Neural Activation Semantic Models: Computational lexical semantic models of localized neural activations

Nikos Athanasiou
School of ECE NTUA
Athens, Greece
el12074@central.ntua.gr

Elias Iosif *
IFF, University of Nicosia
Nicosia, Cyprus
iosif.e@unic.ac.cy

Alexandros Potamianos
School of ECE NTUA
Athens, Greece
apotam@gmail.com

Abstract

Neural activation models that have been proposed in the literature use a set of example words for which fMRI measurements are available in order to find a mapping between word semantics and localized neural activations. Successful mappings let us expand to the full lexicon of concrete nouns using the assumption that similarity of meaning implies similar neural activation patterns. In this paper, we propose a computational model that estimates semantic similarity in the neural activation space and investigates the relative performance of this model for various natural language processing tasks. Despite the simplicity of the proposed model and the very small number of example words used to bootstrap it, the neural activation semantic model performs surprisingly well compared to state-of-the-art word embeddings. Specifically, the neural activation semantic model performs better than the state-of-the-art for the task of semantic similarity estimation between very similar or very dissimilar words, while performing well on other tasks such as entailment and word categorization. These are strong indications that neural activation semantic models can not only shed some light into human cognition but also contribute to computation models for certain tasks.

1 Introduction

The human ability of translating concepts into words and back depends on the ability of mind to decode and encode meaning. This mental process, which is not currently completely understood, has captivated the interest of both neuroscientists and computational linguists (Haxby et al., 2001; Ishai et al., 1999; Cree and McRae, 2003; G. Kanwisher et al., 1997; Just et al., 2010). Specifically, when a person experiences a visual stimulus of a concept, reads, speaks or writes a word, particular neuronal regions in the brain are activated (Pulvermller, 2001).

Various studies have been carried out to explore brain encoding and decoding mechanisms when a stimulus is present, as detailed next. For visual stimuli, studies have shown that is feasible to discriminate and reconstruct images using patterns of neural activity, mainly found in the visual cortex (Kay et al., 2008; Thirion et al., 2006; Nishimoto et al., 2011; Naselaris et al., 2009; Miyawaki et al., 2008), the part of brain responsible for visual information processing. Other works have demonstrated the relationship between cognitive perception and speech (Benson et al., 2001; Formisano et al., 2008). Regarding textual stimuli, researchers have shown distributed semantic maps of words are present in our brains (Huth et al., 2012; Sudre et al., 2012).

Lexical semantics are based on the assumption that similar words appear in similar contexts (Harris, 1954). Based on that assumption, two different approaches for building semantic models have been proposed. The first approach is to encode word semantics, by applying dimensionality reduction of context-word occurrence matrix which was computed using large corpora (Deerwester et al., 1990; Bullinaria and Levy, 2012). The second approach replaces thse "counting" by predictive models (Baroni et al., 2014) based on neural networks (Bengio et al., 2000; Mnih and Hinton, 2007; Mikolov et al., 2013;

*The research was performed when the author was a postdoctoral researcher at School of ECE, NTUA in Athens, Greece.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/

Turian et al., 2010; Collobert and Weston, 2008). Counting models calculate and weight context vectors, while predictive models learn word vectors by guessing the context in which these words tend to appear.

In pursuance of enriching such lexical semantic models with cognitive information, as well as discovering the cognitive representation of word semantics, a number of studies have attempted to examine the mapping between semantic representation of computational and cognitive models. In prior work, it has been shown that semantic of words are related to activation potentials in regions of the brain and that decoding between neural activations and semantic content (Mitchell et al., 2008; Palatucci et al., 2009; Sudre et al., 2012; Murphy et al., 2012; Pulvermller, 2001) is possible. Furthermore, neural activations are shown to have predictive power with respect to semantics at the word (Mitchell et al., 2008; Jelodar et al., 2010) and sentence (Jing et al., 2017; Anderson et al., 2017) level. Computational studies that aim to explore the influence of neural activations in word representations have shown that by incorporating neural activations when training lexical semantic models can improve their generalization ability despite the small amount of neural activation data used (Fyshe et al., 2014; Ruan et al., 2016). These works show that a strong relationship exists between computational semantic models and neural representations. However, it remains to be seen how cognitive semantic representations, including localized neural activation patterns can help improve the performance of computational semantic models, especially for complicated classification and recognition tasks.

