Non-Complementarity of Information in Word-Embedding and Brain Representations in Distinguishing between Concrete and Abstract Words

Kalyan Ramakrishnan
Indian Institute of Technology Madras
kalyan0821@yahoo.com

Fatma Deniz[†‡]
[†]University of California, Berkeley
[‡]Electrical Engineering and Computer Science
Technische Universität Berlin
fatma@berkeley.edu

Abstract

Word concreteness and imageability have proven crucial in understanding how humans process and represent language in the brain. While word-embeddings do not explicitly incorporate the concreteness of words into their computations, they have been shown to accurately predict human judgments of concreteness and imageability. Inspired by the recent interest in using neural activity patterns to analyze distributed meaning representations, we first show that brain responses acquired while human subjects passively comprehend natural stories can significantly distinguish the concreteness levels of the words encountered. We then examine for the same task whether the additional perceptual information in the brain representations can complement the contextual information in the word-embeddings. However, the results of our predictive models and residual analyses indicate the contrary. We find that the relevant information in the brain representations is a subset of the relevant information in the contextualized word-embeddings, providing new insight into the existing state of natural language processing models.

1 Introduction

Language comprises concrete and abstract words that are distinctively used in everyday conversations. Concrete words refer to entities that can be easily perceived with the senses (e.g., "house", "blink", "red"). On the other hand, abstract words refer to concepts that one cannot directly perceive with the senses (e.g., "luck", "justify", "risky"), but relies on the use of language to understand them (Brysbaert et al., 2014).

This categorization of words based on their concreteness is rooted in theoretical accounts in cognitive science. One such account is the Dual Coding Theory (Paivio, 1971, 1991), according to which two separate but interconnected cognitive systems represent word meanings, i.e., a non-verbal system that encodes *perceptual* properties of words and a verbal system that encodes *linguistic* properties of words. Concrete concepts can be easily imagined and are represented in the brain with both verbal and non-verbal codes. Abstract concepts are less imaginable and are represented with only verbal codes. For example, one can readily picture as well as describe the word *bicycle* (e.g., "has a chain", "has wheels"), but relies more on a verbal description for the word *bravery*.

The concreteness of words has since been used as a differentiating property of word meaning representations. Previous studies in natural language processing (NLP) have examined the word-embedding spaces of concrete and abstract words and showed: (i) distinct vector representations of the two categories within and across languages (Ljubešić et al., 2018), and (ii) high predictability of concreteness scores from pre-trained word-embeddings (Charbonnier and Wartena, 2019).

Neurolinguistic studies have shown an extensive, distributed network of brain regions representing the conceptual meaning of words (Mitchell et al., 2008; Wehbe et al., 2014; Huth et al., 2016). Among these, regions more closely involved in sensory processing have been shown to respond favorably to concrete words (Binder et al., 2005) over abstract words. Hill et al. (2014) argued that concrete and abstract concepts must be represented differently in the human brain by showing through a statistical analysis that concrete concepts have fewer but stronger associations in the mind with other concepts, while abstract concepts have weak associations with several other concepts.

Wang et al. (2013) showed that functional Magnetic Resonance Imaging (fMRI) signals of brain activity recorded as subjects attempted to decide which two out of a triplet of words were most similar contained sufficient information to classify the concreteness level of the word triplet, providing

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 1–11

further evidence of the dissimilar representations of the two categories in the brain. However, it remains an open question whether the brain responses within the semantic system can directly predict concreteness levels in the more challenging setting of *naturalistic* word stimuli (e.g., words encountered while reading a story). Moreover, given the human brain's expertise in generating and processing *perceptual* as well as *linguistic* information, one could expect the brain representations to provide information that complements the word-embeddings purely learned from linguistic contexts, improving their predictive capability. We address both these questions in this paper.

While several related works exist, the following limitations prompted a new study: (i) Anderson et al. (2017) indirectly decoded the brain representations for concrete and abstract nouns with the help of word-embeddings and convolutional neural network image representations. Instead of building a predictive model, the authors used a similarity metric to determine which signal in a pair of fMRI signals corresponds to which word in a pair of words. However, a direct, supervised decoding approach (as adopted here) would provide more substantial evidence about the strengths and weaknesses of the different information modalities. (ii) Brysbaert et al. (2014) found word concreteness scores to be highly correlated with both *visual* and *haptic* perceptual strength. However, multi-modal methods (Anderson et al., 2017; Bhaskar et al., 2017) have incorporated only visual features (as the second source of information) instead of general *perceptual* features into their predictions. By incorporating brain representations in our models, we do not miss out on such perceptual information (e.g., the adjectives "silky", "crispy", and "salty" are concrete but not as imagery-inducing as the adjective "blue"). (iii) In contrast to previous studies that have required participants to actively imagine a randomly presented word stimulus[1] (before being given a few seconds to "reset" their thoughts) during the brain data acquisition task (Anderson et al., 2012; Wang et al., 2013; Anderson et al., 2017), we adopt a task where participants would read highly engaging natural stories (without unnatural pauses), enabling them to process the word stimuli in a more realistic context.

