Constructing semantic models from words, images and emojis

Armand S. Rotaru (armand.rotaru.14@ucl.ac.uk)

Gabriella Vigliocco (g.vigliocco@ucl.ac.uk)

Experimental Psychology Department

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. 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 and affective representations, derived from state-of-the-art existing models, by choosing models that best fit available human semantic data and extending the number of concepts they cover. Crucially then, we assess whether adding affective representations (obtained from a neural network model designed to predict emojis from co-occurring text) improves the model's ability to fit semantic similarity/relatedness judgements from a purely linguistic and linguistic-visual model. We find that, given specific weights assigned to the models, adding both visual and affective representations improve performance, with visual representations providing an improvement especially for more concrete words, and affective representations improving especially the fit for more abstract words.

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

Introduction

Distributional models of semantics, such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), Topics (Griffiths, Steyvers, & Tenenbaum, 2007) and Continuous Bag-of-Words (CBOW; Mikolov, Chen, Corrado, & Dean, 2013), to name just a few, have been the focus of numerous studies investigating semantic representations and processes. Such models are based on the *distributional hypothesis* (Harris, 1954), according to which "words that occur in similar contexts tend to have similar meanings" (Turney & Pantel, 2010). For example, the words "comet" and "asteroid" are likely to be semantically similar, given that they both occur in linguistic contexts related to astronomy. In contrast, it is reasonable to assume that the words "poem" and "doughnut" share little semantic content, since they occur in very different contexts (i.e., literary vs culinary, respectively). Despite the success of distributional, linguistic models in accounting for behavioural effects in a variety of semantic tasks, it is well recognized that their cognitive plausibility is limited, among other things, because they suffer from the *symbol grounding problem* (Harnad, 1990). This problem refers to the fact that the meaning of symbols (i.e., words) is computed based on other symbols (i.e., other words), but without connecting those symbols to their real-world referents. For instance, distributional representations might tell us that "hammer" and "nail" are semantically linked, but they do not capture the sensory and motor information obtained by interacting with the two objects (e.g., the shape of the hammer, the texture of its handle, the motion used when hitting the nail, or the sound produced when the nail is struck). 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 (see Meteyard, Cuadrado, Bahrami, & Vigliocco, 2012).

Recent computational models of semantics have attempted to reconcile distributional and embodied theories, by combining linguistic and perceptual (i.e., visual) representations (see Andrews, Frank & Vigliocco, 2015, for a discussion). The underlying assumption is that the two classes of representations capture complementary aspects of meaning. For example, using a semantic clustering task, Riordan and Jones (2011) showed that distributional models focus mostly on information about actions, functions, and situations, but not on the perceptual properties of objects. Instead, such properties are better captured by featural models, which are rich in perceptual information.

The fact that language and vision provide complementary sources of information is best illustrated by the finding that multimodal (i.e., textual-visual) models outperform both purely linguistic and purely visual models, in a wide range of tasks, such as similarity rating (Bruni, Tran, & Baroni, 2014; Kiela, Verő, & Clark, 2016), free association (Silberer & Lapata, 2012), and semantic categorization (Bruni, Tran, & Baroni, 2011; Silberer & Lapata, 2014). The improvement brought about by combining modalities is not tied to the use of a particular dataset (e.g., in the case of similarity/relatedness rating task, Kiela, Verő, & Clark, 2016, showed that comparable results are obtained for the SimLex999 and MEN datasets; Bruni, Tran, & Baroni, 2014; Hill, Reichart, & Korhonen, 2015). Also, the superiority of multimodal models over unimodal ones does not seem to depend on the model architecture, as it has been observed for numerous types of linguistic models, such as HAL (Lund & Burgess, 1996; used in Bruni, Tran, & Baroni, 2014), Topics (Griffiths, Steyvers, & Tenenbaum, 2007; used in Silberer & Lapata, 2012), and Skip-gram (Mikolov, Chen, Corrado, & Dean, 2013; used in Kiela & Bottou, 2014), as well as for multiple types of visual models, such as SIFT (Lowe, 2004; used in Bruni, Tran, & Baroni, 2011), PHOW (Bosch, Zisserman, & Muñoz, 2007; used in Kiela, Hill, Korhonen, & Clark, 2014), and AlexNet (Krizhevsky, Sutskever, & Hinton, 2012; used in Kiela & Bottou, 2014). Furthermore, very similar results are obtained regardless of the method by which the modalities are combined (e.g., by concatenating the representations; Bruni, Tran, & Baroni, 2011; or by feeding the representations to a neural network; Silberer & Lapata, 2014).

However, empirical work has shown that semantic representations are not only grounded in sensory-motor experience (of which models based on static images can only provide a very imprecise impression), 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), and it affects language processing. Studies have found that words with affective features (valenced words) are processed differently than neutral words. In particular, once factors such as length and familiarity are taken into account, both negative and positive words are processed faster than neutral words in lexical decision tasks (a relatively shallow task in which subjects are required to distinguish between words and nonwords; Kousta, Vinson & Vigliocco, 2009; see Vinson, Ponari & Vigliocco, 2014, for a discussion) and this effect is modulated by the frequency of the words (Kuperman, Estes, Brysbaert & Warriner, 2014). Although an advantage for negative words has not been consistently reported, especially using tasks such as the Emotional Stroop task where negative words tend to lead to slower RTs (but see Larsen, Mercer, & Balota, 2006, for a review), a general difference in processing emotional vs. neutral words is well established. Importantly, Kousta et al. (2011) found that a much larger number of abstract than concrete concepts are valenced (have positive or negative emotional associations, and this is true even when words directly denoting emotions are excluded) 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. Finally, Ponari, Norbury, and Vigliocco (2018) showed that emotionally valenced words are learnt earlier and better recognized by children up to 9 years of age. On the basis of these results and within a general embodiment framework, Kousta et al (2011) argued that semantic representations do not only embed sensorimotor properties but also emotional properties. These emotional properties may be especially important for abstract concepts (e.g., *religion, society, idea*), however, as emotional associations are not limited to abstract concepts, these would play a general role in semantic representation. It is important to note here that the claim is limited to state that affective features are more important for abstract than concrete concepts, by virtue of there being more abstract than concrete words with affective connotations. Of course, other types of information, including motor-based information (as highlighted in approach-avoidance models) can also be important in the representation of abstract words, as suggested by a number of studies (e.g., Desai, Reilly, & van Dam, 2018; Dreyer & Pulvermüller, 2018; Ghio, Vaghi, & Tettamanti, 2013; Mazzuca et al., 2018). Likewise, emotional connotation does play a role in some concrete words' representation (e.g., *knife, kitten*).