Motivated by the aforementioned studies that show correlation between localized neural activations and word semantics, we propose a computational model for semantic similarity that utilizes predicted neural activations learned from a small set of concrete nouns. The proposed model is applied to a variety of natural language processing tasks. The neural activation prediction model used here for lexical expansion is that proposed in (Mitchell et al., 2008). In our list of experiments, we first compare the performance of the proposed neural activation model for a concrete noun semantic similarity task and show that for certain word pairs it outperforms the state-of-the-art. Then we evaluate the performance of neural activation vectors for a word classification, sensory modality (sense) classification and textual entailment task. The fusion of neural and traditional word embedding vectors are shown to outperform the state-of-the-art. To our knowledge, this is the first time brain imaging data are successfully used for the aforementioned tasks.

2 Related Work

A significant body of literature investigates neural activations by mapping word semantics to fMRI data. Most of them have in common a basic idea published in (Mitchell et al., 2008). In this work, a model is introduced that maps low dimensional word coocurence vectors to neural activations. The approach is validated in a neural activation-based word classification task. This work shows that the mapping between lexical semantic spaces constructed via computational lexical semantic algorithms and 3D neural activations representations measured via fMRIs is possible.

A first variant of the aforementioned model was introduced in (Jelodar et al., 2010), where the use of WordNet features was investigated for constructing the lexical semantic space. Word classification results reported showed similar performance to (Mitchell et al., 2008), however, by fusion of the two lexical semantic models improved classification results where achieved.

A second approach, introduced in (Levy and Bullinaria, 2012), extends the work in (2008) by increasing the number of fMRI voxels used in the neural activation vectors and the number of featurs (dimension) of the lexical semantic model showing additional performance improvement.

Algorithms that count word coocurences and utilize hand-crafted features for constructing lexical semantic models can be found in (Bullinaria and Levy, 2012; Baroni and Lenci, 2010).Moreover, various lexical semantic models that predict a word based on its context have also elaborated (Bengio et al., 2000; Mikolov et al., 2013; Collobert and Weston, 2008). Word prediction models tend to perform better in natural language processing tasks such as analogy, similarity, synonym detection, concept taxonomy (Baroni et al., 2014) and sentiment analysis (Socher et al., 2011; Socher et al., 2013). However, their relationship with cognitive lexical representation is not yet well understood, at least to a degree that would allow us to improve current computation lexical semantic models. Along these lines, there have

been two main lines of work. In (Fyshe et al., 2014), neural activations were integrated in the training procedure of lexical semantic models in order to learn word embeddings that include latent neural information. Although a small number of words was used to bootstrap the neural activation representations, it have been shown that their model can predict unseen words and generalizes well across different topics. Ruan et al. (2016), have shown that neural activations for different parts of the brain are correlated with word embeddings especially skip-grams. A semantic model was also proposed for training word embeddings as a first step towards including cognitive information in a word vector representation.

3 Approach

3.1 Neural activations prediction

The neural activation prediction model used here is that proposed in Mitchell (2008)[1] First activation potentials are measured from fMRI images. We consider voxels to be 3D pixels created by MRI scanning software depicting brain state. Every voxel v is associated with a $TN \times V'$ array, M, of neural activation values (blood flow), where V' is the number of voxels, T is the number of trials, N is the number of stimuli, in our case different nouns. The first step is to select the most stable (salient) voxels, V to include in the neural activation model. The stability score s_v for voxel v is computed as the average pairwise Pearson correlation ϱ for all the different row combinations of M, as follows:

$$s_v = \frac{2}{TN(TN-1)} \sum_{i=1}^{TN} \sum_{j=i+1}^{TN} \varrho(M_{i,:}, M_{j,:}), \quad \forall\, v = 1 \cdots V' \tag{1}$$

where $M_{i,:}$ is the ith row of matrix M. High stability scores, s_v, as described in Eq. 1, indicate that corresponding voxels have consistent representations across different trials and nouns.

Next, the neural activation predictive model proposed in (Mitchell et al., 2008) is defined.

For this purpose we identify a set of m seed words s_1, s_2, ..., s_m, and a function $f_i(w)$ that estimates the association between seed word s_i and word w. The core assumption of the model is that words that are closely associated have similar neural activation patterns, thus the mapping from the associative (semantic) space to the neural activation (voxel) space is estimated as follows:

$$y_v(w) = \sum_{i=1}^{m} c_{v,i}\, f_i(w), \quad \forall\, v = 1 \cdots V, \tag{2}$$

where $y_v(w)$ is the activation of voxel v for word w, $f_i(w)$ is a scalar value that reflects the association between the i^{th} seed word s_i and the word w, m is the number of seed words (semantic features), V is the total number of voxels, and $c_{v,i}$ is a learned weight ranging between 0 and 1. A set of 25 verbs (seed words) was identified in (Mitchell et al., 2008) as semantic features s_i; seed words were manually selected according to psycholinguistic criteria. The similarity function $f_i(w)$ was set to the (normalized) co-occurrence frequency of the i^{th} seed word and w, estimated on a corpus. The weights $c_{v,i}$ were estimated using the fMRI data on words w using linear or ridge regression estimation. Once $c_{v,i}$ have been estimated, Eq. 2 can be used to predict the neural activation of unseen words[2].

