Multimodality of Meaning: Evaluating Alignment Between Transformer & Cortical Representations of Abstract & Concrete Concepts

Torrey Snyder
University of Amsterdam
torrey.snyder@student.uva.nl

Abstract

In recent years, a variety of multi-modal models have been proposed with demonstrable success on many downstream language-and-vision tasks, exhibiting a greater degree of alignment with human semantic judgments, as well as brain activations, in comparison to language-only models (Oota et al., 2022a; Pezzelle et al., 2021). This research implemented brain encoding models to compare how well these multimodal models' word embeddings align with semantic network activations in the brain. Of particular interest was the differential predictive accuracy between concrete and abstract words. Dual coding theory claims that concrete semantic representations involve more visual information than their primarily linguistic abstract counterparts (Paivio, 1991), and therefore it was expected that the enhanced representational alignment conferred by multimodality would primarily manifest for concrete, rather than abstract, words. In line with prior research by Oota et al. (2022b), findings presented here provide further empirical evidence that multimodal embeddings can more accurately predict brain activity than their language-only counterparts. Notably, the observed difference in embeddings' predictive accuracies between concrete and abstract concepts was unexpectedly small, suggesting that the enhanced neural alignment conferred by multimodality generalizes across both abstract and concrete words.

1 Introduction

1.1 Dual Coding Theory

First proposed by Paivio (1991), dual coding theory postulates that information processing in the brain is implemented across two separate channels: verbal and visual. One suggested manifestation of this bifurcation is in the distinct semantic encodings of concrete and abstract concepts, with concrete concept representations postulated to contain more visual information and abstract concept representations thought to contain more verbal information (Paivio, 1991). fMRI studies have provided some empirical support for this theory, revealing that the processing of abstract concepts in the brain is left-lateralized, while the processing of concrete concepts is bilateral (Binder et al., 2005). This experimental finding is consistent with the expectation that concrete words activate anatomically distinct brain regions not activated during the processing of abstract words.

Tang et al. (2021) sought further confirmation of this theory by constructing a computational model of how visual and linguistic information could be integrated to form semantic representations. In their methodological framework, visual and linguistic representations were first modeled as separate concept embedding spaces and later concatenated. Comparing encoding model performance between different semantic embedding spaces, Tang et al. (2021) demonstrated that the most neurally aligned embeddings contained a combination of visual and linguistic information, with more concrete concepts best modeled by embeddings containing a higher degree of visual information. Notably, Tang et al. (2021) also found that even highly abstract concept representations were not purely linguistic, but instead also contained some amount of visual information, acquired from associated concrete concepts. These findings provide the basis for the experimental investigation presented here, which compares the neural alignments of embeddings generated by transformer-based multimodal models. Unlike in the semantic embedding spaces constructed by Tang et al. (2021), in these models, the integration of visual and linguistic information is not achieved by mere concatenation, but instead through more complex cross-modal supervision mechanisms (Li et al., 2019; Radford et al., 2021; Tan and Bansal, 2020).

1.2 Symbol Grounding Problem

The symbol grounding problem, first articulated by Harnad (1990), describes how symbolic representations of concepts are made meaningful through their connections to real-world referents. As noted by Bender and Koller (2020), language models trained only on linguistic form lack connection to the real-world referents they denote. This lack of grounding limits the cognitive plausibility of these language models. Resolving the symbol grounding problem is therefore critical for constructing models that can faithfully replicate the human capacity for language understanding.

Given this assertion that symbolic representations, such as language, must be grounded in sensory percepts (Harnad, 1990), word embeddings generated using multimodal data, rather than linguistic form alone, should provide richer concept representations that are more faithful replications of semantic networks in the human brain (Baroni, 2016). Prior work by Pezzelle et al. (2021) provides some empirical evidence of this. In their study, Pezzelle et al. (2021) evaluated the extent to which the embeddings generated by multimodal transformers align with human semantic intuitions, finding that the Vokenization model in particular exhibits robust alignment with human judgments. Pezzelle et al. (2021) suggest that the model's token-level approach to visual supervision may contribute to more well-defined multimodal word representations compared to the sentence-level approach implemented by other multi-modal models.

Meanwhile, work by Oota et al. (2022a) evaluating brain encodings across multiple multi-modal transformers found that embeddings generated from VisualBERT (Li et al., 2019) - a single-stream model which jointly encodes text and visual input using cross-modal attention - are more predictive of fMRI responses than, among other multi-modal transformers, CLIP. CLIP implements separate image and text encoders that are trained jointly to produce image and text embeddings such that the cosine similarities between image-text embedding pairs are maximized (Radford et al., 2021). Given that these studies evaluated model performance on different evaluation metrics (alignment with human judgments vs. fMRI responses), it was uncertain how the neural alignments of Vokenization-generated embeddings would directly compare to those of VisualBERT or CLIP - each of which implement a different approach to visuo-linguistic information integration. This was one particular line of inquiry that motivated the present work.