Getting back to computational models, while many models of semantic representation 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). In this study, the authors compared two classes of semantic models, reflecting either external or internal language, respectively. External models assume that language is expressed as a collection of verbal messages, produced by a speech community and captured by written or spoken text corpora. In contrast, internal models consider language to consist of a set of mental representations, the contents of which can be made explicit by using tasks such as feature generation (McRae, Cree, Seidenberg, & McNorgan, 2005) and free association (Nelson, McEvoy, & Schreiber, 2004). As a result, internal language models (more specifically, models based on free association norms) should already contain a large amount of visual and affective information, which means that such models would benefit little from the addition of visual and affective representations. However, as indicated by the symbol grounding problem, this type of information would largely be absent from external language models (more specifically, distributional models, trained on word corpora), and therefore the performance of such models should improve considerably when visual and affective representations are added. The authors tested this hypothesis by comparing the ability of the two types of models to account for triadic comparison ratings (i.e., out of three words, which two are the most related) and semantic similarity/relatedness ratings (i.e., how similar/related two words are). For both tasks, the authors found that including visual and emotional information led to little or no improvement for the internal model, but a moderate positive effect for the external model. Furthermore, the performance of the internal model was as least as good as that of the external model, in line with the assumption that the information contained by the internal model is richer and more multimodal than that reflected by the external model. Also, comparable results were found for different external models (i.e., in terms of architecture and/or corpus size), similarity/relatedness datasets, as well as taxonomic levels (i.e., words at the basic level and across different levels).

Here, we develop a quite different multimodal model of semantics that incorporates linguistic, visual and emotional information from corpora of text, images and emojis, and test this multimodal model against existing datasets of ratings of semantic similarity/relatedness of words. We choose our emotional and visual models as arguably the best models of their class, to fit human semantic similarity/relatedness ratings, and we ensure they have as large coverage as possible. We expect that the multimodal model integrating linguistic, visual and emotional information will outperform a purely linguistic model, as well as models that combine linguistic-visual and linguistic-emotional information. In addition, we expect that adding visual 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). As in previous models, our work uses visual and emotional data that can only be considered as providing a static window into the embodied sensory-motor and affective states of the agent, rather than truly embodied information.

The success of previous work, and its alignment with human data, however, makes us confident that although the obvious limitations, the type of visual (and emotional) representations we employ provides a valid proxy.

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 (3,500 pairs of verbs; Gerz et al., 2016), MEN (3,000 pairs of nouns, verbs, and adjectives; Bruni, Tran, & Baroni, 2014), and SL[1] (7,576 pairs of nouns; Silberer & Lapata, 2014). We chose these datasets mainly because they are some of the largest currently available, but also because the word pairs they contain are very diverse in terms of concreteness and valence, as well as parts of speech. With respect to word pair concreteness (Brysbaert, Warriner, & Kuperman, 2014), SimLex999 (*M* = 3.61, *SD* = 1.09, range = 1.3-5) and SimVerb3500 (*M* = 3.08, *SD* = 0.69, range = 1.45-4.76) cover a broad range of values, whereas MEN (*M* = 4.4, *SD* = 0.48, range = 1.79-5) and SL (*M* = 4.83, *SD* = 0.14, range = 3.63-5) consist predominantly of concrete words.

Model selection

Language Model.

We chose GloVe (Pennington, Socher, & Manning, 2014) as our language model. GloVe is trained on a corpus of 6 billion words, using 300-dimensional representations. This model was proposed as a solution to certain (potential) shortcomings of two classes of popular distributional models, namely global matrix factorization models, such as LSA (Landauer & Dumais, 1997), and local context window models, such as CBOW and Skip-gram (Mikolov, Chen, Corrado, & Dean, 2013). According to the authors, the first class of models perform poorly on word analogy tasks, denoting a sub-optimal vector space structure, while the second class of models do not exploit global co-occurrence information. GloVe has been shown to have a performance better than, or equal to, several

[1] The SL norms contain both semantic and visual similarity ratings. To make the analyses comparable across the different datasets, we employ only the semantic similarity data.

other state-of-the-art distributional models, such as vLBL and ivLBL (Mnih & Kavukcuoglu, 2013), HPCA (Lebret & Collobert, 2014), as well as CBOW and Skip-gram, in tasks that involve solving analogies, predicting similarity ratings, and recognizing named entities. This makes GloVe one of the best linguistic models available.

Emotion Model.

