Modelling semantics by integrating linguistic, visual and affective information

Armand S. Rotaru (armand.rotaru.14@ucl.ac.uk) **Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)**
Faculty of Brain Sciences, University College London,
WC1H 0DS, London, United Kingdom

Abstract

A number of recent models of semantics combine linguistic information, derived from text corpora, and visual information, derived from image collections, demonstrating that the resulting multimodal models are better than either of their unimodal counterparts, in accounting for behavioural data. However, first, while linguistic models have been extensively tested for their fit to behavioural semantic ratings, this is not the case for visual models which are also far more limited in their coverage. More broadly, empirical work on semantic processing has shown that emotion also plays an important role especially for abstract concepts, however, models integrating emotion along with linguistic and visual information are lacking. Here, we first improve on visual representations by choosing a visual model that best fit semantic data and extending its coverage. Crucially then, we assess whether adding affective representations (obtained from a neural network model designed to predict emojis from co-occurring text) improves model's ability to fit semantic similarity/relatedness judgements from a purely linguistic and linguistic-visual model. We find that adding both visual and affective representations improve performance, with visual representations providing an improvement especially for more concrete words and affective representations improving especially fit for more abstract words.

Keywords: language; vision; emotion; distributional models; multimodal models; similarity/relatedness; concreteness.

Introduction

Despite the success of distributional, linguistic models in accounting for behavioural effects in a variety of semantic tasks, all these models suffer from the *symbol grounding problem* (Harnad, 1990). As a solution to this problem, embodied theories of semantics (e.g., Glenberg, Graesser, & de Vega, 2008) have argued that the sensory-motor representations generated by our experiences with the world play an important role in determining word meaning. Recent computational models of semantics reconcile distributional and embodied theories, by combining linguistic and perceptual (i.e., visual) representations. The fact that language and vision provide complementary sources of information is best illustrated by the finding that multimodal, linguistic-visual models outperform both purely linguistic and purely visual models, in a wide range of tasks (see Bruni, Tran, & Baroni, 2011; 2014; Kiela, Verő, & Clark, 2016).

However, empirical work has shown that semantic representations are not only grounded in sensory-motor experience but also in emotion. A vast literature supports the finding that emotion plays a significant and pervasive role in human cognition (for a review, see Dolan, 2002). Emotion is an important factor in memory (Blaney, 1986; Eich, Macaulay, & Ryan, 1994), and in processing words (e.g., Kousta, Vinson, & Vigliocco, 2009). Kousta et al. (2011) found that a much larger number of abstract than concrete concepts are valenced (have positive or negative emotional associations) and by virtue of being valenced, they are processed faster than neutral matched words. Vigliocco et al. (2014) further showed that because of their greater affective associations, abstract words processing engages the limbic emotional system and Ponari, Norbury, and Vigliocco (2018) showed that emotionally valenced words are learnt earlier and better recognized by children up to 9 years of age. Within a general embodiment framework, the hypothesis is that semantic representations do not only embed sensorimotor properties but also emotional properties. Emotional properties may be especially important for abstract concepts (e.g., *religion*, *society, idea*), however, emotional associations are not limited to abstract words and therefore, we argue, they play a general role in semantic representation.

While many models have integrated linguistic and visual information, only one previous study has considered emotional information along with visual and linguistic information (De Deyne, Navarro, Collell, & Perfors, 2018). De Deyne et al. examined the change in performance for distributional models of semantics, when adding visual and emotional information. They tested the assumption that external language models (i.e., distributional models, trained on word corpora) are relatively poor at representing visual and affective information, in comparison to internal language models (i.e., models based on free association norms). They found that adding visual and emotional information led to little or no improvement for internal language models, but a moderate positive effect for external language models. Here, we develop a quite different multimodal model of semantics that incorporates linguistic, visual and emotional information from corpora of text, images and emoticons, and test the multimodal model against existing datasets of ratings of

semantic similarity/relatedness of words. We use a state-of-the-art emotion model (DeepMoji) and we improve the coverage of the visual model we use. While state-of-the art distributional language models (Pereira et al. 2016) have large coverage of words and have been widely tested for their ability to fit human semantic similarity/relatedness data, this is not the case for visual models. Thus, before being able to develop models that embed linguistic, visual and emotional information, we extend the coverage of existing visual models and carry out their evaluation in order to decide which one to use for our multimodal models. We expect that the integrated model will outperform a purely linguistic, as well as models that combine linguistic-visual and linguistic-emotional information. In addition, we expect that adding visual or emotional representations will especially be beneficial for more concrete concepts whereas emotional information will especially be beneficial for more abstract concepts, in line with the empirical evidence reviewed above (and with initial findings from De Deyne et al, 2018).

