger, butter, rice, meat, meal, sauce, dinner, pie, raspberry, lunch, sushi, bean, mustard, pepper, seed, salt, soup, cheese, tomato, hot, berry, potato, dessert, strawberry, salad, cardboard, food, bone, lemon, burn, frost, chocolate, bread, turkey, sandwich, spoon, pizza, chicken, shell, candy, peel, cooking, bubble, knife, fruit, fish, donut, cake, apple, ice, banana, orange |
| plant      | produce | butter, cheese, bread, chicken, soup, sauce, dessert, beef, salad, meat, cake, steak, tomato, potato, pizza, flour, milk, meal, vinegar, bacon, pie, cooking sushi, sandwich, breakfast, burger, menu |
| plant      | vascular | flower, flowers, tree, blossom, dandelion, foliage, fruit, weed, cactus, lily, bloom, shade, leaf, grass, sunflower, poppy, vine, plant, garden, iris, grow, daisy, oak, bulb, rust, herb, moss, tulip, palm, maple, root, tall, bush, seed, family |
| organ      | body parts | skin, spine, neck, bone, chest, throat, shoulder, wrist, stomach, ear, jaw, cheek, lips, nose, eyes, eye, limb, toe, belly, skull, abdomen, finger, teeth, elbow, cord, whiskers, knee, thumb, tooth, muscle, ankle, tail, paws, lip, brain, flesh, leg, body, calf, heart, blood, tongue, brow, pain, tear, blade, mouth, liver, gut, arm, marrow, curled, canine, feathers, foot, vein, hip, cancer |
| part       | body parts | skin, spine, neck, bone, chest, throat, shoulder, wrist, stomach, ear, jaw, cheek, lips, nose, eyes, eye, limb, toe, belly, skull, abdomen, finger, teeth, elbow, cord, whiskers, knee, thumb, tooth, muscle, ankle, tail, paws, lip, brain, flesh, leg, body, calf, heart, blood, tongue, brow, pain, tear, blade, mouth, liver, gut, arm, marrow, curled, canine, feathers, foot, vein, hip, cancer |
| artifact   | verbs   | bring, get, come, want, go, keep, take, know, find, say, give, make, understand, put, listen, enjoy, feel, leave, think, learn, imagine, gather, believe, fail, arrange, add, lose, create, way, hear, send, meet, collect, carry, avoid, buy, remain, allow, appear, might, enter, arrive, seem, entertain, break, steal, receive, stop, stand, build, locked, compare, retain, sell, handle, danger, eat, wander, face, unhappy, protect, please, pray, become, walk, expand, travel, plenty, greet, inspect, comfort, huge, possess, dominate, attach, roam, participate, speak, step, drawn, construct, replace, divide, great, living |

Table 6: Examples from the 20 K-means clusters in $E_s$.

| WordNet lb. | Own lb. | Members |
|------------|---------|---------|
| food       | food    | butter, cheese, bread, chicken, soup, sauce, dessert, beef, salad, meat, cake, steak, tomato, potato, pizza, flour, milk, meal, vinegar, bacon, pie, cooking sushi, sandwich, breakfast, burger, menu |
| nutriment  | plants  | flower, flowers, tree, blossom, dandelion, foliage, fruit, weed, cactus, lily, bloom, shade, leaf, grass, sunflower, poppy, vine, plant, garden, iris, grow, daisy, oak, bulb, rust, herb, moss, tulip, palm, maple, root, tall, bush, seed, family |
| foodstuff  | classical architecture | tower, building, marble, staircase, fountain, doorway, roof, chapel, steeple porch, ceiling, mural, glass, wall, brick, statue, stone, arch, monument, dome, window, gravestone, sculpture, aisle, tiles, gate, interior, painted, decoration, concrete, church, graveyard, cathedral, curtain, painting, palace, clock, grave, portrait, choir, architecture, pyramid, memorial, square, castle, skyscraper, museum, cemetery, temple, organ |
| vascular   | body parts | skin, spine, neck, bone, chest, throat, shoulder, wrist, stomach, ear, jaw, cheek, lips, nose, eyes, eye, limb, toe, belly, skull, abdomen, finger, teeth, elbow, cord, whiskers, knee, thumb, tooth, muscle, ankle, tail, paws, lip, brain, flesh, leg, body, calf, heart, blood, tongue, brow, pain, tear, blade, mouth, liver, gut, arm, marrow, curled, canine, feathers, foot, vein, hip, cancer |
| plant      | part    | skin, spine, neck, bone, chest, throat, shoulder, wrist, stomach, ear, jaw, cheek, lips, nose, eyes, eye, limb, toe, belly, skull, abdomen, finger, teeth, elbow, cord, whiskers, knee, thumb, tooth, muscle, ankle, tail, paws, lip, brain, flesh, leg, body, calf, heart, blood, tongue, brow, pain, tear, blade, mouth, liver, gut, arm, marrow, curled, canine, feathers, foot, vein, hip, cancer |
| organ      | artifact | skin, spine, neck, bone, chest, throat, shoulder, wrist, stomach, ear, jaw, cheek, lips, nose, eyes, eye, limb, toe, belly, skull, abdomen, finger, teeth, elbow, cord, whiskers, knee, thumb, tooth, muscle, ankle, tail, paws, lip, brain, flesh, leg, body, calf, heart, blood, tongue, brow, pain, tear, blade, mouth, liver, gut, arm, marrow, curled, canine, feathers, foot, vein, hip, cancer |
| part       | verbs   | bring, get, come, want, go, keep, take, know, find, say, give, make, understand, put, listen, enjoy, feel, leave, think, learn, imagine, gather, believe, fail, arrange, add, lose, create, way, hear, send, meet, collect, carry, avoid, buy, remain, allow, appear, might, enter, arrive, seem, entertain, break, steal, receive, stop, stand, build, locked, compare, retain, sell, handle, danger, eat, wander, face, unhappy, protect, please, pray, become, walk, expand, travel, plenty, greet, inspect, comfort, huge, possess, dominate, attach, roam, participate, speak, step, drawn, construct, replace, divide, great, living |