Computational models of emotion derive affective word representations by predicting affective labels/classes associated with the documents in a corpus. Two broad classes of computational models can be distinguished, namely traditional, bag-of-words models, and modern, neural network models. Traditional models are typically trained to perform sentiment classification (i.e., using positive vs negative ratings; Maas et al., 2011), emotion classification (i.e., using basic emotions; Mohammad & Kiritchenko, 2015), or mood classification (i.e., using moods such as *amused*, *frustrated*, *lethargic*, *thirsty*, etc.; Leshed & Kaye, 2006). One important shortcoming of such models is that they do not take syntax into consideration: for instance, given that the words "lack" and "flaws" have a negative connotation, the expression "a lack of flaws" would be classified as being (strongly) negative, rather than positive. Another problem is that the dimensionality of the representations is relatively low (i.e., typically less than 50), being limited by either the nature of the nature of the affective information, namely the number of classes, or by the computational resources needed to train the models. Modern models are usually based on recurrent neural networks (Abdul-Mageed & Ungar, 2017; Tai, Socher, & Manning, 2015), as well as feedforward neural networks (Tang et al., 2014) and convolutional neural networks (Kim, 2014). They are trained on the same tasks as the traditional models, as well as on the emoji classification task (i.e., determining the type of emoji present in a document), which exploits a rich source of affective information. Unlike traditional models, recurrent neural networks have the advantage of being sensitive to word order, given that they process text in a sequential manner and have an internal representation (or memory) of the words processed before the current word. An additional benefit of using neural network models is the fact that the dimensionality of the representations is high (typically greater than 100), since it is no longer tied to the number of classes.

In order to select an appropriate model for our study, we compared four of the most recent and well-performing models of emotion. The first model (CNN) is the convolutional model by Coman, Nechaev, and Zara (2018; 300

dimensions). The second model (GRU) and the third model (LSTM) are recurrent neural networks, namely the gated recurrent unit model and the long short-term memory model by Çöltekin and Rama (2018; 128 dimensions). The fourth model (DeepMoji) is the stacked LSTM model by Felbo et al. (2017; 256 dimensions). All the models were trained on the Twemoji dataset (Cappallo et al., 2019), consisting of 15 million tweets, each containing one or more emojis. From the full corpus, we kept only the tweets associated with emojis of facial expressions, as they are reliable and unambiguous indicators of emotion. This resulted in a subset of almost 10 million tweets. The task of the models was to predict the emoji(s) co-occurring with each tweet. After training, we tested the models' ability to account for similarity/relatedness ratings, for 12,659 word pairs covered by all the models and generated by combining the four sets of ratings, and then linearly scaling the values to the range [0,1]. Model performance was measured using the Spearman correlation between the cosine similarity of the model representations, and the ratings from the norms. The results are shown in Fig. 1.

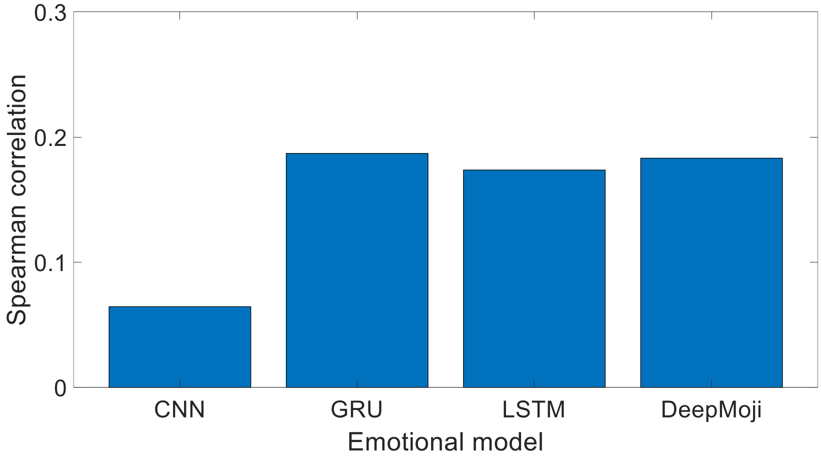

Figure 1. Spearman correlations between model cosine similarities and subjective similarity/relatedness ratings.

All the correlations are significant (p < .0001)[2], which indicates that the models capture some of the semantic information reflected in the subjective similarity/relatedness ratings. Given that we want to find the best model, we apply the Fisher Z-Transformation and then run two-tailed Z-tests for all the 6 possible pairings of models. Only

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

the differences between CNN and all the other models are significant ($p < .0001$), and they reveal that GRU, LSTM, and DeepMoji are comparable in performance, while being better than CNN.

As a result, our model of choice is DeepMojj, but trained over a large corpus of 1.2 billion tweets, as made publicly available by the authors of the model. This version of the model has been shown to obtain state-of-the-art performance in tasks involving emotion and sentiment analysis, as well as sarcasm detection. The 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 representations 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 emojis, and thus it is able to represent complex patterns of word similarity, driven by richer emotional information than that captured by subjective norms. The emojis employed by the model, as well as their frequency, are shown in Fig. 2.