To summarize, our objectives with this paper are twofold. First, we investigate how well human brain representations can predict the concreteness levels of words encountered in natural stories using simple, supervised learning algorithms. Second, we investigate whether brain representations encode information that may be missing from word-embeddings trained on a text corpus in making the concrete/abstract distinction. We believe that answering such questions will shed light on the current state of human and machine intelligence and on the ways to incorporate human language processing information into NLP models.

2 Related Work

A few studies have shown that the concreteness (and imageability) of words can be directly predicted with high accuracy from precomputed word-embeddings using supervised learning algorithms. Recently, Charbonnier and Wartena (2019) used a combination of word-embeddings and morphological features to predict the word concreteness and imageability values provided in seven publicly available datasets. Ljubešić et al. (2018) extended the idea to perform a cross-lingual transfer of concreteness and imageability scores by exploiting pre-trained bilingual aligned word-embeddings (Conneau et al., 2017).

Multi-modal models that use both linguistic and perceptual information have been shown to outperform language models at various NLP tasks, such as learning concrete or abstract word embeddings (Hill and Korhonen, 2014; Lazaridou et al., 2015), concept categorization (Silberer and Lapata, 2014), and compositionality prediction (Roller and Schulte im Walde, 2013). However, Bhaskar et al. (2017) found that the concreteness of nouns could be predicted equally well from the textual, visual, and combined modalities. This suggests that the textual and visual modalities independently provided reliable, non-complementary information to represent both concrete and abstract nouns.

Several studies have addressed the idea of decoding neural activity patterns recorded in subjects when presented with certain textual or visual stimuli. Anderson et al. (2017) applied linguistic and visually-grounded computational models to decode the fMRI representations of a set of concrete and abstract nouns. They, too, reported no decoding advantage for multi-modal combinations over the linguistic model. Anderson et al. (2012) demonstrated that fMRI signals contained sufficient information to perform a 7-way classification of a

[1] e.g., one word would be presented every $10s$.

set of words into WordNet-based (Miller, 1995) taxonomic categories.

Lately, there has been an increasing research interest at the intersection of neuroimaging and language models (Jain and Huth, 2018; Abnar et al., 2019; Gauthier and Levy, 2019; Hollenstein et al., 2019; Toneva and Wehbe, 2019; Jain et al., 2020; Caucheteux and King, 2020; Schrimpf et al., 2020). In an interesting study, Schwartz et al. (2019) finetuned the BERT language model to predict the fMRI responses of text-reading participants to obtain representations that encode brain-activity-relevant semantic information. While the modified representations could better predict neural activity and even generalize to new participants, this inclusion of brain-relevant bias *did not* improve or degrade the model's performance on downstream NLP tasks.

3 Data Collection

3.1 Stimulus and fMRI data

We briefly describe the functional Magnetic Resonance Imaging (fMRI) data-collection procedure here and refer the reader to Deniz et al. (2019) for specific details.

Nine participants were asked to read 11 autobiographical narrative stories taken from *The Moth Radio Hour* podcast. We used six participants' data in our experiments. The stories are each 10-15 minutes long and were chosen to cover a wide range of topics. Each story was first aligned to its transcript by applying the UPenn Forced Aligner (Yuan and Liberman, 2008) and Praat (Boersma and Weenink, 2001) on the narration audio. Timestamps for word-occurrences were then obtained from Praat's TextGrid as a list of entries of the form (w_i, t_i) representing the ith word and its onset time, respectively. Using this *word-representation* list for each story, each word in the story was displayed one-by-one at the center of a screen for a duration equal to its duration in the spoken version.

Each fMRI scan consists of a sequence of voxel-responses[2] acquired at a fixed repetition-time ($TR = 2.0045s$) with a voxel-size of $2.24 \times 2.24 \times 4.1mm$. A separate scan was conducted for each subject and presented story (all analysis was done within subjects). The acquired volumetric fMRI responses for each subject were first preprocessed to correct for motion and then aligned to the first scan's temporal average, using the FMRIB Linear Image Registration Tool (FLIRT) from FSL v5.0 (Jenkinson et al., 2002; Jenkinson and Smith, 2001). A Savitzky–Golay filter (Schafer, 2011) with a $120s$ window was applied to remove low-frequency voxel-response drift from the signal. Finally, the voxel-responses for each story were z-scored separately so that they have zero mean and unit variance across all acquisitions for the story.

We note that an equivalent analysis could be carried out through a listening task since the elicited brain representations have been shown to be largely invariant to the stimulus modality (Deniz et al., 2019).

