



A transformer-Based neural language model that synthesizes brain activation maps from free-form text queries

Gia H. Ngo^{a,1,*}, Minh Nguyen^{a,1}, Nancy F. Chen^{b,2}, Mert R. Sabuncu^{a,c,2}

^a School of Electrical & Computer Engineering, Cornell University, USA

^b Institute for Infocomm Research (I2R), A*STAR, Singapore

^c Department of Radiology, Weill Cornell Medicine, USA

ARTICLE INFO

Keywords:

Coordinate-based meta-analysis
Transformers
Information retrieval
Image generation

ABSTRACT

Neuroimaging studies are often limited by the number of subjects and cognitive processes that can be feasibly interrogated. However, a rapidly growing number of neuroscientific studies have collectively accumulated an extensive wealth of results. Digesting this growing literature and obtaining novel insights remains to be a major challenge, since existing meta-analytic tools are constrained to keyword queries. In this paper, we present *Text2Brain*, an easy to use tool for synthesizing brain activation maps from open-ended text queries. *Text2Brain* was built on a transformer-based neural network language model and a coordinate-based meta-analysis of neuroimaging studies. *Text2Brain* combines a transformer-based text encoder and a 3D image generator, and was trained on variable-length text snippets and their corresponding activation maps sampled from 13,000 published studies. In our experiments, we demonstrate that *Text2Brain* can synthesize meaningful neural activation patterns from various free-form textual descriptions. *Text2Brain* is available at <https://braininterpreter.com> as a web-based tool for efficiently searching through the vast neuroimaging literature and generating new hypotheses.

1. Introduction

A rapidly growing number of functional magnetic resonance imaging (fMRI) studies have given us important insights into the mental processes that underpin behavior. However, individual studies are often power-restricted (Carp, 2012; Button et al., 2013), since the number of subjects and mental processes that can be interrogated in a single experiment is limited (Church et al., 2010). One approach to digest the vast literature and synthesize across many studies is to perform a meta-analysis of the reported results, such as the coordinates of the most significant effects (e.g., 3D location of peak brain activation in response to a task). These meta-analyses usually require expert curation of relevant experiments (e.g. Costafreda et al., 2008; Minzenberg et al., 2009; Shackman et al., 2011). A critical technical challenge here is the consolidation of synonymous terms. Importantly, over time, different denominations might be used in different contexts or invented to refine existing ideas. For instance, “self-generated thought”, one of the most highly studied functional domains of the human brain (Smallwood,

2013), can be referred to by varying terms, such as “task-unrelated thought” (Andrews-Hanna et al., 2014).

The selection of reported results for meta-analysis can be automated on data scraped from the published literature (Yarkoni et al., 2011; Dockès et al., 2020; Rubin et al., 2017). Two popular examples of this direction are Neurosynth (Yarkoni et al., 2011) and more recently Neuroquery (Dockès et al., 2020). Neurosynth utilizes automated keyword search to retrieve relevant studies and statistical tests to find summary brain activation maps corresponding to the keywords. Unlike Neurosynth, Neuroquery is a predictive model that synthesizes activation maps from keywords in the input query. Despite their differences in modeling, both Neurosynth and Neuroquery only support queries consisting predefined keywords. Furthermore, Neurosynth does not explicitly handle long queries, while Neuroquery relies on superficial lexical similarity via word co-occurrences for inference of longer or rarer queries. We propose an alternative approach named *Text2Brain*, which builds on recent neural language models and permits more flexible free-form text queries. *Text2Brain* captures a more fine-grained and

* Corresponding author.

E-mail address: ghn8@cornell.edu (G.H. Ngo).

¹ Indicates equal contribution.

² Indicates equal contribution.

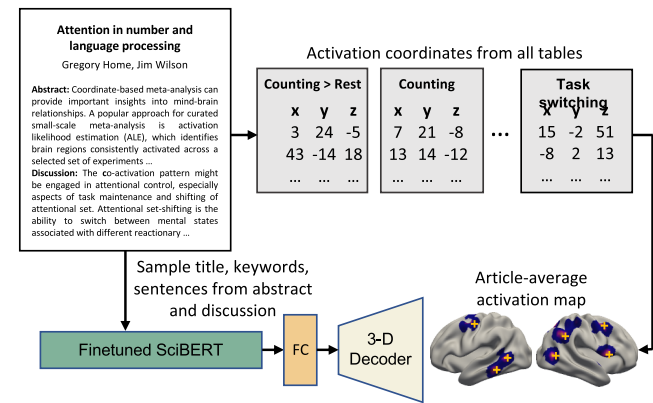


Fig. 1. Overview of data preprocessing, the Text2Brain model, and training procedure. All activation maps are 3D volumes, but projected to the surface for visualization.

Table 1

High-level comparison of approaches to meta-analytic brain maps generation.