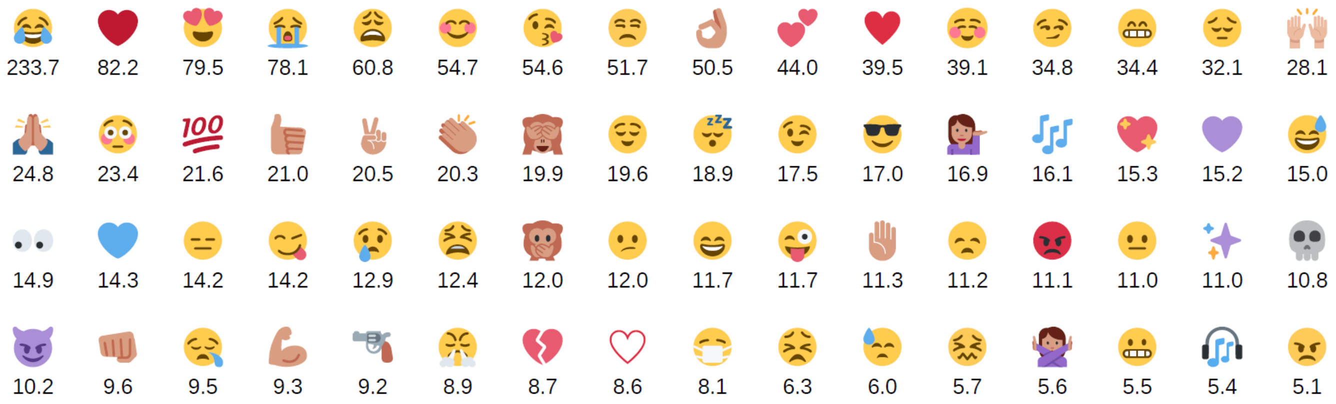

Figure 2. Emojis covered by the DeepMoji model, together with their frequency, in millions. Permission pending.

In order to obtain a more detailed understanding of the emotional information captured by the DeepMoji model, we used PCA and extracted the first 10 principal components from the model representations. Then, we computed the (absolute) Spearman correlations between each component and the affective norms collected by Mohammad (2018), covering subjective ratings of word valence, arousal, and dominance. Our analysis included 13,678 words common to both the model and the norms. The results are shown in Fig. 3.

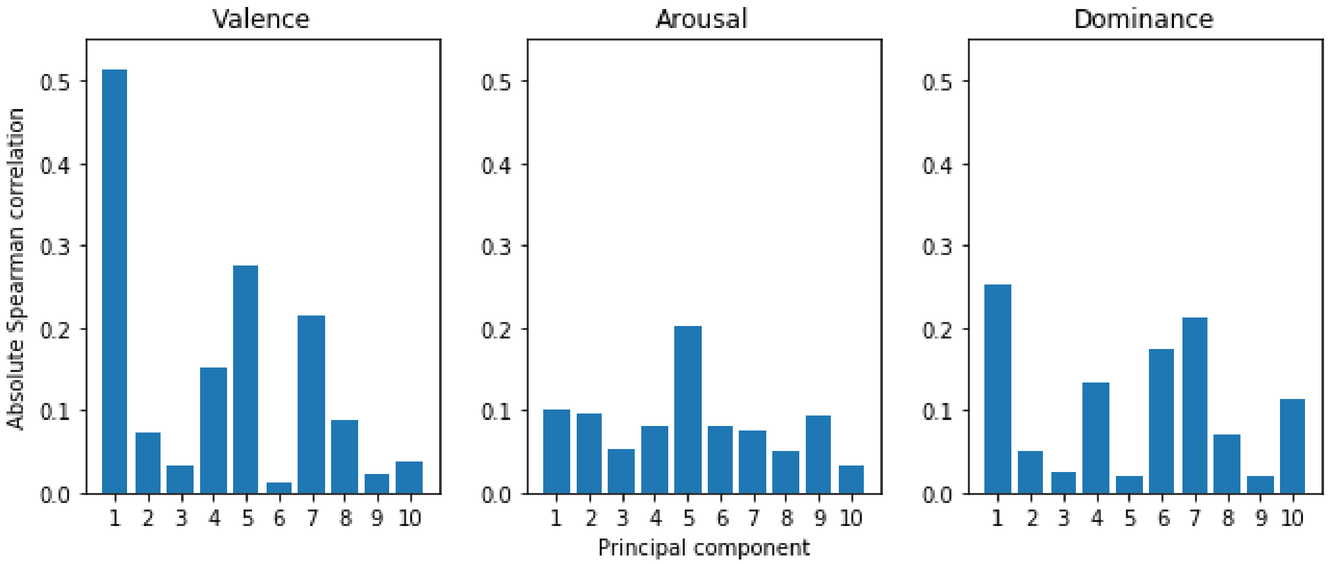

Figure 3. Absolute Spearman correlations between model components and valence, arousal, and dominance ratings.

All the 10 components correlate significantly with valence (except for P6, where *p* = .19), arousal, and dominance. This provides additional evidence that the DeepMoji model is sensitive to affective dimensions of word meaning. Since PC1, PC5, and PC7 seem to be most strongly correlated with the subjective ratings, we also decided to find the words corresponding to the most extreme values for each of the three components. The words, as well as their mean affective ratings, are shown in the Table 1.

Table 1. Words with the lowest and highest values for PC1, PC5, and PC7, as well as the words' mean affective ratings. The words with the lowest values are written in boldface, for better visibility.

	PC1 (low)	PC1 (high)	**PC5 (low)**	PC5 (high)	**PC7 (low)**	PC7 (high)
Words	**Headache**	Follow	**Bitch**	Happiest	**Naked**	Important
	Sad	Birthday	**Beg**	Hope	**Stalk**	Appreciate
	Sucks	Happy	**Hood**	Smile	**Imagine**	Great
	Stressed	Direction	**Hoe**	Gift	**Secret**	Deserve
	Tired	Amazing	**Thug**	Dream	**Sunshine**	Follow
	Fail	Proud	**Savage**	Happiness	**Addicted**	Merry
	Crying	Happiest	**Fuck**	Tour	**Nervous**	Respect
	Impossible	Pizza	**Hell**	Notice	**Drunk**	Proud
	Migraine	Cute	**Pussy**	Happy	**SOS**	Inspiration
	Pain	Sexy	**Ass**	Reading	**Beg**	Supporting
Mean valence	**0.14**	0.88	**0.25**	0.86	**0.47**	0.85
Mean arousal	**0.62**	0.65	**0.75**	0.49	**0.56**	0.55
Mean dominance	**0.29**	0.68	**0.44**	0.60	**0.46**	0.75