3.2 Concreteness Ratings

We used the dataset collected by Brysbaert et al. (2014), consisting of concreteness ratings for 39,954 English words. Each word was rated by around 25 participants (recruited through Amazon Mechanical Turk) on a 1-5 scale so that the most concrete words are assigned the highest score of 5, and the most abstract words are assigned the lowest score of 1. For each word, the average rating (and standard deviation) across all raters was recorded.

3.3 Word-Embeddings

We extracted the 768-dimensional activations from the final hidden layer of the Generative Pre-trained Transformer (GPT-2) (Radford et al., 2019) to obtain contextualized representations for the words in the stories. The reasons for selecting GPT-2 in this work are due to the findings of Schrimpf et al. (2020). First, GPT-2 was constrained to use unidirectional attention in the same way humans process text in a left-to-right fashion. Second, the authors find that models best matching human language processing are precisely those trained for a *next* word prediction objective (such as the GPT family).

4 Data Preparation

Rating and Vectorizing Using the word-representation for each story and a list of the fMRI acquisition-times (identical for all subjects), we partitioned the words into disjoint *chunks* so that all words in a chunk correspond to the same acquisition. Therefore, all words read by the subjects within a duration of 1 TR from the start of the acquisition pulse were included in the same chunk.

[2] voxel = volumetric pixel.

We used GPT-2 to vectorize each word in a story by supplying all words in the story leading up to it[3] as context and extracting the network's hidden layer representation corresponding to the last input position. To rate the words in the story, we first lowercased and lemmatized them and then used the Brysbaert et al. (2014) concreteness dataset to assign a rating to each word in a chunk. Only around 7% of all words in the stories were not covered by the dataset and were dropped before subsequent analysis.

We stored the ith preprocessed functional image of each subject as an N_b-dimensional *voxel-response vector* $\vec{b}_i$, where N_b denotes the number of voxels for that subject's brain. Typical values for N_b were found to lie in the 70k-90k range (with a mean of 80976 and a standard deviation of 6173, across subjects).

Downsampling Since the rate at which the text stimulus was presented to the subjects (the narration rate) is higher than the rate of fMRI data acquisition (2.0045s per acquisition), several words may occur within the TR corresponding to a single acquisition and will all fall under the same chunk. Therefore, we downsampled the stimulus to match the acquisition rate before further analysis by averaging out the concreteness ratings (r_w) and word-embeddings ($\vec{e}_w$) within each TR. Thus, the *chunk-rating* and *chunk-embedding* for chunk C_i are given by:

$$r_i = \frac{1}{|C_i|} \sum_{w \in C_i} r_w$$

$$\vec{e}_i = \frac{1}{|C_i|} \sum_{w \in C_i} \vec{e}_w$$

Stacking We temporally stacked the voxel-response vectors, chunk-embeddings and chunk-ratings, first within each story and then across all 11 stories to obtain (i) a per-subject *voxel-response matrix* $B \in \mathbb{R}^{T \times N_b}$, (ii) an *embedding matrix* $E \in \mathbb{R}^{T \times D}$, and (iii) a *rating vector* $\vec{r} \in \mathbb{R}^T$, where T denotes the total number of fMRI acquisitions across all stories per subject, and D denotes the dimensionality of the word-embedding space. $D = 768$ for GPT-2, and 11 stories with an average duration close to 12.5 min per story gives $T = 4028$.

[3] or as many as allowed by the model's capacity.

5 Predictive Models

5.1 Word-Embedding based model

We consider the task of classifying words as *concrete* or *abstract* (based on their concreteness ratings) using the word-embeddings (chunk-embeddings, $\vec{e}_i$) as explanatory variables. For this, we first defined a *concreteness threshold* τ as follows: any word is labeled *concrete* if its assigned rating is strictly greater than τ, and is labeled *abstract* otherwise. We take $\tau = 3$.

We then segregated the data into *well-defined* classes by discarding any chunks that were found to consist of a mixture of concrete and abstract words (as defined above). This retains roughly 42% of all chunks ($T^s < T$), resulting in the following *strict* counterparts to the embedding matrix and rating vector obtained in Section 4: (i) $E^s \in \mathbb{R}^{T^s \times D}$, and (ii) $\vec{r}^s \in \mathbb{R}^{T^s}$, with the superscript ^{s} denoting that only chunks satisfying the strictly concrete/abstract property are being considered. We binary-encoded $\vec{r}^s$ into the boolean vector $\vec{y}^s \in \{0,1\}^{T^s}$, so that $y^s_i = 1$ if the corresponding chunk is strictly concrete and $y^s_i = 0$ otherwise. Our specific choice for the concreteness threshold ($\tau = 3$) produces a dataset that is approximately balanced between the two classes and is a natural choice for a 1-5 scale.[4]

We learned the $E^s \rightarrow \vec{y}^s$ mapping for each subject through $L2$-regularized logistic regression. We trained on 75% of the available data and picked the best value for the regularization parameter C through 5-fold cross-validation. We report the accuracy, recall, and F1 score of the classifier in our results.