2 Related Work

The implementation of brain encoding models to predict fMRI responses has become an increasingly frequent methodological technique in recent years, particularly in the research programme of computational cognitive neuroscience. Earlier studies primarily implemented encoding models that predict brain activity from representations of single-mode stimuli, either visual or text (Allen et al., 2022; Schrimpf et al., 2021). Yet the human brain processes inputs from diverse modalities, such as vision and audio, in parallel. Therefore, brain encoding models trained on multimodal data should provide a richer account of how the brain integrates multiple channels of sensory information to construct grounded semantic representations.

Prior work by Pereira et al. (2018) provided this multimodal dataset, and the functional magnetic resonance imaging (fMRI) data collected in that study was used in the present work for brain encoding analysis. In the original fMRI study, subjects were asked to read a word and think about its meaning in the context of either sentences or pictures. Pereira et al. (2018) implemented single-concept decoding, investigating the extent to which 300-dimensional semantic vectors (representations derived using GloVe) could be decoded from brain imaging data. These subject-specific decoding models were trained on the 5000 most informative fMRI voxels. The "informativeness" score of each voxel was determined using a ridge regression model, with more informative voxels yielding higher correlations between predicted and true values for each semantic vector dimension. Pereira et al. (2018) found that the most informative voxels were distributed across 4 networks: (i) frontotemporal language-specific network, (ii) the default mode network (DMN), (iii) task-positive (TP) network, (iv) vision network. The highest degrees of predictive accuracy observed by Pereira et al. (2018) were in the language and vision networks, which were the regions subsequently selected for the brain encoding analysis presented here.

Recent work by Wang et al. (2022) demonstrated that the multimodal transformer model CLIP (Radford et al., 2021) better encodes neural responses in the visual cortex in comparison with unimodal models such as BERT or ImageNet-trained ResNet. In

this study, Wang et al. (2022) extracted image features generated from CLIP, which encodes visual concepts via supervision from natural language captions. CLIP jointly trains an image encoder and text encoder to maximize the cosine similarity between corresponding image and text embeddings, using a linear projection to map each encoder's representation to a multi-modal embedding space (Radford et al., 2021). Wang et al. (2022) then used voxelwise encoding models based on these CLIP features to predict brain responses to real-world images from the Natural Scenes Dataset. It was found that CLIP explains greater unique variance in higher-level visual areas compared to models trained only with image/label pairs (ResNet) or text (BERT).

3 The Present Work

In this work, word embeddings extracted from transformer-based models (RoBERTa, CLIP, VisualBERT, Vokenization) were mapped to voxelwise activations via a brain encoding model to compare the representational alignments of these embeddings with fMRI responses. The present work built on the prior encoding study of Oota et al. (2022a) by expanding the set of multimodal transformers to be evaluated to include the Vokenization model (Tan and Bansal, 2020), found to be the best performing model in the aforementioned research conducted by Pezzelle et al. (2021). This research further extends the work conducted by Oota et al. (2022a) by comparing brain encoding performance across different perceptual contexts (linguistic vs. visual processing). While Oota et al. (2022a) restricted their brain encoding analysis to the pictures context, the present work compares brain encoding performance across both experimental contexts used by Pereira et al. (2018) (pictures & sentences). The final extension of Oota et al. (2022a)'s research introduced in this work is the comparison of brain encoding performance between concrete and abstract words. In summary, this work aimed to investigate:

(a) whether multi-modal transformers have higher accuracy than text-only models in predicting brain activity,

(b) whether this higher predictive accuracy only holds for concrete concepts,

(c) if the addition of visual information to embeddings improves their predictive accuracy for visual brain regions

(d) if the Vokenization model (best-performing in Pezzelle et al. 2021) has greater predictive accuracy than VisualBERT (best-performing in Oota et al. 2022a),

(e) whether the type of context (pictures vs. sentences) in which each target word is presented affects the encoding models' predictive accuracy.

This research sought to provide empirical support for both Harnad's assertion of the necessity of symbol grounding and Paivio's dual coding theory by investigating whether multimodal models, compared to a text-only model (RoBERTa), indeed achieve higher predictive accuracies for concrete words in comparison to abstract words.

As described previously, grounded cognition theories claim that a concept's semantic representation is constructed using associated perceptual information (Harnad, 1990). Yet distributional word embeddings' success - despite their lack of perceptual access to real-world referents - suggests that sufficiently functional semantic representations can be learned from language alone (Patel and Pavlick, 2021). However, dual coding theory claims that the robustnes of these embeddings can vary depending on whether the represented concept is abstract or concrete. In this theoretical formulation, concrete concept representations are postulated to contain a high degree of visual information (Paivio, 1991). Therefore, multimodal embeddings of concrete concepts were expected to achieve a greater degree of neural alignment compared to their unimodal (text-only) counterparts. On the other hand, dual coding theory claims that abstract concepts are represented primarily by linguistic information. In this case, unimodal embeddings for abstract concepts should were expected to achieve comparable predictive accuracies with their multimodal counterparts.