3.2 From neural activations to similarity

Based on the hypothesis that similar words have similar neural activations, we propose a model to estimate word similarities based on neural activations predicted using Eq. 2. We evaluated various metrics for computing semantic similarity from neural activations. We present only a top performing metric

[1]The features in (2008) are attractive because of their simplicity and low-dimensionality, and generalize well for lexical expansion to a large lexicon compared to other works described in Section 2 that potentially perform slightly better.

[2]Although Eq. 3 can be used to perform lexical expansion an any token w for which $f_i(w)$ can be computed, the proposed framework (choice of $f()$ and associated fMRI data) is meant for concrete words and typically only neural activations for concrete words are reported in the literature.

formulated as the weighted square distance, namely:

$$S(w_1, w_2) = \sum_{v=1}^{V} b_v \ (y_v(w_1) - y_v(w_2))^2, \tag{3}$$

where $S(w_1, w_2)$ is the semantic similarity between words w_1 and w_2, V represents the number of voxels used in the predicted neural image, $y_v(w)$ is the activation of a voxel for word w, and b_v is a learned weight of the contribution of a particular voxel to the similarity metric. In Figure 1 the predictions

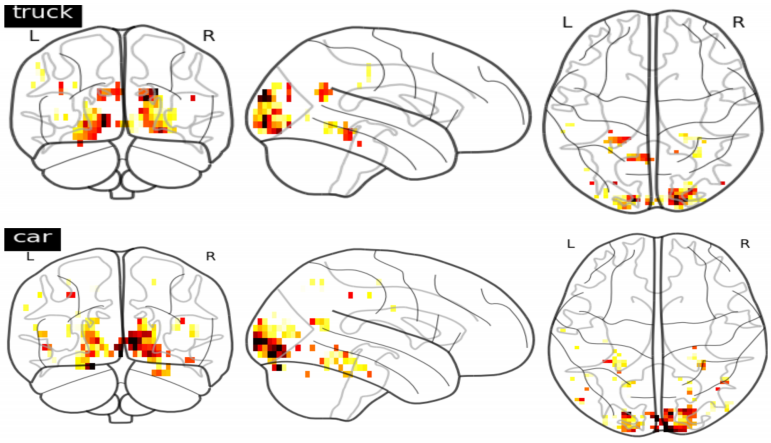

Figure 1: Neural activation images for two similar concrete nouns including the 500 most stable voxels (participant P1). This figure shows just one horizontal slice in Montreal Neurological Institute (MNI) space of the three-dimensional image.

of neural activations for two highly similar nouns in fMRI dataset are presented. Visualizations were created from 500 voxels to gather insight for our computational model. Observe that both brain images have similar neural activations both in terms of which parts of the brain are activated and the activation values. Although, we don't utilize localization information, our weighting schema implicitly detects activation patterns by variations in b_i coefficients.

4 Experimental Data and Preprocessing

We built the neural activations prediction model as in (Mitchell et al., 2008) using fMRI data for 60 concrete nouns. We calculate the semantic features $f_i(w)$ for each concrete noun, w, by counting its co-occurences with 25 manually selected verbs in a large corpus created in (Iosif et al., 2016) by aggregating results of web queries to Yahoo. The fMRI data used in our experiments were collected and processed by (Mitchell et al., 2008) and are publicly available. In this dataset, each one of the participants was presented 60 concrete nouns (for 6 times each) through a line drawing which was labeled with the corresponding noun. Each out of nine subjects, was asked to think about properties of the presented noun during scanning procedure. Finally, a vector representation of the whole cortex neural activation was extracted. Further details about the dataset can be found in (Mitchell et al., 2008) and its supplementary website[3]. Prior to training both training and test data are averaged across trials and the final neural activations of each noun are mean normalized. The following datasets have been used for semantic similarity, taxonomy creation, sense classification and entailment tasks. The code of the present work is publicly available[4].