An important variable in cognitive processing is the frequency with which words are encountered in language. High-frequency words are often perceived and processed faster than low-frequency words (van Heuven et al., 2014). Thus, word frequency could be a confounding variable to our objective if its distribution over the concrete words significantly differs from its distribution over the abstract words encountered in the stories. To check if this is the case, we computed the distribution of SUBTLEX-US (Brysbaert and New, 2009) word frequencies separately over all concrete vs. abstract words encountered by the subjects. However, a Kolmogorov-Smirnov test showed that the computed distribution over the concrete words was *not*

[4] Out of all strictly concrete/abstract chunks, 52% were labeled concrete, and 48% were labeled abstract.

significantly different from the distribution over the abstract words ($ks = 0.056, p = 0.063$).

5.2 Voxel-Response based model

Voxel Selection With up to 90,000 voxel-responses recorded per fMRI acquisition, not all voxels may be relevant to our objective of predicting the concreteness of word stimuli (Binder et al., 2005).

A standard voxel selection method is to manually determine regions of interest (ROIs) in the brain by analyzing the fMRI responses recorded in an auxiliary functional localizer task (Fedorenko et al., 2010) and select voxels from only these regions. However, this comes at the risk of being too restrictive. For example, one might inadvertently exclude regions in the brain encoding relevant sensory processing information in favor of regions encoding linguistic information. Given our objective to investigate whether brain representations contain any such additional information over word-embeddings, we avoided ROI-based methods for voxel selection.

We instead selected voxels based on their fractions of potentially-explainable response variance across time steps. This may be estimated separately for each voxel by recording different versions of its (time-varying) response corresponding to repeated presentations (Hsu et al., 2004) of the same stimulus-sequence. Assume that one story is repeatedly presented N times to a given subject and b represents a voxel being analyzed. If $b_t^{(n)}$ represents its response at time step t corresponding to the nth repetition, then its mean response across repetitions is $b_t = \frac{1}{N}\sum_{m=1}^{N} b_t^{(m)}$. The following equations estimate the fraction of potentially-explainable variance for b assuming the voxel-responses are z-scored across all time steps for the story:

$$ev(b) = \frac{1}{N}\sum_{n=1}^{N}[1 - \underset{t}{Var}(b_t^{(n)} - b_t)]$$

$$\overline{ev}(b) = ev(b) - \frac{1}{N-1}(1 - ev(b))$$

Thus, $\overline{ev}(b)$ is analogous to the adjusted R^2 of a (perfect) model that always predicts the mean response (b_t) across repetitions. A larger value indicates that the voxel responds consistently to repetitions of the same stimulus. Each subject was presented the last story $N = 2$ times, and the top-V voxels with the highest $\overline{ev}$ values were retained. From this, we obtain the desired reduced form $\hat{B} \in \mathbb{R}^{T\times V}$. The optimal number of semantic voxels V was chosen separately for each subject to maximize performance on the validation set (described next).

Prediction Task Blood-oxygen-level-dependent (BOLD) signals in the brain typically persist for 8-10s after stimulus onset (Ashby, 2019). Since each chunk covers nearly $2s$ of stimulus presentation, we expect the response to each chunk to be jointly encoded by the first, second, third, and fourth (reduced) voxel-response vectors that *follow* the current acquisition. However, including the first or fourth acquisition significantly degraded predictive performance. We posit that this degradation occurs because the voxel-response vectors recorded one or four TRs after the current acquisition are more prone to be directly affected by words falling in chunks preceding or succeeding the chunk of interest.

With this observation, we modeled the brain's representation of the stimulus in chunk C_i to be of the form $f(\hat{\vec{b}}_{i+2}, \hat{\vec{b}}_{i+3})$, where $\hat{\vec{b}}_{i'}$ represents the reduced voxel-response vector from the i'th acquisition. We therefore constructed the *reduced+delayed* voxel-response matrix $\hat{B}^{+} \in \mathbb{R}^{T\times 2V}$ by replacing each row of $\hat{B}$ with the concatenation of the *second* and *third* rows that succeed it.[5]

For classification, we first discarded chunks that are not strictly concrete/abstract and obtained $\hat{B}^{+s} \in \mathbb{R}^{T^s\times 2V}$. We then used regularized logistic regression to learn the per-subject $\hat{B}^{+s} \rightarrow \vec{y}^s$ mapping. The training procedure is identical to the one followed in Section 5.1.

Statistical Significance We determined the statistical significance of our classification results using a label-permutation method (Ojala and Garriga, 2009) with cross-validated accuracy as the chosen test statistic. Here, the distribution of a test statistic under the null hypothesis (that data and labels are independent) is estimated by training and evaluating the classifier on several randomized versions of the original data (by permuting classification labels). The p-value is then calculated as the proportion of randomized samples where the classifier performs better than it does on the original sample. We ran 100 iterations per subject.