4 Brain Imaging Dataset

The present study used fMRI data collected and preprocessed by Pereira et al. (2018) to train brain encoding models. While the Pereira et al. (2018) study implemented a decoding model, the present study inverted this paradigm - rather than predicting semantic vectors from voxel activations, here voxel activations were predicted from semantic vectors. The Pereira et al. (2018) dataset is comprised of two experimental contexts: (1) linguistic stimuli (sentences) and (2) visual stimuli (pictures). In each context, participants were shown a concept word alongside the contextual stimuli (either sentences

or pictures) with the aim of observing brain activation when participants retrieved relevant word meanings using the contextual information. In the original Pereira et al. (2018) study, 15 subjects were presented the stimuli (6 sentences/images) corresponding to 180 concepts. In their study, fMRI data was processed using the FMRIB software library (FSL). Data each scanning session were corrected for motion, slice timing, and bias field inhomogeneity (Pereira et al., 2018). Further temporal pre-processing included high-pass filtering, in which low-frequency noise is removed from the data. Blood oxygenation-level dependent (BOLD) data collected by Pereira et al. (2018) for each subject in each scanning session was represented with a matrix of 96 x 96 voxels. Given that the most informative (predictable) voxels were located in the vision and language networks, these were the regions of interest selected in the brain encoding analysis presented here.

5 Methodology

5.1 Abstract/Concrete Word Designation

Given that the present study aimed to investigate the comparative model performances between concrete and abstract words, the original set of 180 words was reduced to 132 words (63 abstract, 69 concrete). Each word's designation as abstract/concrete was established by its behaviorally-determined concreteness score, based on data collected by Brysbaert et al. (2014), in which participants assigned concreteness ratings (ranging from 1-5) to presented words. The delineation between abstract/concrete for the particular set of words presented in the Pereira et al. (2018) study was achieved by first calculating the mean concreteness score within the original word list and subsequently defining concrete words as all words half a standard deviation above the mean and abstract words as all words half a standard deviation below (Hendrikx and Beinborn, 2020).

5.2 Pre-trained Transformer Embeddings

5.2.1 Text-only Baseline

RoBERTa only encodes text stimuli. Similar to its predecessor BERT (Devlin et al., 2018), RoBERTa is trained using a masked language objective, in which the model is tasked with predicting the original vocabulary ID of the masked token based only on its context (Liu et al., 2019). In this way, the model learns contextualized word representations. Given that each word occurred in 6 different contexts (sentences), the 6 corresponding vector outputs were averaged to generate a single static embedding for each word. This context combination mechanism, in which multiple contextualized vector representations are collapsed to a single static representation, is similar to the aggregation method first proposed by Bommasani et al. (2020) and also implemented by Pezzelle et al. (2021).

5.2.2 Multi-modal transformers

CLIP projects both text and visual features to a latent space, whereby both textual and image embeddings have identical dimensions, such that the dot product between the projected image and text features can then be calculated and used as a similarity score, with the objective of maximizing this value (Radford et al., 2021). As with RoBERTa, the average across each set of 6 vector representations was taken to generate a single embedding per modality (language and vision). Given that CLIP generates separate embeddings for image and text features, to construct a joint representation, the element-wise multiplication fusion technique was implemented (Jaafar and Lachiri, 2023).

VisualBERT implicitly aligns elements of the input text and regions in the input image via self-attention (Li et al., 2019). Visual embeddings are computed by summing three embeddings: (a) a visual feature representation of the bounding region, computed by a convolutional neural network, (b) a segment embedding indicating it is an image embedding, and (c) a position embedding (Li et al., 2019). Following the methodological implementations of (Oota et al., 2022a), the visual feature representation was comprised of region proposals as well as bounding box regression features extracted from Fast R-CNN (Ren et al., 2015) as image features. Multimodal embeddings are generated by aligning these vision embeddings with text input (sentences). As with the other models, the average across 6 vector representations was used to generate a single multimodal embedding for each concept.

Vokenization While the aforementioned multimodal models apply natural language supervision to vision, Vokenization (Tan and Bansal, 2020) is a visually-supervised language model. This model incorporates a novel technique "vokenization" that extrapolates multimodal alignments to language-only data by contextually mapping language tokens

to their related images ("vokens"). Supervised by these generated vokens, significant improvement over the purely self-supervised language model on multiple language tasks has been observed (Tan and Bansal, 2020). Consistent with the embedding extraction methodology for the other models, given 6 input sentences per word, a single multimodal embedding is generated by taking the average across each contextualized representation.