Table 7: Examples from the 20 K-means clusters in $E_L$. 
| WordNet lb.   | Own lb. | Members                                      |
|--------------|---------|----------------------------------------------|
| bird         | birds   | seagull, gull, goose, duck, pelican, swan, mallard, stork, eagle, flamingo |
| aquatic bird |         |                                              |
| seabird      |         |                                              |
| vascular plant | plants | weed, bunch, maple, cancer, iris, poppy, dandelion, leave, flower, rose, foliage, grow, plant, cactus, spring, tulip, ivy, palm, lily, leaf, daisy, tree, root, wheat, wool, raspberry, tobacco, flowers, blossom, butterfly, sunflower, cotton, herb, violet, oak, moss, strawberry, nest, dew, berry, rice, branch, coal |
| plant grow   |         |                                              |
| food         | food    | sushi, meal, sandwich, pie, breakfast, lunch, food, supper, flour, cereal, sweet, dessert, m'dinner, subway, diet, cake, date, steak, sauce, bread, copper, nuts, bacon, cooking, beef, meat, bakery, knitting, eat, potato, salad, donut, pizza, burger, coffee, soup, bean, cheese, vitamin, fruit, pumpkin, rock, marrow, market, timber |
| nutrient     |         |                                              |
| substance    |         |                                              |
| artifact     | colours, | texture, fabric, cloth, metal, rain, concrete, paper, suds, rough, words, stone, wall, square, dense, leather, quote, wood, frost, mud, noise, text, purple, carpet, blue, tiles, dirt, droplets, red, sand, fog, formula, mist, pattern, handwriting, green, straw, linen, asphalt, stripes, crowd, marble, yellow, black, brown, grey, grass, white |
| change cover |         |                                              |
| travel       | vacation| island, view, reflection, harbor, nice, side, sea, summer, tropical, pollution, port, aircraft, pier, travel, surfers, journey, sunny, coast, flying, morning, ocean, seashore, horizon, mare, holiday, lake, surf, shore, vacation, bay, airport, cliff, sunlight, air, river, storm, ship, fishing, beach, desert, harbour, puddle, flight, sailing, evening, sunrise, skyline, vessel, lighthouse, dawn, sunset, rocket, mountain, whale, underwater, boat, swimming, swim, plane, dusk, jet, cloud, sky, airplane, ski |
| change object|         |                                              |

Table 8: Examples from the 20 K-means clusters in $E_V$.

| Cross-modalities | $>0.2$ | $>0.3$ | $>0.4$ | Max |
|------------------|--------|--------|--------|-----|
| Clust.           | K-m Agg| K-m Agg| K-m Agg| K-m Agg |
| $E_S - E_L$      | 12     | 6      | 1      | 0    |
|                  | 0.358  | 0.347  |        |      |
| $E_S - E_V$      | 8      | 9      | 2      | 4    |
|                  | 0.363  | 0.5    |        |      |
| $E_L - E_V$      | 9      | 9      | 5      | 5    |
|                  | 0.479  | 0.467  |        |      |

| K-means – Agglomerative | $>0.2$ | $>0.3$ | $>0.4$ | $>0.5$ | $>0.6$ | $>0.7$ | Max |
|-------------------------|--------|--------|--------|--------|--------|--------|-----|
| Clust.                  | K-m Agg| K-m Agg| K-m Agg| K-m Agg| K-m Agg|
| $E_L - E_L$             | 18     | 14     | 9      | 4      | 2      | 1      | 0.79|
| $E_S - E_S$             | 16     | 15     | 13     | 9      | 5      | 1      | 0.729|
| $E_V - E_V$             | 23     | 16     | 13     | 8      | 4      | 0      | 0.644|

Table 9: Number of cluster pairs out of $20^2$ with Jaccard similarities above thresholds of [0.2, . . . , 0.7]. Last column shows the maximum similarity.
Figure 6: Estimated Mutual Informations: $I(E_L, E_V)$ (red) and $I(E_L, E_S)$ (blue) for different $E_L$ corpus sizes.

Figure 7: Estimated Mutual Informations: $I(E_L, E_V)$ (red) and $I(E_L, E_S)$ (blue) for different word frequency ranges.