Methods

Datasets of behavioural data

We use four datasets of similarity/relatedness ratings to carry out evaluation of the models. The datasets are: SimLex999 (999 pairs of nouns, verbs, and adjectives; Hill, Reichart, & Korhonen, 2015), SimVerb3500 (3500 pairs of verbs; Gerz et al., 2016), MEN (3000 pairs of nouns, verbs, and adjectives; Bruni, Tran, & Baroni, 2014), and SL (7576 pairs of nouns; Silberer & Lapata, 2014). We chose these norms mainly because they are some of the largest datasets currently available, but also because the word pairs they contain cover are very diverse in terms of concreteness and valence, as well as parts of speech. In terms of word pair concreteness, SimLex999 (M = 3.62, SD = 1.07) and SimVerb3500 (M = 3.1, SD = 0.7) cover a broad range of values, whereas MEN (M = 4.4, SD = 0.49) and SL (M = 4.83, SD = 0.14) consist predominantly of concrete words.

Model choice

Language Model. Our language model of choice is GloVe (Pennington, Socher, & Manning, 2014), trained on a corpus of 6 billion words, using 300-dimensional representations. GloVe has been shown to have a performance better than, or equal to, several other state-of-the-art distributional models (Pereira, Gershman, Ritter, & Botvinick, 2016), which makes it one of the best linguistic models available.

Emotion Model. The emotion model that we use is DeepMoji (Felbo et al., 2017), trained on 1.2 billion tweets. This model has been shown to obtain state-of-the-art performance in tasks involving emotion and sentiment analysis, as well as sarcasm detection. DeepMoji is similar to a number of recent approaches, which employ emotional expressions co-occurring with text fragments, such as positive/negative emoticons (Deriu et al., 2016), hashtags (e.g., #anger, #joy; Mohammad, 2012), or mood tags (Mishne, 2005). This model is very different from the one by De Deyne et al. (2018), which was constructed by concatenating valence, arousal, and potency ratings, for men and women separately (i.e., 6 dimensions), from the study by Warriner, Kuperman, and Brysbaert (2013), with valence, arousal, and dominance ratings, from the study by Mohammad (2018). DeepMoji provides better represetations for our purposes than ratings because firstly, a model trained over a corpus of tweets, rather than subjective ratings, makes the emotion model more comparable to the linguistic and visual models, both trained over corpora. Secondly, DeepMoji covers 50,000 words, whereas the combined affective norms cover less than 14,000 words. Finally, the model operates with 256-dimensional vector representations, and is trained to predict the occurrence of 64 types of emoticons, and thus it is able to represent complex patterns of word similarity, driven by richer emotional information than that captured by subjective norms.

Visual Model. To select the best model, we compared five models, based on their performance in predicting subjective similarity/relatedness ratings. The first model (K&B) is the convolutional model employed by Kiela and Bottou (2014; 6144 dimensions), trained on the ESP Game dataset (Von Ahn & Dabbish, 2004), using the mean of the feature vectors per each word. The second, third, and fourth models are AlexNet (Krizhevsky, Sutskever, & Hinton, 2012; 4096 dimensions), GoogLeNet (Szegedy et al., 2015; 1024 dimensions), and VGG-19 (Simonyan & Zisserman, 2014; 4096 dimensions), trained on images obtained from Google Image Search, following the approach used by Kiela, Verő, and Clark (2016). The fifth model uses SIFT descriptors (Lowe, 2004), computed over the NUS-WIDE dataset (Chua et al., 2009; 500 dimensions). The models were tested on similarity/relatedness ratings for 7611 word pairs, covered by all models and obtained by merging the four sets of ratings. Before merging, the scores in each set were linearly rescaled to fall in the interval [0,1], to make them comparable across datasets. The performance of the models was evaluated using the Spearman correlation between the cosine similarity of the model representations, and the similarity/relatedness ratings from the norms. The results are shown in Fig. 1.