5.3 Encoding Models

In line with the methodological framework implemented by Oota et al. (2022a), fMRI encoding models were trained using ridge regression to predict fMRI brain responses (collected by Pereira et al., 2018) for regions of interest activated in response to presentation of concept stimuli - sentences or pictures associated with target concept. The objective of each ridge regression-based encoding model was to predict fMRI voxel activations given an input embedding. These embeddings were obtained using multi-modal transformers, and, for comparison purposes, a pre-trained text transformer. The following brain networks were selected as regions of interest (ROIs): language - left and right hemispheres (regions include: lateral medial temporal gyrus, lateral prefrontal temporal gyrus, lateral inferior temporal gyrus) vs. vision - body, object, face, scene (regions include: primary visual cortex, fusiform body area, lateral occipital cortex, fusiform face area, parahippocampal place area) (Oota et al., 2022a). Given that these brain regions can be divided into those that specialize in visual processing and those that specialize in linguistic processing, it was expected that the predictive accuracies across these brain regions would be influenced by the experimental context (sentences vs. pictures). To train and test from a single dataset (BOLD data for 1 subject), K-fold (K=10) cross-validation was implemented, in which all the data samples from K-1 folds were used for training, and evaluation was performed using samples of the left-out fold. Given that there were 15 subjects (N=15) in the original study, resulting in 15 sets of fMRI data, 2V2 accuracy scores for each ROI were averaged across participants to generate a mean predictive accuracy for each ROI.

5.4 Evaluation Metric: 2V2 Accuracy

Using the evaluation methodology implemented by Toneva et al. (2020), given 2 left-out samples, with predictions b1 and b2 and corresponding ground truth (observed voxel activations) B1 and B2, 2 scores are calculated: score1 is computed as the sum of the cosine distances between (b1, B1) and (b2,B2) and score 2 is computed as the sum of the cosine distances between (b1 and B2) and (b2, B1). This score 2 corresponds to random prediction, as the predicted activation of voxel 1 should be closer to the observed activation of voxel 1, rather than an independent voxel 2. An indicator function is applied that returns 1 if score1 $<$ score2 and 0 otherwise. In this evaluation metric, chance performance is 0.5.

6 Results

6.1 Inter-Model Comparison: Multi-Modal vs. Text-Only

Presented in Fig. 2 are the 2V2 accuracy results for each model (RoBERTa, CLIP, VisualBERT, Vokenization), across both abstract and concrete word sets. As expected, multi-modal models do indeed outperform the text-only RoBERTa. In particular, multimodal models CLIP and VisualBERT achieve higher predictive accuracies in vision areas (Object, Body, Face, Scene). In each set of words (abstract & concrete) in the sentences context, CLIP embeddings outperform those of RoBERTa, VisualBERT and Vokenization. In the pictures context, however, VisualBERT embeddings outperform those of CLIP in visual regions.

A higher correlation across the visual and language brain regions achieved by multi-modal embeddings demonstrates that the integration of visual and linguistic representations indeed increases neural alignment. Multi-modal transformers outperform the baseline across all 7 regions of interest (language-left hemisphere, language-right hemisphere, vision-primary, vision-body, vision-face, vision-object, vision-scene). Interestingly, in the pictures context, Vokenization's performance is notably lower than the other multimodal models. This comparatively poor performance might be a consequence of Vokenization's visually-supervised approach to language processing. While this learning mechanism may improve its performance on linguistic tasks, Vokenization's embeddings may be comparatively poorer representations of visual information.

6.2 Abstract vs. Concrete

The 2V2 accuracies of multimodal embeddings for abstract words words was greater than anticipated,

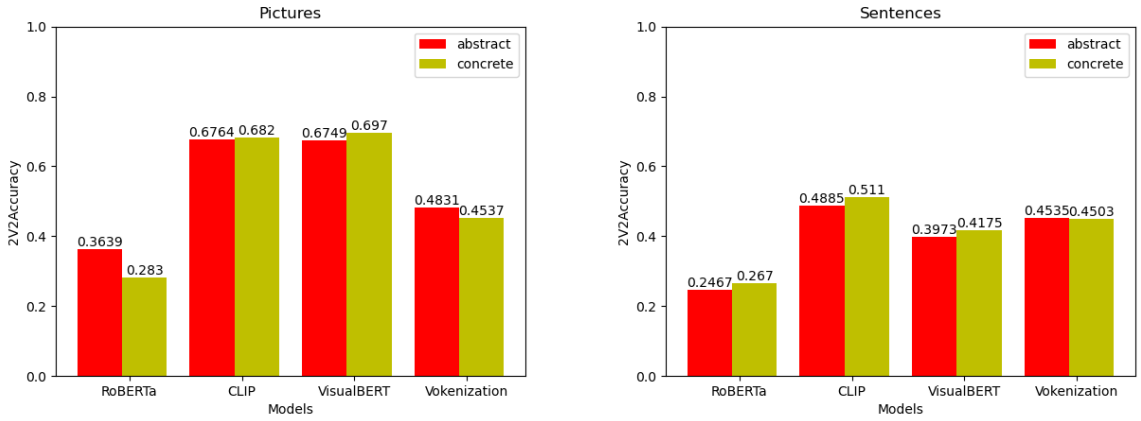

Figure 1: Inter-model comparisons of predictive accuracy between pictures (left) and sentences (right) paradigms for brain regions involved in visual scene perception (PPA). 2V2accuracy for RoBERTa and Vokenization embeddings does not appear to be influenced by the experimental context (pictures vs. sentences), while for both CLIP and VisualBERT embeddings, there is a consistent drop in accuracy from picture to sentence stimuli. In the sentences paradigm, Vokenization outperforms VisualBERT; however, note that this predictive accuracy is still poor (< 0.5).