[5] For rows that are ≤ 3 positions from the end, we used zero-padding for consistent dimensions.

6 Comparing the Representations

6.1 Combined model

First, we combined the word-embedding and voxel-response stimulus representations (obtained in Section 4 and Section 5.2) for each subject, by stacking the word-embedding matrix (E) and the reduced+delayed voxel-response matrix ($\hat{B}^{+}$) along the feature dimension to obtain the combined stimulus matrix $C \in \mathbb{R}^{T \times (D+2V)}$. Limiting the data to *strict* chunks yields the matrix $C^s \in \mathbb{R}^{T^s \times (D+2V)}$, which was then used for the classification task.

The rationale behind combining representations is the following. If the information encoded by the word-embedding and voxel-response representations were indeed complementary, the combined model should fare better at the prediction task than the two individual models because it now has access to information that was missing in either representation.

The classification task (predicting $\vec{y}^s$) and its training procedure are identical to those described in Section 5.1.

6.2 Residual Classification

Next, we attempted to *remove* the information present in each representation from the other and then train the classification model using the resulting representation. This procedure is described below.

1. *Removing voxel-response information from word-embeddings:* For each subject, we learned a linear mapping $L \in \mathbb{R}^{2V \times D}$ from $\hat{B}^{+s}$ to E^s through multivariate ridge regression (Haitovsky, 1987). We then computed the residuals $E^s_r \in \mathbb{R}^{T^s \times D}$ in a cross-validated manner as follows, and used the residuals for the classification task:

$$E^s_r = E^s - \hat{B}^{+s} \cdot L$$

2. *Removing word-embedding information from voxel-responses:* For each subject, we learned the linear mapping $L' \in \mathbb{R}^{D \times 2V}$ from E^s to $\hat{B}^{+s}$ through multivariate ridge regression. We then computed the residuals $\hat{B}^{+s}_r \in \mathbb{R}^{T^s \times 2V}$ in a cross-validated manner as follows, and used the residuals for the classification task:

$$\hat{B}^{+s}_r = \hat{B}^{+s} - E^s \cdot L'$$

Statistical Significance To statistically validate that any observed decrease in a residual model's performance compared to the corresponding non-residual model is really due to shared information between the representations (and not due to overfitting/chance), we adopted a "residual-permutation" procedure similar to that in Section 5.2.

Here, an empirical null distribution is created by training and evaluating each residual model above with several randomized versions of whichever representation is to be *regressed out*. The randomization is performed by permuting this representation over all time steps. The p-value is then calculated as the fraction of such residual models with cross-validated accuracies *lower* than that of the true (non-randomized) residual model. We ran 100 iterations per subject.

7 Results

We use the abbreviations **E** for the word-embedding based model, **B** for the voxel-response based (brain) model, **E+B** for the combined-representation model, **E-B** for the word-embedding model with voxel-response information removed, and **B-E** for the voxel-response model with word-embedding information removed. Figure 1 shows the classification accuracies of all models across the six subjects.

7.1 Individual models

Table 1 shows the average accuracy, recall, and F1 score of *E* and *B*.

B achieved an average classification accuracy of 69% and F1 score of 71%, and performed significantly higher than chance under the label-permutation test ($p \leq 9 \times 10^{-3}$) for each subject. This indicates that the fMRI signals triggered due to words encountered by subjects in natural stories encode enough information to significantly distinguish their concreteness levels under the current predictive framework. Evidently, this information must be useful above and beyond the noise present in the fMRI data unique to the data acquisition process. To our knowledge, the ability to classify the concreteness of *naturalistic* word stimuli from their induced brain representations in a direct, supervised fashion has not been shown in the existing literature.

E achieved a comparatively higher classification accuracy of 87%, which is in agreement with existing research (in non-naturalistic settings) on the pre-

Model	Performance (Mean ± S.D.)		
	Accuracy	Recall	F1 score
E	**0.87**	0.88	0.87
B	$\mathbf{0.69 \pm 2.5\%}$	$0.77 \pm 2.6\%$	$0.71 \pm 2.4\%$
E+B	$\mathbf{0.86 \pm 1.9\%}$	$0.86 \pm 2.6\%$	$0.85 \pm 2.0\%$

Table 1: Classification metrics across the six participants for the word-embedding based (*E*), voxel-response based (*B*) and combined (*E+B*) models.