MEN: For the semantic similarity task we train and evaluate our model on the MEN dataset (Bruni et al., 2014) which consists of 3000 word pairs (2000 for training set and 1000 pairs for test set). Each word pair is associated with a similarity score. This score was computed by averaging the similarities

[3] http://www.cs.cmu.edu/~tom/science2008/
[4] https://github.com/athn-nik/neural_asm

that provided by human annotators. We hand-labeled the dataset to keep only concrete nouns because the neural activation prediction model is trained only on concrete nouns. This resulted in 1114 pairs (562 unique words) in the training set and 524 pairs (438 unique words) in the test set. The similarity scores were normalized between zero and one. We also created 2 subsets of MEN of 39 highly similar and 79 highly dissimilar word pairs using a thresholding technique, where pairs with similarity score over 0.85 and under 0.1 belong in the first and second subset respectively.

ESSLLI: For the taxonomy creation task, we evaluate our model on the ESSLLI dataset (Baroni et al., 2008). It consists of a three-level hierarchy (2-3-6 classes). The lowest level of hierarchy contains 6 classes of concrete nouns (birds, land animals, fruit, greens, vehicles, tools), the middle 3 classes (vegetables, artifacts, animals) while the upper class is distinguished between living beings and objects.

Sensicon: For sense classification, we use the Sensicon (Tekiroglu et al., 2014) dataset. Sensicon is a dictionary which contains 22684 English words and associates each word with 5 numerical scores and a part of speech annotation. The scores correspond to the relevance of the word to each of the 5 senses, namely vision, hearing, taste, smell and touch. In order to use these scores for the sense classification task we selected nouns who have non-zero scores in only one sense results in 1011 words.

SNLI dataset: For the entailment task, we used the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015) which contains around 570K sentence pairs with three labels: entailment, contradiction and neutral. We preprocessed the initial dataset to keep only training and testing examples that have at least two or three concrete nouns that are also in the MEN dataset for both premise and hypothesis. This resulted in 30,498 training and 592 test samples for the case of at least three common words and 171,528 training and 3201 test samples for the case of at least two common words with MEN.

5 Experiments on Semantic Similarity

In this section we present baseline results on the neural activation prediction model of (Mitchell et al., 2008) and investigate the performance of the proposed neural activation semantic model of Eq. 3 for a semantic similarity task on the MEN dataset. Our goal here is first to reproduce the results in (Mitchell et al., 2008) and then to investigate the properties of neural activation embeddings semantic models compared to traditional embedding models, e.g., word2vec in (Mikolov et al., 2013).

5.1 Neural activations prediction

As proposed in (Levy and Bullinaria, 2012), we choose the 500 most stable voxels from the fMRI images. Then, $c_{v,i}$ is estimated as in Eq. 2 by applying linear regression per voxel across different words with regularization. To evaluate the neural activation prediction model we used cosine similarity in order to evaluate if our prediction for the possible pair of test words is correct or not. Correct prediction means that sum of the cosine similarities of the correct matched pairs is greater than the false matched pair as shown next:

$$cos(\vec{i_1}, \vec{p_1}) + cos(\vec{i_2}, \vec{p_2}) > cos(\vec{i_2}, \vec{p_1}) + cos(\vec{i_1}, \vec{p_2}), \tag{4}$$

where $\vec{i}$ is the actual image and $\vec{p}$ is the predicted image of 500 voxels. The dataset, which consists of 60 nouns, is split in train set (the rest 58 nouns) and test set (2 nouns) for all possible 1770 combinations using cross-validation.That process is followed for every one of 9 participants. The results are shown reported in Table 1 when using linear and ridge regression to estimate $c_{v,i}$ in Eq. 2. Results in Table 1 are on average consistent with the baseline results reported in (Mitchell et al., 2008). We achieved higher performance for some participants and lower for others. This can be attributed to the different tools we used to implement the system (as we used scikit-learn (Pedregosa et al., 2011) and (Mitchell et al., 2008) used Matlab). Moreover, we experimented with the number of voxels and our results agree with the findings reported in (Levy and Bullinaria, 2012).

Participant ID	Linear Regression	Ridge Regression	Mitchell et. al
1	0.79	0.84	0.83
2	0.75	0.82	0.76
3	0.63	0.76	0.78
4	0.63	0.79	0.72
5	0.61	0.78	0.78
6	0.58	0.65	0.85
7	0.58	0.75	0.73
8	0.65	0.68	0.68
9	0.57	0.68	0.82
Mean	0.64	0.75	0.77

Table 1: Baseline Model Results

5.2 Similarity task

For the semantic similarity task, we applied Eq. 3 for the word pairs of the MEN dataset. The $y_v(.)$s of Eq. 3 were computed using Eq. 2 utilizing up to 250 voxels. We exploited the training subset of MEN for learning the b weights of Eq. 3 using linear regression. Those weights were used for computing the similarities for the test subset of MEN. The Spearman correlation coefficient between the human similarity scores (ground truth) and the similarity scores computed by Eq. 3 was used as evaluation metric. The results are presented in Table 2, where we compare the performance of the proposed neural model (averaged across participants) against the performance yielded by the w2vec word embeddings (Mikolov et al., 2013) trained on the GoogleNews corpus.