with no notable differences from the 2V2 accuracies for concrete words. Furthermore, this similarity between predictive accuracies for abstract and concrete word embeddings held for brain regions involved in visual processing. As noted previously, the results of Tang et al. (2021) indicate that even highly abstract concepts should contain some visual information, due to their linguistic associations to concrete words. These findings suggest that abstract concepts are perhaps more visually grounded than originally described by dual coding theory.

This enhanced neural alignment of multimodal representations, which generalizes across concrete and abstract words, may be a consequence of the inherent multimodality of meaning - cortical representations of even highly abstract concepts may involve some degree of visual information. While this is consistent with the aforementioned results from Tang et al. (2021), this finding is harder to reconcile with the original implications of dual coding theory, which suggested a stronger bifurcation of visual & linguistic information between concrete and abstract word representations. However, it is also important to note that the 2V2 evaluation metric used in the present work is not a direct measure of fMRI responses that would enable comparison between neural activity across brain regions. Therefore, it could still be the case that abstract words yield higher activations in the language network than in the vision network (and vice versa for concrete words). While this would be a worthwhile neuroimaging study, as a brain encoding analysis, this is beyond the scope of the present work.

6.3 Language vs. Visual Brain Regions

As can be observed in Fig. 2, multimodal models yielded higher predictive accuracies than the unimodal model in visual brain regions (ROIs are listed along the x-axis of each subplot). Interestingly, and unexpectedly, this enhancement extended for both abstract and concrete words. Furthermore, compared with the text-only model, the multimodal models also achieved a higher predictive accuracy in brain regions involved in linguistic processing. Again, this improvement manifested for both abstract and concrete words.

While multimodal embeddings' enhanced predictive accuracy was observed in both language and visual brain regions, this enhancement was not necessarily equal across both brain networks. In the pictures context, embeddings from CLIP and VisualBERT (the two best-performing models) have higher predictive accuracy in visual brain regions than in language regions. For both models, activation of brain regions responsible for visual scene perception (PPA) was most predictable. The enhanced encoding performance in this region in particular is noteworthy, because this type of high-level visual processing involves the perception and identification of real-world referents, and thus would be highly relevant for symbol grounding.