When comparing the ratings for two sets of extreme words, the results indicate significant differences in valence (all $|t| > 4.29$, $p < .001$) and dominance ($|t| > 2.28$, $p < .035$), for each of the three components. However, the two sets of extreme words differ significantly, in terms of arousal ($t = 2.83$, $p = .011$), only for PC5. These results are consistent with those from the correlation analysis, and suggest that the DeepMoji model is most sensitive to valence, followed by dominance, followed by arousal. In addition, PC1, PC5, and PC7 seem to capture different aspects of emotion. This is most evident when looking at the words from the lower extreme (i.e., the ones displayed in boldface): for PC1, they seem to be related to bodily causes and effects of emotion (e.g., *headache*, *migraine*, *pain*, *crying*, *tired*), and belong to polite language; for PC5, they refer mostly to individuals (e.g., *bitch*, *hoe*, *thug*, *savage*), and belong to vulgar language; for PC7, they include actions (e.g., *stalk*, *beg*), physical states (e.g., *addicted*, *drunk*), and pieces of information (e.g., *SOS*, *secret*), all belonging to polite language.

Visual Model.

A wide variety of visual models, trained over image datasets, have been developed in the last 20 years. One of the earliest, as well as most popular approaches to creating visual representations (e.g., Bruni, Tran, & Baroni, 2011; Feng & Lapata, 2010; Kiela, Hill, Korhonen, & Clark, 2014) is the bag-of-visual-words (i.e., BoVW; Sivic & Zisserman, 2003) approach, inspired by work on linguistic, distributional models of semantics. Within the BoVW approach, some of the most widely used feature vectors are obtained from Scale Invariant Feature Transform (SIFT; Lowe, 2004) descriptors, which have the advantage of being invariant to translations, rotations and rescalings, as well as to changes in perspective and illumination.

More recent approaches to image representation rely on deep convolutional networks (e.g., AlexNet; Krizhevsky, Sutskever, & Hinton, 2012; GoogLeNet; Szegedy et al., 2015; VGG-19; Simonyan & Zisserman, 2014), typically trained on supervised object recognition tasks. Such networks mimic the hierarchical nature of visual perception, by employing convolutional layers, which have units that receive input only from limited, spatially contiguous regions of the previous layer (i.e., the receptive fields of the units).

Just as we did for the emotion model, to select the best model for our study, we compared five of the most popular visual models for object recognition, 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; 4,096 dimensions), GoogLeNet (Szegedy et al., 2015; 1,024 dimensions), and VGG-19 (Simonyan & Zisserman, 2014; 4,096 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 7,611 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. 4.

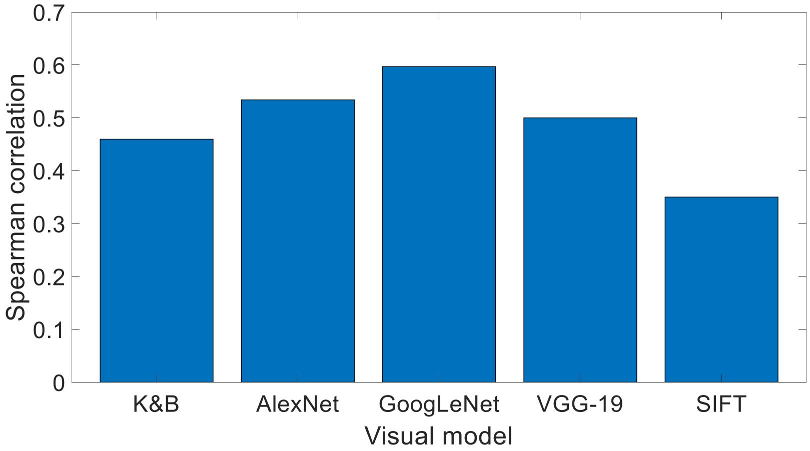

Figure 4. Spearman correlations between model cosine similarities and subjective similarity/relatedness ratings.

All the correlations are significant (*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 is not clear which weight would produce the best 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. The results are shown in Fig. 5.