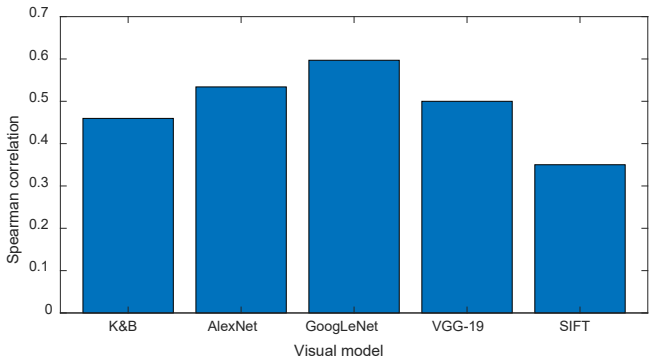


Figure 1. Spearman correlations between model cosine similarities and subjective similarity/relatedness ratings.

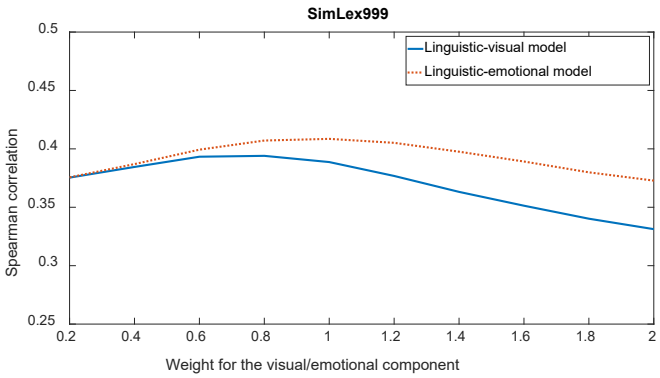
All the correlations are significant[1] ($p < .001$), suggesting that model-based similarities are reliable predictors of subjective similarity/relatedness ratings. Since we want to find the best model, we apply the Fisher Z-Transformation and then run two-tailed Z-tests for all the 10 possible pairings of models. All the differences are significant ($p < .004$), and they reveal that GoogLeNet has the highest performance, followed by Alexnet, VGG-19, K&B, and SIFT. Thus, we use GoogLeNet.

Results

We tested whether linguistic-visual and linguistic-emotional models are indeed better than a purely linguistic one, as well as whether it is the case that linguistic-visual-emotional models are better than linguistic-visual, linguistic-emotional and purely linguistic ones. We also examined whether the models behave differently for concrete and abstract word pairs.

Linguistic-visual and linguistic-emotional models vs purely linguistic model.

To evaluate the change in goodness of fit associated with adding a visual component to the purely linguistic model, we began by normalizing the linguistic and the visual representations to unit length. Next, we concatenated the linguistic representations with the visual ones, assigning a weight of 1 to the linguistic components, and weights from 0.2 to 2, in steps of 0.2, to the visual components. Both here and in our further analyses, we tested various weights, since it was not clear which weight would produce optimal results. Finally, for each of the four similarity/relatedness datasets, we compared the 10 resulting linguistic-visual models with the purely linguistic model, by normalizing the correlations and using two-tailed Z-tests. The same type of analyses were run for the linguistic-emotional models. Results are in Fig. 2.