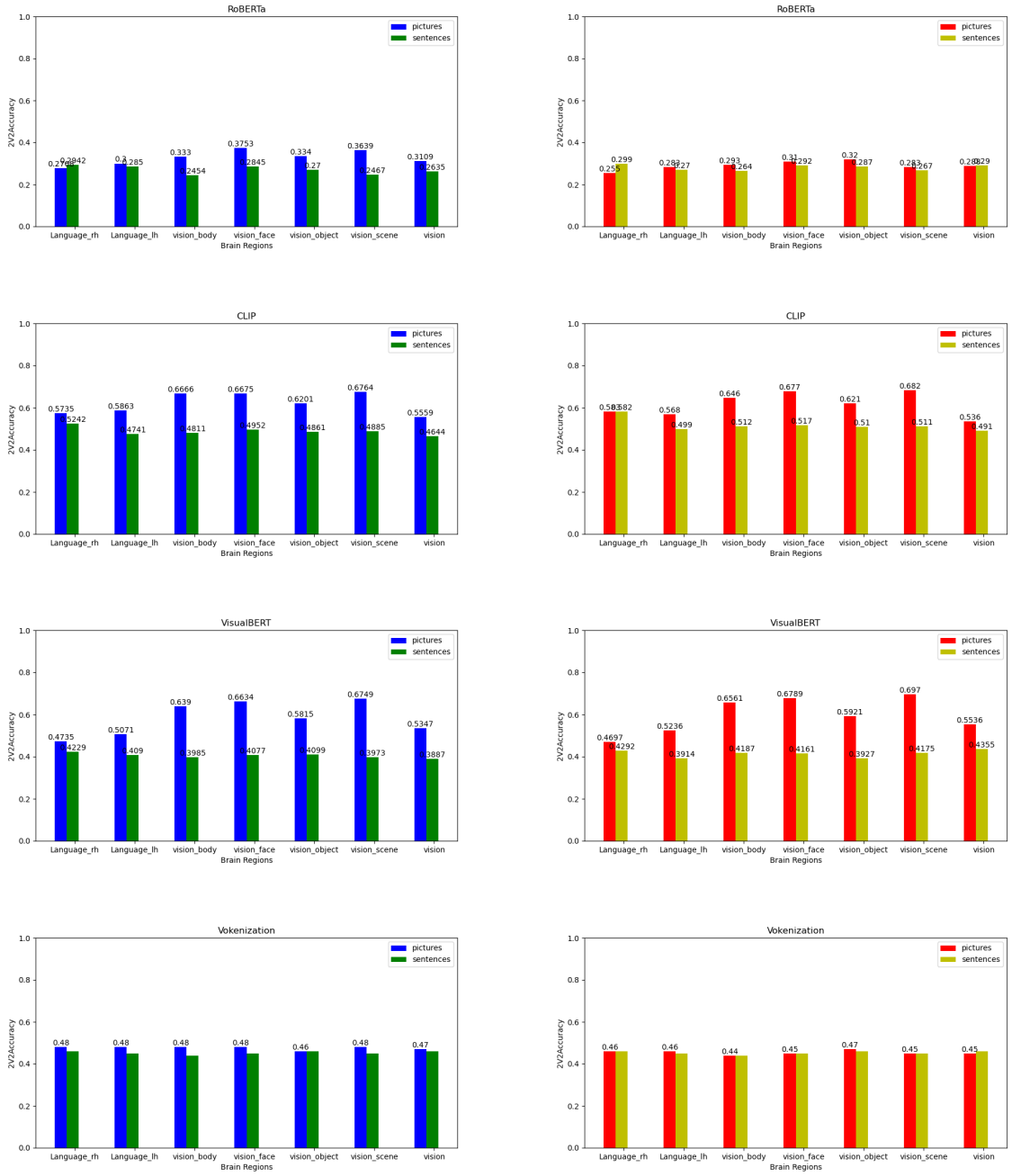

Figure 2: Model performances across brain regions for abstract (left) and concrete (right) words. The predictive accuracy of RoBERTa and Vokenization embeddings does not change notably across brain regions. However, both CLIP and VisualBERT exhibit greater predictive accuracy in higher-order visual regions (for both abstract and concrete words). While there is no notable difference across abstract and concrete words, it is observed that the type of stimuli (pictures vs. sentences) does impact the 2V2 accuracy of CLIP and VisualBERT.

While it is intuitive that visual brain region activation would be more predictable in the pictures context, it was expected that this enhanced predictability would be restricted to concrete words. However, this higher predictability in visual brain regions extended to both concrete and abstract words. Notably, this differential predictive accuracy between language and visual brain regions is not observed in RoBERTa or Vokenization, both of which achieved comparatively poor predictive accuracy in the pictures context.

6.4 Vokenization vs. VisualBERT

The comparatively poor predictive accuracies (< 0.5) across brain regions of the Vokenization-generated embeddings seem contrary to the results of Pezzelle et al. (2021). That study implemented a judgment-based evaluation metric, in which performance was measured by the alignment between embeddings and concreteness ratings derived from human judgments (Brysbaert et al., 2014). The findings presented here suggest that this concreteness rating cannot be inferred from the fMRI responses collected by Pereira et al. (2018). Perhaps this reflects that the tasks from the two studies were fundamentally incomparable, involving distinct modes of processing. Unlike the Pereira et al. (2018) task, which involved multi-modal processing of both visual and linguistic stimuli, the task of assigning concreteness scores to words is fundamentally a linguistic one. Even though the original Brysbaert et al. (2014) study defined concrete words as words that can be experienced "directly through one of the five senses" - a definition which suggests some invocation of multimodal perception - the only stimuli presented to the raters in this classification task was text. Therefore, given this contextual difference between the original Brysbaert et al. (2014) rating task and the study presented here (unimodal vs. multimodal stimuli), a more valid analysis of the Pezzelle et al. (2021) study and the present findings would restrict comparison to the sentences (linguistic-only stimuli) context. Indeed, in this context, Vokenization does outperform VisualBERT. However, even in this context, for each ROI, Vokenization-generated embeddings still yield poor predictive accuracies (< 0.5). Notably, in this unimodal (sentences) context, CLIP-generated embeddings are the only representations to yield predictive accuracies greater than 0.5. It should be pointed out that the original Pezzelle et al. (2021) study does not include a comparison with CLIP-generated embeddings, so it is uncertain whether these embeddings would also be more aligned with human-derived concreteness scores.

While Vokenization-generated embeddings may have been slightly more neurally aligned than VisualBERT embeddings in the sentences context, this did not hold in the pictures context. Instead, VisualBERT-generated embeddings were more predictive of visual brain region (body, face, scene, V1) activations than both Vokenization and CLIP embeddings. This is consistent with the findings of Oota et al. (2022a). However, Oota et al. (2022a) also observed this enhanced predictive accuracy in the language (both left and right hemisphere) regions, which is not replicated here. Instead, the present findings show that, in both language regions, CLIP-generated embeddings yield the highest predictive accuracies. This inconsistency could be explained by the differences in data sets. In the present study, the original word list of 180 was reduced to 69 (concrete word encoding) and 63 (abstract word encoding). Therefore, the superior neural alignment of CLIP-generated embeddings observed here may not necessarily be a reflection of a more cognitively plausible neural network architecture or objective function, but may merely be a consequence of the highly data-dependent nature of brain encoding analyses.