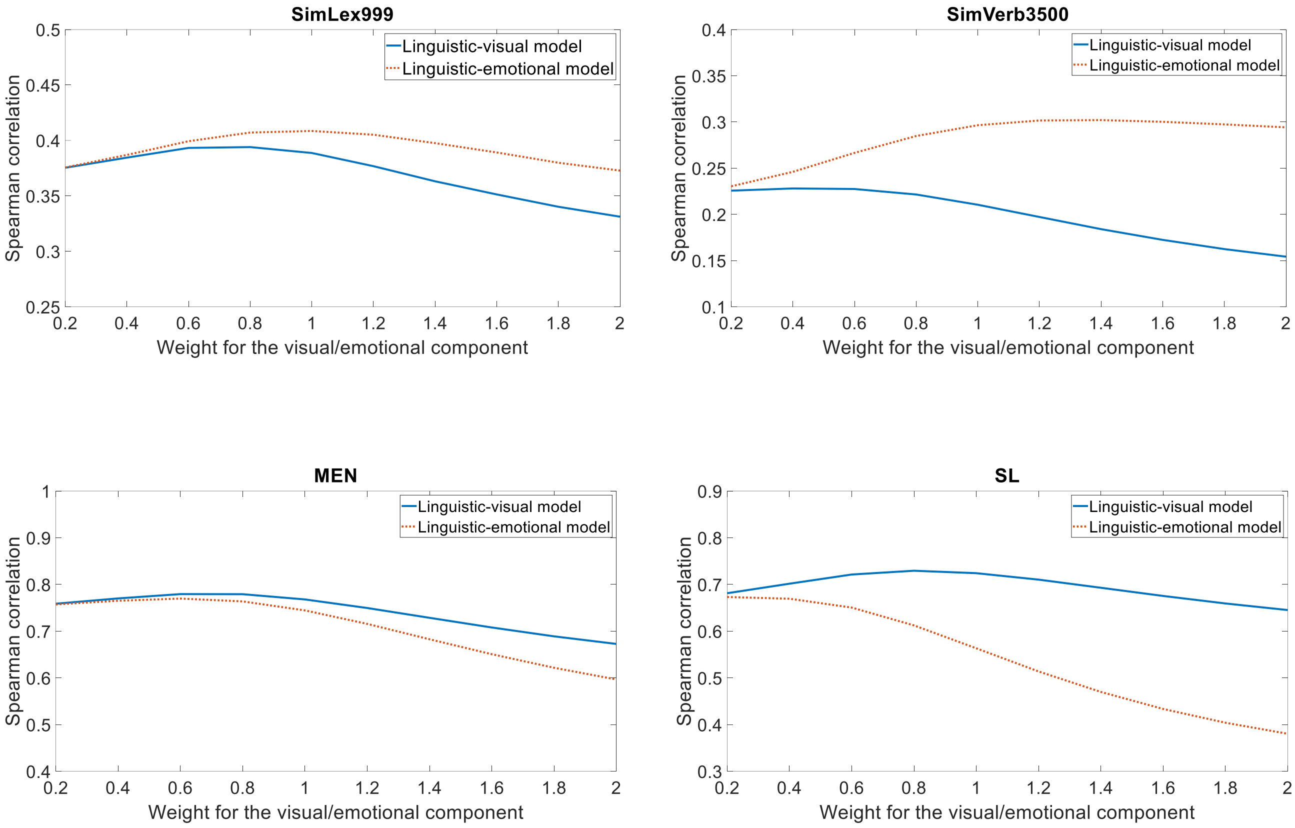

Figure 5. Model performance for the linguistic-visual and linguistic-emotional models. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2

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, & 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[3], 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 our predictions above that adding the emotional component may be especially useful for abstract concepts.

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

[3] As a caveat, it is worth keeping in mind that, in terms of performance, certain distributional models are more sensitive to corpus size than others (e.g., De Deyne et al., 2018; Sahlgren & Lenci, 2016).

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. 6 and Table 2.

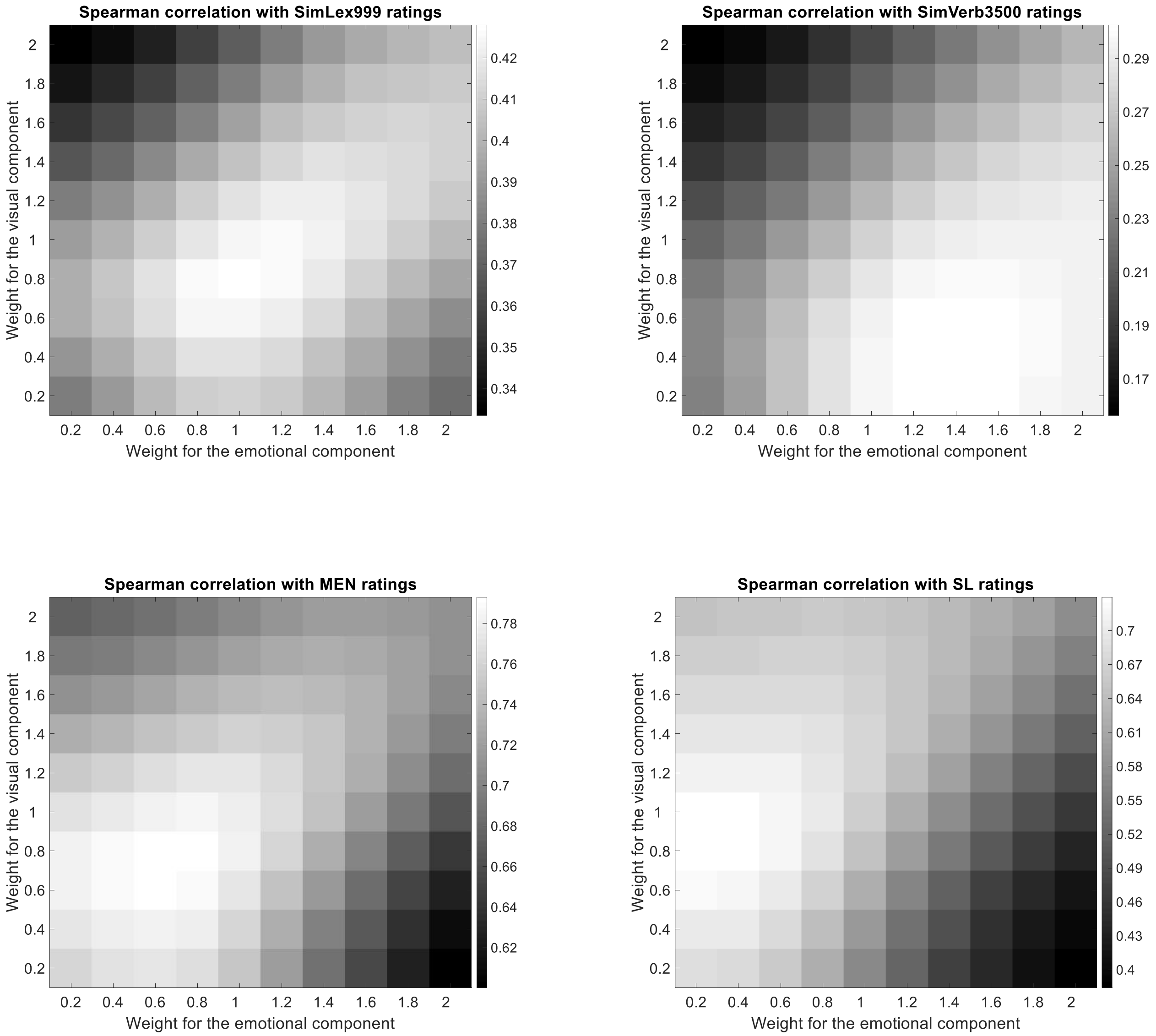

Figure 6. Model performance for the linguistic-visual-emotional model. The weights assigned to the visual/emotional component vary from 0.2 to 2, in steps of 0.2