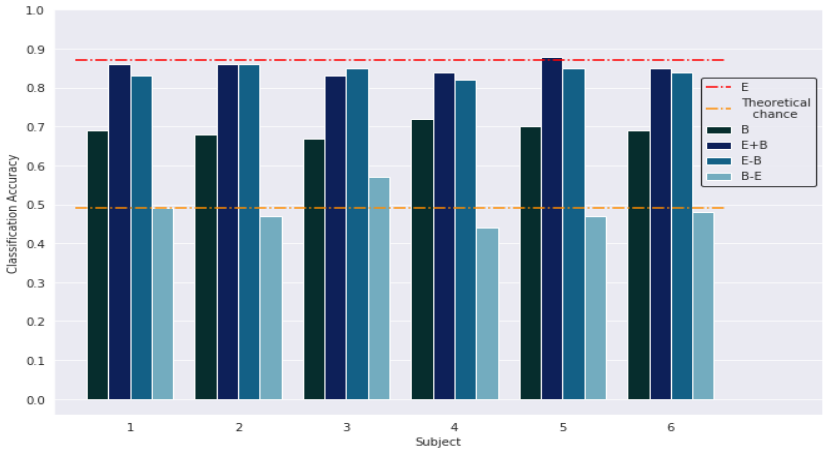

Figure 1: Variation in classification accuracies of all models over the six subjects' data.

dictability of word concreteness and imageability using word-embeddings as explanatory variables (Charbonnier and Wartena, 2019; Ljubešić et al., 2018).

7.2 Comparative models

Table 1 shows the average accuracy, recall, and F1 score of *E*, *B*, and *E+B*.

As argued in Section 1, we expect the additional sensory processing information encoded in the voxel-responses to complement the linguistic/contextual information encoded in the word-embeddings. Consequently, the combined model should fare better at distinguishing the concreteness of words in the stories.

However, our results indicate otherwise. The performance of *E+B* ($86 \pm 1.9\%$) was not significantly different from *E* (87%) under a 1-sample t-test ($t = -2.33, p = 0.07, df = 5$, 2-tail), meaning the combined model is only as good as the word-embedding based model at the task considered. Therefore, the information present in the voxel-responses relevant to differentiating between concrete and abstract words is already well-encoded by the word-embeddings, and the former does not complement the latter. On the other hand, the performance of *E+B* ($86 \pm 1.9\%$) was significantly higher than *B* ($69 \pm 2.5\%$) under a paired t-test ($t = 17.77, p = 5 \times 10^{-6}, df = 5$, 1-tail). This indicates that the word-embeddings may even contain useful extra information above that in the fMRI signals (note that we already demonstrated the effectiveness of our predictive framework in significantly distinguishing word-concreteness purely from fMRI signals). We explore this idea further next.

Table 2 shows the average accuracy, recall, and F1 score of the residual models *E-B* and *B-E*.

The results of the residual analyses are surprising. First, *E-B* achieved an average accuracy of 84%, which was significant under the residual-permutation test ($p \leq 9 \times 10^{-3}$) for each subject. The performance of *E-B* ($84 \pm 1.7\%$) was also significantly lower than *E* (87%) across subjects under a 1-sample t-test ($t = -4.71, p = 2.6 \times 10^{-3}, df = 5$, 1-tail). This shows that removing the voxel-response information from the word-embeddings marginally affects its ability to classify word concreteness. More strikingly, *B-E* achieved an average accuracy of 48%, which is lower than the theoretical chance accuracy of 50% (see Figure 1). This result was significant under the residual-permutation test ($p \leq 9 \times 10^{-3}$) for each subject, ruling out the possibility that the

Residual Model	Performance (Mean ± S.D.)		
	Accuracy	Recall	F1 score
E-B	$\mathbf{0.84 \pm 1.7}\%$	$0.85 \pm 2.4\%$	$0.84 \pm 1.4\%$
B-E	$\mathbf{0.48 \pm 9.1}\%$	$0.60 \pm 5.8\%$	$0.55 \pm 5.6\%$

Table 2: Classification metrics across the six participants for the two residual models.

Misclassified example	Ground-truth label
... And so at the earliest opportunity ...	abstract
... with this kind of curious compassion. And ...	abstract
... to suggest I might find myself on such a wayward path ...	abstract
... . Kind of blissfully unaware of what was ...	abstract
... start to get a little tricky. My husband ...	abstract
... couple amens and some applause and then everybody ...	concrete
... you know, for hundred dollars a night maybe ...	concrete

Table 3: Examples of chunks frequently misclassified by the voxel-response model. The exact phrase falling within the chunk is in dark color. We find that a majority of such misclassifications come from the abstract category.

huge performance decrease was merely caused by overfitting/chance. Across subjects too, the performance of *B-E* ($48 \pm 9.1\%$) was significantly lower than *B* ($69 \pm 2.5\%$) under a paired t-test ($t = -8.52, p = 1.8 \times 10^{-4}, df = 5$, 1-tail).

Therefore, while removing the word-embedding information from the voxel-responses fully *eliminates* the latter's predictive capability (a 30% decrease), going the other way around only has a marginal effect on predictive performance (a 3% decrease). These results show not only that the fMRI signals do not provide complementary information to the word-embeddings in making the concrete/abstract distinction, but that the relevant information in the voxel-responses is really a *subset* of the relevant information in the word-embeddings. This is a surprising result, considering the task was to distinguish a property of words theorized to fundamentally affect how the human brain represents language. We summarize our findings and provide some additional observations about this work next.