6.5 Sentences vs. Pictures

As can be observed in Fig. 1, while the predictive accuracy of both RoBERTa and Vokenization embeddings did not notably change between context types, CLIP and VisualBERT embeddings were better at predicting brain activations in the pictures task compared to the sentences task. This stimuli-dependent difference is consistent with the original findings of Pereira et al. (2018). In their study, pairwise accuracy of their decoding models was higher for the pictures paradigm than sentences.

It is worth remembering here that the text components of the multimodal embeddings were generated from sentence-level contextualizations. But in the Pereira et al. (2018) pictures experimental context, each word was essentially a static representation, presented with no linguistic context. Therefore, in this experimental paradigm, static word embeddings generated from GloVe (Pennington et al., 2014) - which uses word co-occurrence statistics, rather than a sentence-level contextual

window - might have had a greater degree of representational alignment with the neural responses in the language network of the brain. Indeed, this was the word embedding method implemented for the original Pereira et al. (2018) brain decoding study. However, in another sense, these presented words were indeed contextualized, but this contextualization was provided in the form of visual, rather than linguistic, data. Given that CLIP- and VisualBERT-generated embeddings still achieved higher predictive accuracies for language-processing regions in the pictures context rather than the sentences context, it does not appear that this distinction negatively affected the embeddings' neural alignment.

One speculative explanation for the comparatively poor predictability in the purely linguistic (sentences) task could be that linguistic processing in the brain is comparatively diffuse, comprised of both functionally specialized and domain-general networks, and therefore activates an extensive range of subregions. (Fedorenko and Thompson-Schill, 2014). This heterogeneity could limit predictability in purely linguistic contexts.

7 Discussion

7.1 Cognitive Implications: Visuo-Linguistic Semantic Representations

While multi-modal embeddings did have greater predictive accuracy than the text-only baseline, there was no notable difference in accuracy between abstract and concrete words. In other words, multi-modal embeddings were no more predictive of concrete than abstract words, contrary to the hypothesis that the increase in accuracy will be primarily restricted to concrete words, given the assumption that abstract word representations involve little visual information. This similarity in predictive accuracies between abstract and concrete words held across both experimental contexts (pictures and sentences).

The finding that this similarity also held in the pictures context is particularly surprising after taking into account the significant differences between the corresponding image sets for abstract and concrete words. For example, the word "ball" has one of the highest concreteness ratings in the Pereira et al. (2018) word list, with a concreteness score of 5.0. Its corresponding images - 3 of which are displayed in Fig. 3 - are characterized by a high degree of similarity, particularly in comparison to the set of images corresponding to the highly abstract word "typical" (concreteness score = 1.52). The high degree of variance between the abstract images would be consistent with the assumption that abstract concept representations contain primarily linguistic information, as the corresponding visual stimuli appear to be less informative than that of concrete concepts. Yet, the findings presented here seem to contradict those intuitions. The comparatively disparate visual information conveyed by the set of corresponding abstract images appear to have little negative effect on the predictive accuracy of their visio-linguistic embeddings, with little difference in 2V2 accuracy between abstract and concrete concepts. Furthermore, the predictive accuracy of CLIP and VisualBERT's visuo-linguistic embeddings is actually higher in the pictures paradigms.

7.2 Natural Language-Supervised Vision Models vs. Visually-Supervised Language Models

The superior predictive accuracy of CLIP is consistent with prior findings from Wang et al. (2022). Their findings indicate that CLIP is much more accurate than single modality models at predicting brain activation in higher-level visual regions. Multimodal loss signals from CLIP's final layer are propagated through all earlier layers of both the visual and language encoders (Radford et al., 2021). Wang et al. (2022) speculate that this may render CLIP a more faithful approximation of human visual processing by endowing the model with some degree of top-down knowledge that is able to influence earlier layers of visual input. This top-down influence of language on vision is apparent in humans during category learning (Conwell et al., 2022). The findings of Wang et al. (2022), combined with the present study's demonstration of CLIP embeddings' superior predictive accuracy, provides some compelling evidence that supervision from natural language leads to representations that are more predictive of cortical activation in high-level visual regions.

On the other hand, the comparatively poor predictive accuracy of the Vokenization model - a visually-supervised language model - suggests that perhaps the transfer of visual information to language processing is less robust than the transfer of linguistic information to visual processing. CLIP and VisualBERT's differential predictive accuracy between the sentences and pictures con-

Figure 3: Subset of corresponding images from Pereira et al. (2018) dataset for abstract word "typical" (top) and concrete word "ball" (bottom)

texts may also be explained by this. This comparatively high predictive accuracy in the pictures paradigm suggests that multimodality (specifically, that achieved using the CLIP or VisualBERT architecture) may be more useful in visual, rather than language, processing. Nevertheless, the superior accuracy of each of the multimodal (CLIP, VisualBERT, Vokenization) embeddings over the unimodal (RoBERTa) embeddings across both (pictures & sentences) contexts confirms that multimodality does improve neural alignment for both visual and linguistic processing. However, these results indicate that the type of cognitive processing being performed may affect the degree of this improvement, with a greater degree of representational alignment in visual, rather than purely linguistic, contexts.