Table 2. *p* values for comparing the best and worst linguistic-visual-emotional model to their corresponding linguistic-visual, linguistic-emotional and purely linguistic models, respectively.

Vis.	Emo.	LVE	LVE	LVE	Vis.	Emo.	LVE	LVE	LVE
weight	weight	vs LV	vs LE	vs L	weight	weight	vs LV	vs LE	vs L
SimLex999 – Best five models					SimLex999 – Worst five models				
0.8	1	.367	.604	.138	2	0.2	.956	.288	.339
1	1.2	.322	.576	.151	2	0.4	.851	.221	.410
0.8	1.2	.401	.584	.154	1.8	0.2	.961	.397	.459
0.8	0.8	.409	.631	.159	2	0.6	.701	.177	.531
1	1	.340	.664	.162	1.8	0.4	.838	.325	.559
SimVerb3500 – Best five models					SimVerb3500 – Worst five models				
0.4	1.4	**.001**	.981	**.001**	2	0.2	.933	.002	.005
0.2	1.4	**.001**	.990	**.001**	2	0.4	.746	< *.001*	.010
0.4	1.2	**.002**	.987	**.001**	1.8	0.2	.926	.007	.013
0.6	1.4	**.001**	.998	**.001**	1.8	0.4	.705	.002	.028
0.2	1.2	**.001**	.988	**.001**	2	0.6	.475	< *.001*	.028
MEN – Best five models					MEN – Worst five models				
0.8	0.6	.179	.027	**< .001**	0.2	2	< *.001*	.834	< *.001*
0.6	0.6	.201	.030	**< .001**	0.4	2	< *.001*	.420	< *.001*

0.8	0.8	.217	.008	**< .001**	0.2	1.8	< *.001*	.816	< *.001*
0.6	0.8	.374	.020	**.001**	0.6	2	< *.001*	.084	< *.001*
0.6	0.4	.387	.030	**.002**	0.4	1.8	< *.001*	.353	< *.001*

SL – Best five models					SL – Worst five models				
0.8	0.2	.963	**< .001**	**< .001**	0.2	2	< *0.001*	.809	< *0.001*
0.8	0.4	.847	**< .001**	**< .001**	0.4	2	< *0.001*	.335	< *0.001*
1	0.2	.938	**< .001**	**< .001**	0.2	1.8	< *0.001*	.773	< *0.001*
1	0.4	.968	**< .001**	**< .001**	0.6	2	< *0.001*	.034	< *0.001*
0.6	0.2	.978	**< .001**	**< .001**	0.4	1.8	< *0.001*	.254	< *0.001*

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 model, setting all the weights to 1. The results are displayed in Fig. 7.

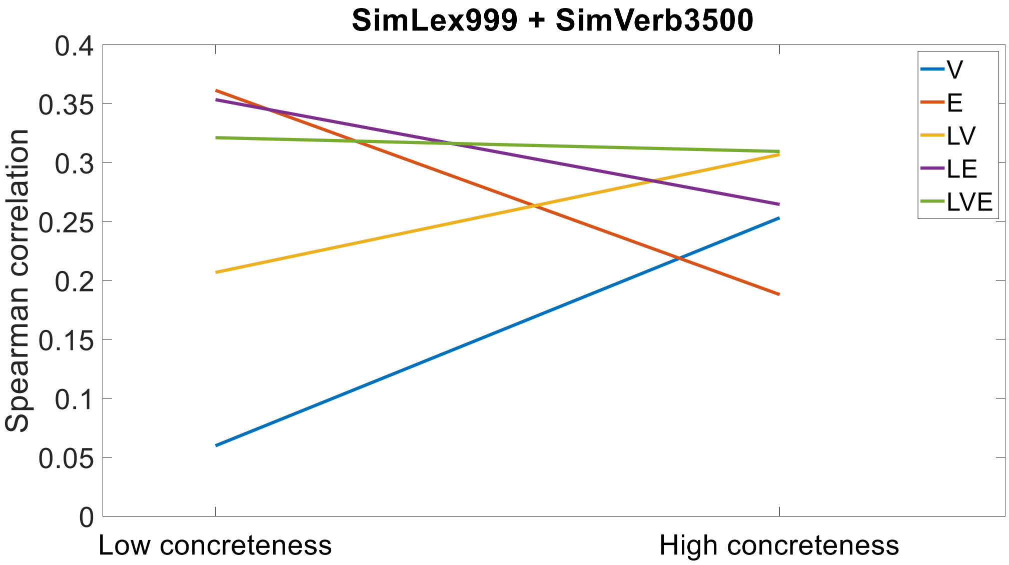

Figure 7. Model performance for low and high concreteness word pairs.