8 Conclusion

This paper has three key findings. First, we showed that words encountered in natural stories could be classified based on concreteness purely from the neural activity elicited as subjects passively comprehended the stories, using a direct, supervised approach.

Second, we showed that in making the concrete/abstract distinction, contextualized word-embeddings (i.e., GPT-2) **do not** benefit from the inclusion of information from the accompanying fMRI signals, despite evidence from several neurolinguistic studies of the human brain exhibiting fundamentally different representations over the two categories.

Finally, we found that while the residual information remaining in fMRI signals after regressing out word-embedding information can no longer distinguish concrete from abstract words, the residual information in word-embeddings beyond the fMRI signals performs significantly at this task. This shows that the information in the voxel-responses important to our prediction task is a **subset** of the corresponding information in the contextualized word-embeddings.

Our results should be interpreted in light of the following observations:

A limitation of our work is that while the voxel-responses and word-embeddings (from GPT-2) considered provide contextualized stimulus representations, the Brysbaert et al. (2014) dataset provides non-contextualized ratings for each word. We partially addressed this discrepancy by formulating the prediction task as a *classification* problem since the available labels are now much more likely to match ground-truth. I.e., it is reasonable to assume that the broad binary concreteness class of a word will rarely be modified by context as much as the continuous scores would. Future work could overcome this limitation by developing the ideas from the recently introduced CONcreTEXT task[6]

[6] https://github.com/lablita/CONcreTEXT

Metric	Model				
	E	*B*	*E+B*	*E-B*	*B-E*
Spearman's ρ (Mean $\pm$ S.D.)	0.85	$\mathbf{0.42 \pm 0.03}$	0.84 ± 0.02	0.80 ± 0.03	$\mathbf{0.09 \pm 0.05}$

Table 4: Spearman's rank-correlation coefficients ($\rho \in [-1, 1]$) between predicted and true ratings across the six participants.

of computing contextualized rating scores. We still report regression results in Table 4 for completeness and observe that they are consistent with our findings (e.g., *B-E* can no longer predict word concreteness as suggested by its near-zero rank-correlation). Finally, we find that repeating our analyses with non-contextualized word2vec embeddings (Mikolov et al., 2013) also yielded *qualitatively* identical results as in Section 7.2, indicating that our three conclusions above hold for word-embeddings more generally.

Another observation is that while *B* ($69 \pm 2.5\%$) significantly distinguishes concrete from abstract words, it still does not perform as well as *E* (87%) at this task. There could be two reasons for this difference. First, *B* does not handle abstract stimuli as well as *E* does. Quantitatively, while *B* achieves a recall of $77 \pm 2.6\%$ on concrete chunks, its recall on abstract chunks is significantly lower at $63 \pm 3.6\%$. On the other hand, *E* shows nearly identical performances over the categories. Table 3 shows some of *B*'s misclassified examples common to as many as four out of six subjects. Out of the 29 such common misclassifications, 19 (65.5%) were found to be abstract. This could indicate that neural activity patterns are not as informative for abstract stimuli as concrete stimuli, which is in agreement with psycholinguistic studies demonstrating verbal processing advantages for concrete concepts over abstract concepts (Holmes and Langford, 1976; Kroll and Merves, 1986; Romani et al., 2008). Second, the temporal resolution of functional Magnetic Resonance Imaging may be too coarse (Gauthier and Levy, 2019; Schwartz et al., 2019) for optimal performance on our task (we had to downsample the stimulus in Section 4). Nevertheless, our findings are important. Applying the current predictive framework on the fMRI signals produced highly significant results, and it is under such a framework that the above conclusions were made. Future work could explore the differences in decoding neural activity from naturalistic stimuli with imaging methods of different temporal resolutions (e.g., EEG, MEG) to determine which method should be used for which kind of task.

To conclude, we believe that this paper will inspire future work to take up the following exciting directions: Which natural language processing tasks may benefit from incorporating human language processing information into the existing frameworks? Are there ways of including such information to expose avenues for improvement in these models?

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem H. Zuidema. 2019. Blackbox meets blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains. *CoRR*, abs/1906.01539.

Andrew Anderson, Tao Yuan, Brian Murphy, and Massimo Poesio. 2012. On discriminating fMRI representations of abstract WordNet taxonomic categories. In *Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon*, pages 21–32, Mumbai, India. The COLING 2012 Organizing Committee.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually Grounded and Textual Semantic Models Differentially Decode Brain Activity Associated with Concrete and Abstract Nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

F Gregory Ashby. 2019. *Statistical analysis of fMRI data*. MIT press.