Subset	**Number of voxels**	**Neural model**	**w2vec**
All Concrete nouns	50	0.43	0.76
	100	0.47	0.76
	150	0.48	0.76
	200	0.48	**0.76**
Most & Least similar	50	0.58	0.73
	100	0.82	0.73
	150	0.82	0.73
	200	**0.88**	0.73
Least similar	50	0.43	0.43
	100	0.44	0.43
	150	0.47	0.43
	200	**0.63**	0.43
Most similar	50	0.28	0.14
	100	0.63	0.14
	150	0.68	0.14
	200	**0.83**	0.14

Table 2: Evaluation results on the concrete nouns subset of the MEN test set, and on most and least similar concrete word subsets.

Overall, the w2vec model outperforms the neural model achieving 0.76 correlation on all concrete nouns. For the neural model, performance increases as more voxels are exploited reaching 0.48 correlation is obtained for at least 150 voxels. In Table 2, the performance is also shown for three subsets of the MEN test set, namely, "Most & Least similar", "Least similar" and "Most similar" concrete nouns.

For all three subsets, the performance achieved by the neural model exceeds[5] the performance of w2vec when at least 100 voxels are used. The performance improvement becomes more pronounced as the number of voxels increases. The best correlation score achieved is 0.88 for the case of the "Most & Least similar" subset for 200 voxels, outperforming the w2vec model (0.73 correlation). Especially for the case of the "Most similar" evaluation subset, we observe a remarkable difference between the two models, i.e., 0.83 vs. 0.14 for the neural and the w2vec model, respectively.

6 Classification and Entailment Experiments

Next we present the performance of the neural activation semantic model for a taxonomy creation (semantic class classification), sensory modality (sense) classification, and lexical entailment task. Neural vectors are averaged across participants and evaluated both standalone and in combination (early or late fusion) with traditional word embedding models.

6.1 Taxonomy Creation

The performance of similarities computed by Eq. 3 were also evaluated on a taxonomy creation task on the ESSLLI dataset (Baroni et al., 2008). Taxonomy creation is performed using the neural activation vectors $\vec{y}(w)$ estimated from Eq. 2 and the coefficient vectors $\vec{b}$ defined in Eq. 3 trained using linear regression on the whole of the MEN dataset. Then, the similarity matrix $S(w_i, w_j)$ is estimated for all pairs in the dataset using Eq. 3 and then the spectral clustering algorithm proposed in (Ng et al., 2001) is applied to obtain the lexical classes. In this work, neural fusion refers to early fusion (vector concatenation) of word vectors and neural activation vectors. We used a purity-based metric for evaluating the quality of the automatically created clusters. The purity P of the taxonomy is defined as in (Baroni and Lenci, 2010):

$$P = \frac{1}{d} \sum_{i=1}^{k} \max_{j}(e_{ij}), \tag{5}$$

where e_{ij} is the number of nouns assigned to the ith cluster that belong to the jth groundtruth class, k is the number of clusters, and d is the total number of concrete nouns included in the dataset. Purity expresses the fraction of nouns that belong to the true class, which is most represented in the cluster, taking values in the range [0,1].

Dataset	Neural Model	w2vec	Neural Fusion
ESSLI (6 classes)	0.61	0.70	**0.71**
ESSLI (3 classes)	0.77	0.95	**0.95**
ESSLI (2 classes)	0.66	**0.77**	0.72

Table 3: Evaluation results for taxonomy creation.

The evaluation results are presented in Table 3 for the neural activation model, the w2vec word embeddings (Mikolov et al., 2013) trained on the GoogleNews corpus and the late fusion of the two models (denoted as neural fusion) with equal weighting of their similarity matrices S. All results shown are computed on $V = 250$ voxels. The neural model performs worse that the w2vec model in all three subtasks (6, 3 or 2 classes), however, the proposed neural fusion achieves the best purity scores for 6 and 3 classes, at 0.71 and 0.95 purity, respectively.

6.2 Sense Classification

For the sensory modality (sense) classification task we use the Sensicon dataset to evaluate the performance of our model regarding sense discrimination. By definition all nouns in Sensicon are concrete nouns since they are associated with a real-world sensory stimulus. Sense classification is performed as

[5]The differences between the similarity scores estimated by our model and the baseline (i.e., w2vec) were found to be statistically significant at 99% level according to paired-sample t-test.

described in the previous section, i.e., the similarity matrix in Eq. 3 is calculated using the weight vector $\vec{b}$ trained on the MEN dataset and then the spectral clustering (Ng et al., 2001) is applied for the five sense categories. The resulting clusters are used for sense classification either between two senses, one versus all or among all five senses.