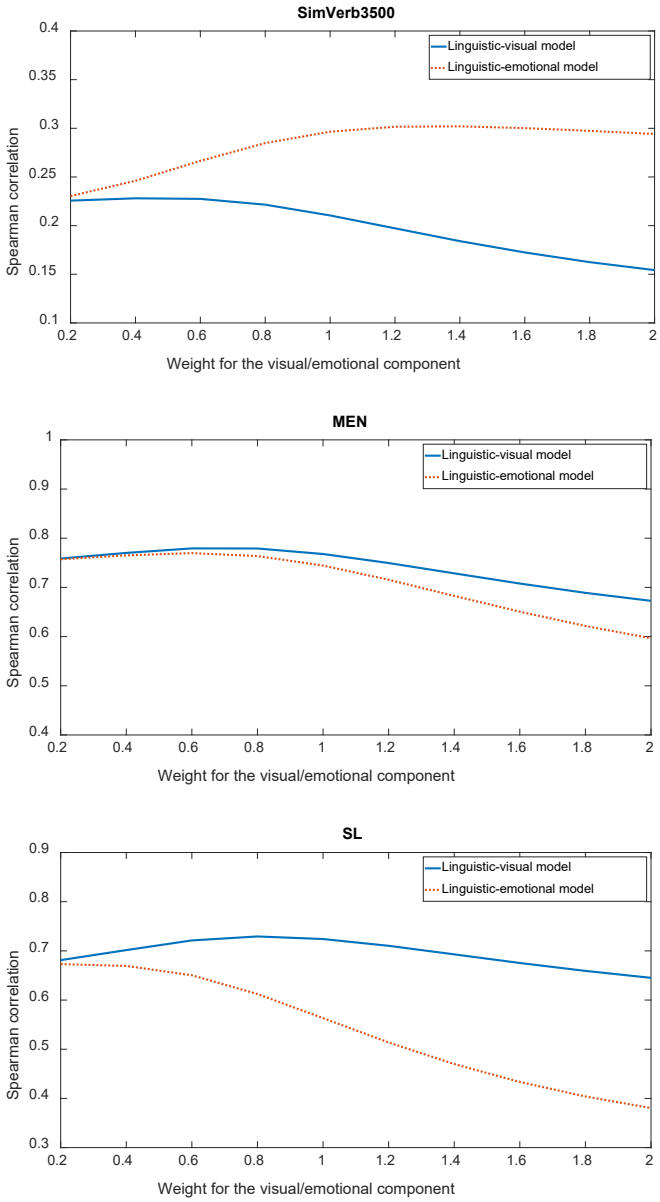


Figure 2. Model performance for the linguistic-visual and linguistic-emotional models. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2

The tests indicate that adding visual information has a significant positive effect only for the SL dataset ($p < .001$), for weights ranging from 0.6 to 1.2, and a significant negative effect for the MEN dataset ($p < .001$), for weights between 1.6 and 2. These results seem to be at odds with previous studies showing that linguistic-visual models always perform slightly better than purely linguistic ones. However, firstly, in almost all the other studies, the authors either weigh the linguistic and visual representations equally, by default (e.g., Kiela, Hill, Korhonen, & Clark, 2014; Silberer, Ferrari, &

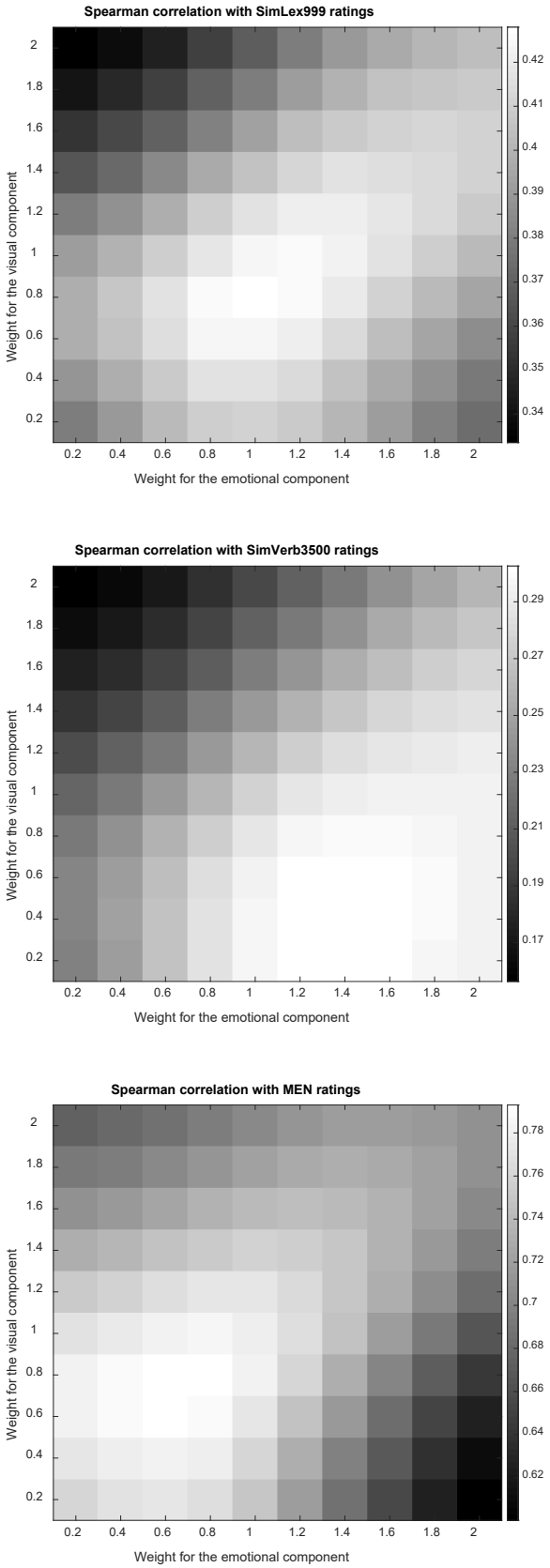
[1] The Bonferroni correction was applied when assessing the statistical significance of all the results presented in this study.