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 ($p < .001$) 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 and emotional models, in order to identify the model(s) better fitting behavioural semantic data. In the case of the emotional models, models based on recurrent neural networks (i.e., GRU, LSTM, DeepMoji) perform comparably, while being better than a convolutional neural network model (perhaps unsurprisingly, knowing that CNNs are not designed for operating with linguistic representations). Here, 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 emoji based on its immediate linguistic context; its capacity to use rich emotional information, as it is trained on tweets containing 64 types of emojis; its high dimensionality, which allows it to encode complex patterns of emotion-based word similarity. Moreover, with respect to the DeepMoji representations, we found that the first ten principal components are significantly correlated with subjective valence, arousal, and dominance ratings, which provides additional support to the hypothesis that the DeepMoji model captures affective information. For the visual models, we found that convolutional neural networks models (i.e., K&B, AlexNet, GoogLeNet, VGG-19), instead, 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 of our work was to develop and evaluate 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. 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 could see when adding this information leads to better or worse performance. We found that adding visual information had a positive effect in 4/40 cases, no significant effect in 33/40 cases, and a negative effect in 3/40 cases. When including emotional information, there was a positive effect in 3/40 cases, no significant effect in 26/40 cases, and a negative effect in 11/40 cases. Finally, when introducing both visual and emotional information, for the best models, the analyses revealed a positive effect in 15/20 cases, no effect in 5/20, and a negative effect in 0/20 cases; in contrast, for the worst models, the results indicated a positive effect in 0/20 cases, no significant effect in 10/20 cases, and a negative effect in 10/20 cases. In general, we found that whether the addition of non-linguistic increases or decreases model performance, or instead has no effect, is determined by the weights attributed to the different types of information, which may have practical value for future modelling.

In addition, it appears that this impact depends on whether the dataset includes predominantly concrete or abstract words. As expected on the basis of previous literature (e.g., Kousta et al., 2011), we found 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. 7). 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.

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 including 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. In De Deyne et al. (2018), external language models derive semantic representations from corpora of language, whereas internal language models derive semantic representations from free associations. Thus, the models differ in whether they use objective or subjective data (based on a metacognitive task), but both might be argued to tap into the same construct. We focus on an objective corpus-based approach, to avoid such potential criticisms. 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 a few dimensions.

Acknowledgements: We are grateful to Marco Marelli for helpful discussions regarding the results of the principal component analysis. The work reported here was supported by Economic and Social Research Council research grant (ES/P00024X/1), European Research Council Advanced Grant (ECOLANG, 743035) and Royal Society Wolfson Research Merit Award (WRM\R3\170016) to GV.

References

Abdul-Mageed, M., & Ungar, L. (2017). Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 718-728).

Bosch, A., Zisserman, A., & Muñoz, X. (2007). Image classification using random forests and ferns. In *Proceedings of the 11th IEEE International Conference on Computer Vision* (pp. 1-8).

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.

Cappallo, S., Svetlichnaya, S., Garrigues, P., Mensink, T., & Snoek, C. G. (2019). New Modality: Emoji Challenges in Prediction, Anticipation, and Retrieval. *IEEE Transactions on Multimedia*, *21*(2), 402-415.

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*.

Çöltekin, Ç., & Rama, T. (2018). Tübingen-Oslo at SemEval-2018 Task 2: SVMs perform better than RNNs in Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation* (pp. 34-38).

Coman, A. C., Nechaev, Y., & Zara, G. (2018). Predicting Emoji Exploiting Multimodal Data: FBK Participation in ITAmoji Task. In *EVALITA 2018 Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

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

Desai, R. H., Reilly, M., & van Dam, W. (2018). The multifaceted abstract brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1752), 20170122.

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

Dreyer, F. R., & Pulvermüller, F. (2018). Abstract semantics in the motor system?–An event-related fMRI study on passive reading of semantic word categories carrying abstract emotional and mental meaning. *Cortex*, *100*, 52-70.

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).

Feng, Y., & Lapata, M. (2010). Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 91-99).

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).

Ghio, M., Vaghi, M. M. S., & Tettamanti, M. (2013). Fine-grained semantic categorization across the abstract and concrete domains. *PloS one*, *8*(6), e67090.

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

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244.

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

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

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).

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

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., Verő, 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).

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746–1751).

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.

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.

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).

Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, A. B. (2014). Emotion and language: Valence and arousal affect word recognition. *Journal of Experimental Psychology: General*, *143*(3), 1065-1081.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Larsen, R. J., Mercer, K. A., & Balota, D. A. (2006). Lexical characteristics of words used in emotional Stroop experiments. *Emotion*, *6*(1), 62-72.

Lebret, R., & Collobert, R. (2014). Word Embeddings through Hellinger PCA. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 482-490).

Leshed, G., & Kaye, J. J. (2006). Understanding how bloggers feel: recognizing affect in blog posts. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems* (pp. 1019-1024).

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

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203-208.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (Volume 1)* (pp. 142-150).

Mazzuca, C., Lugli, L., Benassi, M., Nicoletti, R., & Borghi, A. M. (2018). Abstract, emotional and concrete concepts and the activation of mouth-hand effectors. *PeerJ*, *6*, e5987.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547-559.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, *48*(7), 788-804.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at the International Conference on Learning Representations* (pp. 1-12).

Mnih, A., & Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems* (pp. 2265-2273).

Mohammad, S. M. (2018). 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).

Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, *31*(2), 301-326.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

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).

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

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303-345.

Sahlgren, M., & Lenci, A. (2016). The Effects of Data Size and Frequency Range on Distributional Semantic Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 975-980).

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).

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).

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).

Tai, K. S., Socher, R., & Manning, C. D. (2015). Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1556-1566).

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1555-1565).

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*(1), 141-188.

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.

Vinson, D., Ponari, M., & Vigliocco, G. (2014). How does emotional content affect lexical processing?. *Cognition & Emotion*, *28*(4), 737-746.

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.