Classes	Neural Model	w2vec	Neural Fusion
Vision, Audition	0.55	0.55	**0.57**
Vision, Touch	0.68	0.66	**0.69**
Vision, Taste	0.60	0.60	**0.61**
Audition, Touch	0.59	0.58	**0.59**
Audition, Taste	0.57	0.55	**0.57**
Taste, Touch	0.54	0.54	0.54
Vision, Other	0.68	0.68	0.68
Audition, Other	0.74	0.74	0.74
Touch, Other	0.81	0.81	0.81
Taste, Other	0.78	0.78	**0.79**
Audition, Vision, Smell, Touch, Taste	0.37	0.33	**0.39**

Table 4: Evaluation results for two-way, one-vs-all, and five-way sense classification.

The evaluation results[6] are presented in Table 4 for the neural model, the w2vec model (same as the one used in the previous section) and the late fusion of the two (neural fusion) using equal weighting on the similarity matrices S. The evaluation metric used is the purity of clusters defined in Eq. 5. All results shown are computed on $V = 250$ voxels.

The neural and w2vec models achieve very similar results for two-way classification tasks, with the neural model performing better 0.37 vs 0.33 for five-way sense classification. The neural fusion model outperforms both neural and w2vec models for the majority of the two-way classification tasks and also achieves top performance for the five-way classification task reaching 0.39 purity score. Overall, the neural and neural fusion models show strong performance for this task, which is reasonable given the localization of sensory representations in the human cortex.

These results are also consistent with neuroscientific research (Heed, 2010; Marieb and Hoehn, 2007; Kobayashi, 2006; Ungerleider, 1982; Pickles, 1988; Buck and Axel, 1991).

6.3 Entailment Task

Next, we applied the neural activations to an entailment classification task. We used a Bi-LSTM model proposed in (Rocktäschel et al., 2015) featuring contextual attention (see Figure 2) as our baseline model. Word embeddings for concrete nouns were estimated using GloVe as detailed in (Pennington et al., 2014) and used as input to the Bi-LSTM network. The neural activations vectors were then combined via early fusion (vector concatenation) with GloVe embeddings (Pennington et al., 2014). Evaluation results for GloVe vectors and neural fusion are shown in Table 5 in terms of prediction accuracy. The results are

Dataset(SNLI)	Dimensions (GloVe, neural)	GloVe	Neural Fusion
3-common	(300,250)	68.2	**68.7**
2-common	(300,250)	76.6	**77.7**

Table 5: Entailment task accuracy for GloVe and neural fusion vector input to the Bi-LSTM.

reported for two subsets of the SNLI dataset, namely, 3-common and 2-common (see section 4).

We observe that the top accuracy is achieved by the fusion scheme for both the 3-common and 2-common subsets ($68.7 \pm 0.9\%$ and $77.7 \pm 0.9\%$).

[6]Note that the sense smell is not always present in Table 4 because smell has only four nouns compared in to other senses that contain more than 100 nouns each.