Lapata, 2013), or they only employ the weight that gives the best results for the integration (e.g., Bruni, Tran, & Baroni, 2014; Bruni, Uijlings, et al., 2012), which leaves room for null or detrimental results of linguistic-visual integration, when employing sub-optimal weights. Secondly, we use a linguistic model that is trained over a corpus of 6 billion words, whereas other studies (e.g., Hill & Korhonen, 2014; Kiela & Bottou, 2014; Silberer & Lapata, 2012) typically employ considerably smaller corpora (i.e., containing between 80 and 800 million words). Since smaller corpora lead to a poorer performance of the linguistic model, this leaves more room for a beneficial effect of adding visual information in the other studies, as compared to our study.

Adding emotional information is significantly beneficial only for the SimVerb3500 dataset ($p < .00125$), for weights ranging from 1.2 to 1.6, while it is significantly detrimental for the MEN dataset ($p < .001$), for weights between 1.4 and 2, and for the SL dataset ($p < .001$), for weights between 0.6 and 2. The SimVerb3500 dataset is different from all the others in that it is the only one including only verbs (which are not highly represented in any other dataset). As verbs (words referring to events) are considered to be more abstract, this finding is in line with the view that emotional information is especially important for abstract words (Kousta et al., 2011).

Linguistic-visual-emotional model vs linguistic-visual, linguistic-emotional, and purely linguistic models.

In order to compare the trimodal model with the bimodal and unimodal ones, we again start by normalizing the linguistic, visual, and emotional representations, to unit length. We then construct trimodal models by assigning a weight of 1 to the linguistic components, and weights from 0.2 to 2, in steps of 0.2, to the visual and emotional components, in all pairwise combinations for the last two components. Next, for each dataset, we select the best five and worst five trimodal models, in terms of performance, and compare them to their corresponding linguistic-visual models (i.e., obtained by removing the emotional component), linguistic-emotional models (i.e., obtained by removing the visual component), and purely linguistic model (i.e., obtained by removing both the visual and emotional components). The results are shown in Fig. 3.



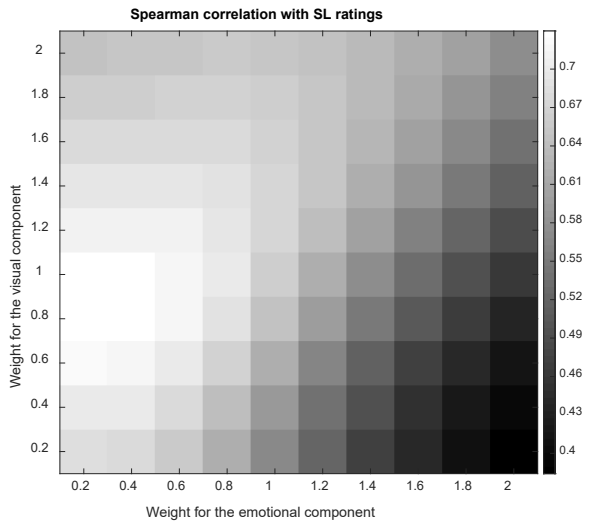


Figure 3. Model performance for the linguistic-visual-emotional model. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2

When comparing the performance of the trimodal models to that of their corresponding linguistic-visual models, the addition of an emotional component has a significant positive effect for the best models on the SimVerb3500 dataset ($p <$ .0016), and a significant negative effect for the worst models on the MEN and SL datasets ($p <$ .001). These results are very similar to those found when comparing the linguistic-emotional models to the purely linguistic one, and might be explained by the fact that verbs, such as those that make up the SimVerb3500 norms, are relatively abstract. In contrast, for concrete nouns, which form the majority of pairs from the MEN and SL norms, emotion should not have a positive effect (the finding of a detrimental effect is unexpected but potentially interesting as may indicate that adding affective information may reduce the separation between different types of words).