Sai Abishek Bhaskar, Maximilian Köper, Sabine Schulte Im Walde, and Diego Frassinelli. 2017. Exploring Multi-Modal Text+Image Models to Distinguish between Abstract and Concrete Nouns. In *Proceedings of the IWCS workshop on Foundations of Situated and Multimodal Communication*.

J. R. Binder, C. F. Westbury, K. A. McKiernan, E. T. Possing, and D. A. Medler. 2005. Distinct Brain Systems for Processing Concrete and Abstract Concepts. *Journal of Cognitive Neuroscience*, 17(6):905–917.

Paul Boersma and David Weenink. 2001. PRAAT, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

M. Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41:977–990.

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

Charlotte Caucheteux and Jean-Rémi King. 2020. Language processing in brains and deep neural networks: computational convergence and its limits. *bioRxiv*.

Jean Charbonnier and Christian Wartena. 2019. Predicting Word Concreteness and Imagery. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 176–187, Gothenburg, Sweden. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *CoRR*, abs/1710.04087.

Fatma Deniz, Anwar O. Nunez-Elizalde, Alexander G. Huth, and Jack L. Gallant. 2019. The Representation of Semantic Information Across Human Cerebral Cortex During Listening Versus Reading Is Invariant to Stimulus Modality. *Journal of Neuroscience*, 39(39):7722–7736.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. 2010. New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects. *Journal of Neurophysiology*, 104(2):1177–1194. PMID: 20410363.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, Hong Kong, China. Association for Computational Linguistics.

Yoel Haitovsky. 1987. On Multivariate Ridge Regression. *Biometrika*, 74(3):563–570.

Felix Hill and Anna Korhonen. 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 (EMNLP)*, pages 255–265, Doha, Qatar. Association for Computational Linguistics.

Felix Hill, Anna Korhonen, and Christian Bentz. 2014. A Quantitative Empirical Analysis of the Abstract/Concrete Distinction. *Cognitive Science*, 38(1):162–177.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A Framework for Cognitive Word Embedding Evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549, Hong Kong, China. Association for Computational Linguistics.

V.M. Holmes and J. Langford. 1976. Comprehension and recall of abstract and concrete sentences. *Journal of Verbal Learning and Verbal Behavior*, 15(5):559 – 566.

Anne Hsu, Alexander Borst, and Frédéric E Theunissen. 2004. Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, 15(2):91–109. PMID: 15214701.

Alexander Huth, Wendy Heer, Thomas Griffiths, Frédéric Theunissen, and Jack Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458.

Shailee Jain and Alexander Huth. 2018. Incorporating Context into Language Encoding Models for fMRI. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. 2020. Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech. In *Advances in Neural Information Processing Systems*, volume 33, pages 13738–13749. Curran Associates, Inc.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. 2002. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage*, 17(2):825 – 841.

Mark Jenkinson and Stephen Smith. 2001. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143 – 156.

Judith F Kroll and Jill S Merves. 1986. Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(1):92.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting Concreteness and Imageability of Words Within and Across Languages via Word Embeddings. In *Proceedings of The Third Workshop*

on Representation Learning for NLP, pages 217–222, Melbourne, Australia. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

George A. Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41.

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.

M. Ojala and G. C. Garriga. 2009. Permutation Tests for Studying Classifier Performance. In *2009 Ninth IEEE International Conference on Data Mining*, pages 908–913.

Allan Paivio. 1971. *Imagery and Verbal Processes*. Holt, Rinehart and Winston.

Allan Paivio. 1991. Dual Coding Theory: Retrospect and Current Status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 45(3):255–287.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Stephen Roller and Sabine Schulte im Walde. 2013. A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, Seattle, Washington, USA. Association for Computational Linguistics.

Cristina Romani, Sheila Mcalpine, and Randi C. Martin. 2008. Concreteness Effects in Different Tasks: Implications for Models of Short-Term Memory. *Quarterly Journal of Experimental Psychology*, 61(2):292–323. PMID: 17853203.

R. W. Schafer. 2011. What Is a Savitzky-Golay Filter? [Lecture Notes]. *IEEE Signal Processing Magazine*, 28(4):111–117.

Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. 2020. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv*.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. Inducing brain-relevant bias in natural language processing models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14123–14133. Curran Associates, Inc.

Carina Silberer and Mirella Lapata. 2014. Learning Grounded Meaning Representations with Autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland. Association for Computational Linguistics.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14954–14964. Curran Associates, Inc.

Walter J. B. van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190. PMID: 24417251.

Jing Wang, Laura B. Baucom, and Svetlana V. Shinkareva. 2013. Decoding abstract and concrete concept representations based on single-trial fMRI data. *Human Brain Mapping*, 34(5):1133–1147.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously Uncovering the Patterns of Brain Regions Involved in Different Story Reading Subprocesses. *PLOS ONE*, 9(11):1–19.

Jiahong Yuan and Mark Liberman. 2008. Speaker identification on the SCOTUS corpus. *Acoustical Society of America Journal*, 123(5):3878.