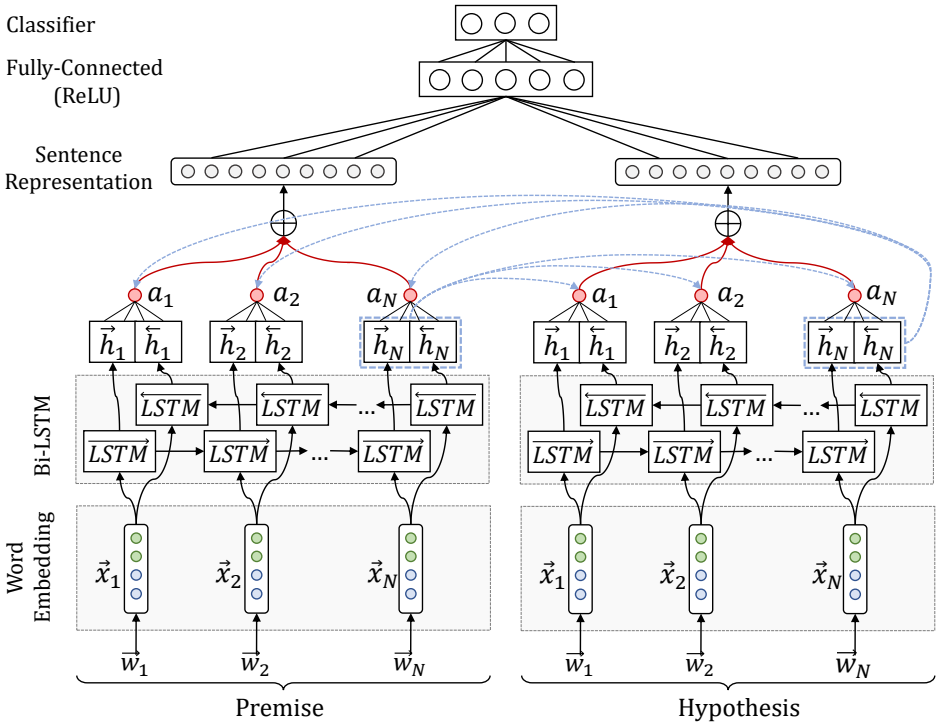

Figure 2: Bi-LSTM with context attention used in our experiments. Words' representations are either pretrained word embeddings or fusion of neural activations and word embeddings.

7 Conclusion

We proposed a simple neural activation semantic model extending the work of (Mitchell et al., 2008). The performance of the neural model was investigated for the tasks of word semantic similarity, taxonomy creation, sensory modality classification and concept entailment.

The analysis of the neural model word performance showed that the proposed model can differentiate between very similar and very dissimilar concrete nouns better than state-of-the-art word embeddings semantic models, while it performs worse overall for the word semantic similarity task. This is a strong indication that the semantic discriminability of neural activation vectors is of a different flavor than that of traditional word embedding vectors, and thus neural vectors can be used to augment state-of-the-art semantic representations. Results on the taxonomy classification, sense classification and entailment task indeed verify the different flavor of neural embeddings. For certain tasks, e.g., sense classification, neural models provide state-of-the-art performance. For other tasks, the fusion of neural and word2vec embeddings provides significant improvement. Overall (predicted) localized neural activation vectors can also be used in conjunction with other semantic representations and deep architectures to improve the results in challenging tasks, like concept entailment.

Although the collection of neuroimaging data has many limitations such as variation across participants, high signal-to-noise ratio and the need of expensive equipment for data capture, it provides an alternative view of how lexical and sensory information in localized in the human brain. Despite the very small dataset used in our experiments, results are encouraging about the value of neural activations patterns for computational tasks. In future work, we will investigate the relative performance of neural activations for a variety of natural language processing tasks, how to build neural vector predictors for the whole dictionary via lexical expansion, as well as more efficient fusion algorithms of neural and traditional word embedding models.

Acknowledgements. This work has been partially supported by EU H2020 BabyRobot project (grant #687831). Also, the authors would like to thank Christos Baziotis and Georgios Paraskevopoulos for many useful discussions.

References

[Anderson et al.2017] Andrew James Anderson, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Mario Aguilar, Xixi Wang, Donias Doko, and Rajeev D. S. Raizada. 2017. Predicting neural activity patterns associated with sentences using a neurobiologically motivated model of semantic representation. *Cerebral Cortex*, 27(9):4379–4395.

[Baroni and Lenci2010] Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

[Baroni et al.2008] M Baroni, S Evert, and A Lenci. 2008. Bridging the gap between semantic theory and computational simulations. In *Proc. of ESSLLI Distributional Semantic Workshop*.

[Baroni et al.2014] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

[Bengio et al.2000] Y Bengio, Rjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:932–938, 01.

[Benson et al.2001] Randall R. Benson, D.H. Whalen, Matthew Richardson, Brook Swainson, Vincent P. Clark, Song Lai, and Alvin M. Liberman. 2001. Parametrically dissociating speech and nonspeech perception in the brain using fmri. *Brain and Language*, 78(3):364 – 396.

[Bowman et al.2015] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

[Bruni et al.2014] Elia Bruni, N Tram, Marco Baroni, et al. 2014. Multimodal distributional semantics. *The Journal of Artificial Intelligence Research*, 49:1–47.

[Buck and Axel1991] Linda Buck and Richard Axel. 1991. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell*, 65(1):175 – 187.

[Bullinaria and Levy2012] John A. Bullinaria and Joseph P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior Research Methods*, 44(3):890–907, Sep.

[Collobert and Weston2008] Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

[Cree and McRae2003] George S Cree and Ken McRae. 2003. Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, 132(2):163.

[Deerwester et al.1990] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

[Formisano et al.2008] Elia Formisano, Federico De Martino, Milene Bonte, and Rainer Goebel. 2008. "who" is saying "what"? brain-based decoding of human voice and speech. *Science*, 322(5903):970–973.