The comparison between the trimodal models and their corresponding linguistic-emotional models reveals that including a visual component is significantly beneficial for the best models on the SL dataset ($p <$ .001), but significantly detrimental for two of the worst models on the SimVerb3500 datasets ($p <$ .001). Again, SL consists only of concrete nouns, for which visual information is very salient, while SimVerb3500 consists only of verbs, the semantics of which is likely not to be properly captured in a few tens of images per word, due to its complexity.

Finally, contrasting the trimodal models with the purely linguistic one, we find that bringing in both visual and emotional information significantly increases performance for the best models on the SimVerb3500, MEN, and SL datasets ($p <$ .0016), while it significantly decreases performance for the worst models on the MEN and SL datasets ($p <$ .001). These results are a combination of the partial results regarding the effects of appending visual and emotional components to the purely linguistic and bimodal models, which indicates little overlap between vision and emotional representation.

Comparing the models for concrete and abstract words

In order to test whether visual content is more important for more concrete words, while emotional content for more abstract words, we first combined the SimLex999 and SimVerb3500 datasets, as they cover a broader range of concreteness ratings than MEN and SL. Then, we divided the merged dataset into a low and a high concreteness subset. More specifically, we selected the bottom 25% and the top 25% of pairs, based on the mean concreteness of each word pair covered by the concreteness norms of Brysbaert, Warriner, and Kuperman (2014). We then tested the performance of the emotional and visual models, the two bimodal models, and the trimodal models, setting all the weights set to 1. The results are displayed in Fig. 4.

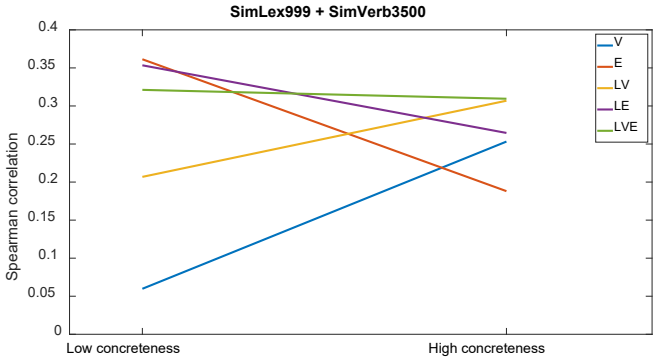


Figure 3. Model performance for low and high concreteness word pairs.

Using one-tailed *Z*-tests, after normalizing the correlations, we found that the performance of the visual model is higher for more concrete pairs, in comparison to the less concrete ones, for the visual ($p <$ .001) and linguistic-visual ($p <$ .01) models. Also, the emotional model has a better performance for the more abstract pairs, as opposed to the less abstract ones. Non-significant results were obtained for the linguistic-emotional and trimodal models. These results seem to suggest that the positive effect of adding visual information should be greatest for datasets consisting mainly of more concrete words, such as MEN and SL, while the beneficial effect of including emotional information should be largest for datasets made up mainly of more abstract words, such as SimLex999 and SimVerb3500.

Discussion

A first goal of this paper was to present an evaluation of visual models in order to identify the model(s) better fitting behavioural semantic data. We found that convolutional neural networks models (i.e., K&B, Alexnet, GoogLeNet, VGG-19) have a better performance than a classical, bag-of-visual-words model (i.e., SIFT), when tested over a large dataset of similarity/relatedness ratings. Among the

convolutional models, GoogLeNet gave the best results, followed by AlexNet, VGG-19, and K&B.

The second, and main goal was to develop models that integrate linguistic, visual and emotional information and to assess their performance against purely linguistic models and models that only include either visual or emotional features. We chose the DeepMoji model for a number of reasons, namely: its state-of-the-art performance in a number of emotional tasks; its distributional nature, since it predicts the occurrence of an emoticon based on its immediate linguistic context; its capacity to use rich emotional information, as it is trained on tweets containing 64 types of emoticons; its high dimensionality, which allows it to encode complex patterns of emotion-based word similarity.