8 Limitations

The present study's exploration of multimodality could be further extended to include multimodal models that integrate audio information as well, such as Akbari et al. (2021)'s recently proposed Video-Audio-Text Transformer (VATT). The present work's language data consists only of text. This limits the study's examination of linguistic processes to reading - a fundamentally visual task. The addition of audio information may facilitate the construction of richer visio-linguistic representations, which may yield higher predictive accuracies for language processing regions in the brain.

As mentioned previously, GloVe embeddings could have been also used as an additional text-only baseline. This modification would have also enabled comparison of neural alignment across another dimension: word co-occurrence vs. sentence-level contextualization derivations of text embeddings.

Overall, the comparison presented here between multi-modal models is worthy of more in-depth investigation. Model architecture and objective function alone cannot account for the observed differences in predictive accuracy because these models also differ in the size of their training datasets. As observed by Conwell et al. (2023), the training dataset is also a determinative factor in a model's alignment with neural data. Further experimental work involving controlled comparisons (in which the training dataset is held constant) between multi- and uni-modal models are warranted. Necessarily, this would restrict comparisons to models that share the same training data. However, this line of research will be crucial to isolating whether multimodality indeed is responsible for multimodal models' enhanced alignment with neural data.

9 Conclusion

The comparatively high predictive accuracy of CLIP embeddings observed in this study prompts speculation into whether its objective function of maximizing the cosine similarity of image-text representations is a reasonable approximation of the brain's construction of semantic representations. The superior predictive accuracy of CLIP embeddings observed in this study is consistent with current research trends in multimodal modeling, in which the use of contrastive language alignment to facilitate more robust predictions of activations in visual processing regions of the brain is widespread (Conwell et al., 2022).

The lack of difference in 2v2 accuracy between concrete and abstract word embeddings is surprising, as it seems to violate the prior assumption that abstract word representations contain primarily linguistic information. Given these prior claims from dual coding theory, the expectation was that multimodal embeddings of abstract words would not necessarily be made more predictive of brain activation by virtue of including visual information. The results presented here seem to suggest that incorporating visual information into word embeddings increases their representational alignment with cortical activations regardless of the concept's abstractness.

While current discourse surrounding the question of whether language models require grounding for understanding has yet to be definitively resolved (Patel and Pavlick, 2021; Pavlick, 2023), the findings presented here provide further empirical

support for the grounded cognition argument that multimodality should enhance the neural alignment of models' semantic representations. Multimodal embeddings achieved higher predictive accuracies than unimodal embeddings across all contexts (pictures and sentences), across all brain regions, and notably, across both abstract and concrete word types. This finding suggests that the enhanced neural alignment achieved by multimodality does not manifest solely for concepts that are easily visualized, but instead generalizes across both concrete and abstract words.

10 Data Accessibility

Code is made available at https://github.com/torreysnyder/Multimodal-Brain

11 References

References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. 2022. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126.

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Jeffrey R Binder, Chris F Westbury, Kristen A McKiernan, Edward T Possing, and David A Medler. 2005. Distinct brain systems for processing concrete and abstract concepts. *Journal of cognitive neuroscience*, 17(6):905–917.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46:904–911.

Colin Conwell, Jacob S Prince, Christopher J Hamblin, and George A Alvarez. 2023. Controlled assessment of clip-style language-aligned vision models in prediction of brain & behavioral data. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. 2022. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, pages 2022–03.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Evelina Fedorenko and Sharon L Thompson-Schill. 2014. Reworking the language network. *Trends in cognitive sciences*, 18(3):120–126.

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

Evi Hendrikx and Lisa Beinborn. 2020. The fluidity of concept representations in human brain signals. *arXiv preprint arXiv:2002.08880*.

Noussaiba Jaafar and Zied Lachiri. 2023. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, 211:118523.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S Bapi. 2022a. Visio-linguistic brain encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 116–133.

Subba Reddy Oota, Jashn Arora, Vijay Rowtula, Manish Gupta, and Raju S. Bapi. 2022b. Visio-linguistic brain encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 116–133, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(3):255.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(2251):20220041.

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. Association for Computational Linguistics.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080.

Jerry Tang, Amanda LeBel, and Alexander G Huth. 2021. Cortical representations of concrete and abstract concepts in language combine visual and linguistic representations. *bioRxiv*, pages 2021–05.

Mariya Toneva, Otilia Stretcu, Barnabás Póczos, Leila Wehbe, and Tom M Mitchell. 2020. Modeling task effects on meaning representation in the brain via zero-shot meg prediction. *Advances in Neural Information Processing Systems*, 33:5284–5295.

Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. 2022. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. *BioRxiv*, pages 2022–09.