[Fyshe et al.2014] Alona Fyshe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2014. Interpretable semantic vectors from a joint model of brain- and text- based meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 489–499, Baltimore, Maryland, June. Association for Computational Linguistics.

[G. Kanwisher et al.1997] N G. Kanwisher, Josh Mcdermott, and Marvin M. Chun. 1997. The fusiform face area: A module in human extrastriate cortex specialized for face perception. 17:4302–11, 07.

[Harris1954] Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

[Haxby et al.2001] James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430.

[Heed2010] Tobias Heed. 2010. Touch perception: How we know where we are touched. *Current Biology*, 20(14):R604 – R606.

[Huth et al.2012] AlexanderG. Huth, Shinji Nishimoto, AnT. Vu, and JackL. Gallant. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210 – 1224.

[Iosif et al.2016] Elias Iosif, Spiros Georgiladakis, and Alexandros Potamianos. 2016. Cognitively motivated distributional representations of meaning. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

[Ishai et al.1999] Alumit Ishai, Leslie G Ungerleider, Alex Martin, Jennifer L Schouten, and James V Haxby. 1999. Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 96(16):9379–9384.

[Jelodar et al.2010] Ahmad Babaeian Jelodar, Mehrdad Alizadeh, and Shahram Khadivi. 2010. Wordnet based features for predicting brain activity associated with meanings of nouns. In *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*, CN '10, pages 18–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Jing et al.2017] Wang Jing, Cherkassky Vladimir L., and Just Marcel Adam. 2017. Predicting the brain activation pattern associated with the propositional content of a sentence: Modeling neural representations of events and states. *Human Brain Mapping*, 38(10):4865–4881.

[Just et al.2010] Marcel Adam Just, Vladimir L. Cherkassky, Sandesh Aryal, and Tom M. Mitchell. 2010. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLOS ONE*, 5(1):1–15, 01.

[Kay et al.2008] Kendrick Kay, Thomas Naselaris, Ryan J Prenger, and Jack Gallant. 2008. Identifying natural images from human brain activity. 452:352–5, 04.

[Kobayashi2006] Masayuki Kobayashi. 2006. Functional organization of the human gustatory cortex. *Journal of Oral Biosciences*, 48(4):244–260.

[Levy and Bullinaria2012] Joseph P Levy and John A Bullinaria. 2012. Using enriched semantic representations in predictions of human brain activity. *Connectionist Models of Neurocognition and Emergent Behavior: From Theory to Applications. Singapore: World Scientific*, pages 292–308.

[Marieb and Hoehn2007] Elaine Nicpon Marieb and Katja Hoehn. 2007. *Human anatomy & physiology*. Pearson Education.

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

[Mitchell et al.2008] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

[Miyawaki et al.2008] Yoichi Miyawaki, Hajime Uchida, Okito Yamashita, Masa aki Sato, Yusuke Morito, Hiroki C. Tanabe, Norihiro Sadato, and Yukiyasu Kamitani. 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915 – 929.

[Mnih and Hinton2007] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648. ACM.

[Murphy et al.2012] Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 114–123, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Naselaris et al.2009] Thomas Naselaris, Ryan J Prenger, Kendrick Kay, Michael Oliver, and Jack Gallant. 2009. Bayesian reconstruction of natural images from human brain activity. 63:902–15, 09.

[Ng et al.2001] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press.

[Nishimoto et al.2011] Shinji Nishimoto, AnT. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and JackL. Gallant. 2011. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641 – 1646.

[Palatucci et al.2009] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418. Curran Associates, Inc.

[Pedregosa et al.2011] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

[Pickles1988] James O Pickles. 1988. *An introduction to the physiology of hearing*, volume 2. Academic press London.

[Pulvermller2001] Friedemann Pulvermller. 2001. Brain reflections of words and their meaning. *Trends in Cognitive Sciences*, 5(12):517 – 524.

[Rocktäschel et al.2015] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

[Ruan et al.2016] Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu. 2016. Exploring semantic representation in brain activity using word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 669–679, Austin, Texas, November. Association for Computational Linguistics.

[Socher et al.2011] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 151–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Socher et al.2013] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

[Sudre et al.2012] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62(1):451 – 463.

[Tekiroglu et al.2014] Serra Sinem Tekiroglu, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An automatically constructed sensorial lexicon. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1511–1521, Doha, Qatar, October. Association for Computational Linguistics.

[Thirion et al.2006] Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. 2006. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104 – 1116.

[Turian et al.2010] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 384–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Ungerleider1982] Leslie G Ungerleider. 1982. Two cortical visual systems. *Analysis of visual behavior*, pages 549–586.