In order to better understand the relative importance of each visual and emotional component, we carried out comparisons in which we parametrically varied the weight of visual and/or emotional information. In this manner, we can see when adding this information leads to better or worse performance. In general, we found that including non-linguistic information has a positive impact. However, first, this impact is modulated by whether the dataset includes predominantly concrete or abstract words. As expected on the basis of previous literature (e.g., Kousta et al., 2011) we see that including visual information is particularly beneficial to more concrete concepts whereas including emotional information is particularly beneficial to more abstract concepts. This is clearly visible when we assess model performance separately for more concrete and abstract words (see Fig 4). It is also clear from the comparison between MEN (only concrete words) and SimVerb3500 (only verbs, hence more abstract): across comparisons, we see that indeed visual information brings more benefit to the former, whereas emotional information brings more benefit to the latter.

Second, the effect is modulated by the weights attributed to the different types of information. While the theoretical interpretation of the differences we found related to weights is not immediate, this finding may have practical value for future modelling.

As mentioned in the introduction, a previous study (De Deyne et al., 2018) also examined the change in performance for distributional models of semantics, when adding experiential (i.e., visual and emotional) information. They found that adding experiential information led to little or no improvement for internal language models, but had a moderate positive effect for external language models. Moreover, they also found that adding visual information had the greatest effect for concrete words, while introducing affective information had the largest impact for abstract words. This finding mirrors our own, when comparing the linguistic-visual and linguistic-emotional models to the purely linguistic model.

However, there are a number of key differences between their approach and ours. Firstly, we avoided the potentially controversial distinction between external and internal language models, focusing on an objective corpus-based approach. Secondly, in a similar vein, we decided to use an emotional model that learns affective information indirectly, by predicting the co-occurrence of emojis and text in a corpus, rather than using emotional representations derived directly from valence, arousal and dominance norms (Mohammad, 2018; Warriner, Kuperman, & Brysbaert, 2013). This also increases the coverage of our model. Finally, since the resulting representations in our model are high-dimensional, they might provide more fine-grained information than representations with only three dimensions.

References

Blaney, P. H. (1986). Affect and memory: A review. *Psychological Bulletin*, *99*(2), 229-246.

Bruni, E., Tran, G. B., & Baroni, M. (2011). Distributional semantics from text and images. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics* (pp. 22-32).

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1-47.

Bruni, E., Uijlings, J., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM International Conference on Multimedia* (pp. 1219-1228).

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Buechel, S., & Hahn, U. (2016, August). Emotion analysis as a regression problem—dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of the 22nd European Conference on Artificial Intelligence* (pp. 1114-1122).

Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*.

De Deyne, S., Navarro, D., Collell, G., & Perfors, A. (2018, November 28). Visual and Affective Grounding in Language and Mind. https://doi.org/10.31234/osf.io/q97f8

Deriu, J., Gonzenbach, M., Uzdilli, F., Lucchi, A., Luca, V. D., & Jaggi, M. (2016). Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (pp. 1124-1128).

Dolan, R. J. (2002). Emotion, cognition, and behavior. *Science*, *298*(5596), 1191-1194.

Eich, E., Macaulay, D., & Ryan, L. (1994). Mood dependent memory for events of the personal past. *Journal of Experimental Psychology: General*, *123*(2), 201-215.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion

and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1616–1626).

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th International World Wide Web Conference* (pp. 406-414).

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). SimVerb-3500: A Large-Scale Evaluation Set of Verb Similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2173–2182).

Glenberg, A. M., Graesser, A. C., & de Vega, M. (Eds.). (2008). *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford University Press.

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

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*(4), 665-695.

Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 255-265).

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 36-45).

Kiela, D., Hill, F., Korhonen, A., & Clark, S. (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* (pp. 835-841).

Kiela, D., Vero, A. L., & Clark, S. C. (2016). Comparing data sources and architectures for deep visual representation learning in semantics. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 447–456).

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473-481.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14-34.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097-1105).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, *60*(2), 91-110.

Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of the ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access* (pp. 321-327).

Mohammad, S. M. (2012, June). #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (pp. 246-255).

Mohammad, S. (2018, July). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 174-184).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Nethods in Natural Language Processing* (pp. 1532-1543).

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3-4), 175-190.

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science*, *21*(2), e12549.

Silberer, C., Ferrari, V., & Lapata, M. (2013). Models of semantic representation with visual attributes. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 572-582).

Silberer, C., & Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1423-1433).

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 721-732).

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations* (pp. 1-14).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, .V, & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-9).

Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: the role of emotion. *Cerebral Cortex*, *24*(7), 1767-1777.

Von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 319-326).

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45, 1191–1207.