	Neurosynth	Neuroquery	Text2Brain
Vocabulary	Fixed	Fixed	Unlimited
Handle of complex query	None	Lexical similarity	Semantic similarity
Predictive models	None	TF-IDF, linear regression	Transformer, convolution

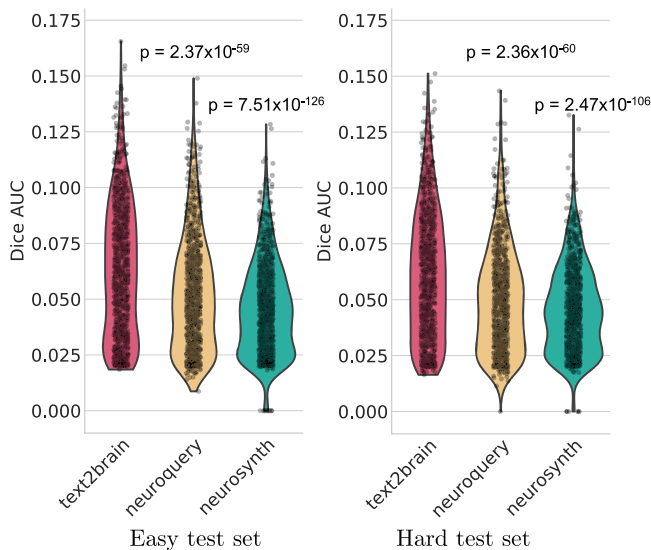


Fig. 2. Evaluation of article-average activation maps predicted from their titles measured in area under the Dice curve (AUC) score. The left and right graph show the Dice AUCs of samples from the easy and hard test sets, respectively (Section 3.1). The p-values are computed from paired-sample t-tests between Text2Brains and each of the 2 baselines.

implicit semantic similarity via vector representations from the neural language model in order to retrieve more relevant studies. Furthermore, in contrast to tools like Neuroquery, our method computes synthesized activation maps via a 3D convolutional neural network (CNN) model, which we empirically demonstrate, can capture coarse and fine details.

We compare Text2Brain's predictions with those from Neurosynth and Neuroquery, where we used article titles as free-form queries. Furthermore, we assess model predictions on independent test datasets, including reliable task contrasts and meta-analytic activation maps of well-studied cognitive domains predicted from their descriptions. Our

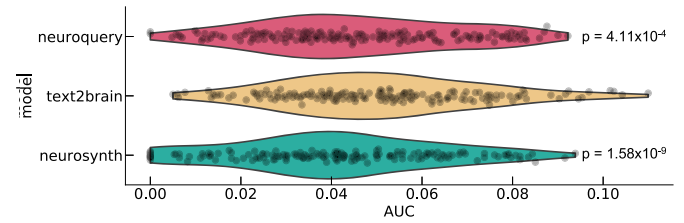


Fig. 3. Dice AUCs of predicted IBC task activation maps from contrasts' description. The p-values are estimated from paired-sample t-tests between Text2Brain against the two baselines.

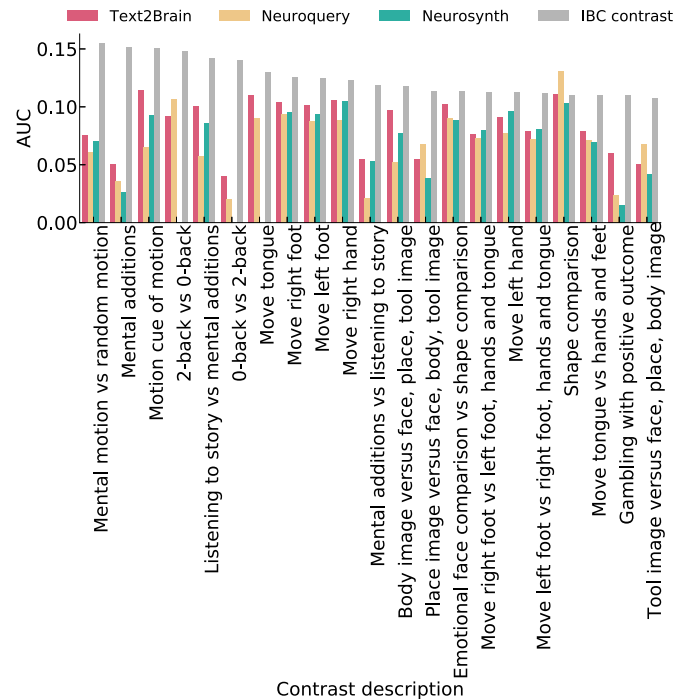


Fig. 4. Dice AUCs of predicted HCP task activation maps from contrasts' description. The graph includes 22 contrasts with the highest HCP-IBC's Dice AUC scores and sorted in decreasing order.

analysis shows that Text2Brain generates activation maps that better match the target images than the baselines tools. Given its flexibility in taking input queries, Text2Brain can be used as an educational aid as well as a tool for synthesizing maps based on published results or generating novel hypotheses for future research. Compared to our conference article (Ngo et al., 2021), we have extensively expanded our results and analysis. Specifically, we have expanded on the model validation on article titles with a different test set (Sections 3.1 and 4.1), added additional evaluation on the contrast maps predicted from their descriptions (Section 4.2). New results and discussion have also been added to this paper, including a high-level conceptual comparison of models (Section 2.7), new experiments on predicting representative meta-analytic results (Sections 3.3 and 4.2), and quantitative analysis of the models' robustness to input queries (Sections 3.4 and 5.2).

2. Datasets and methods

2.1. Model overview

Fig. 1 shows the overview of this work, including data generation, model architecture, and model training. The Text2Brain model has an encoder-decoder architecture that maps text sequences into brain

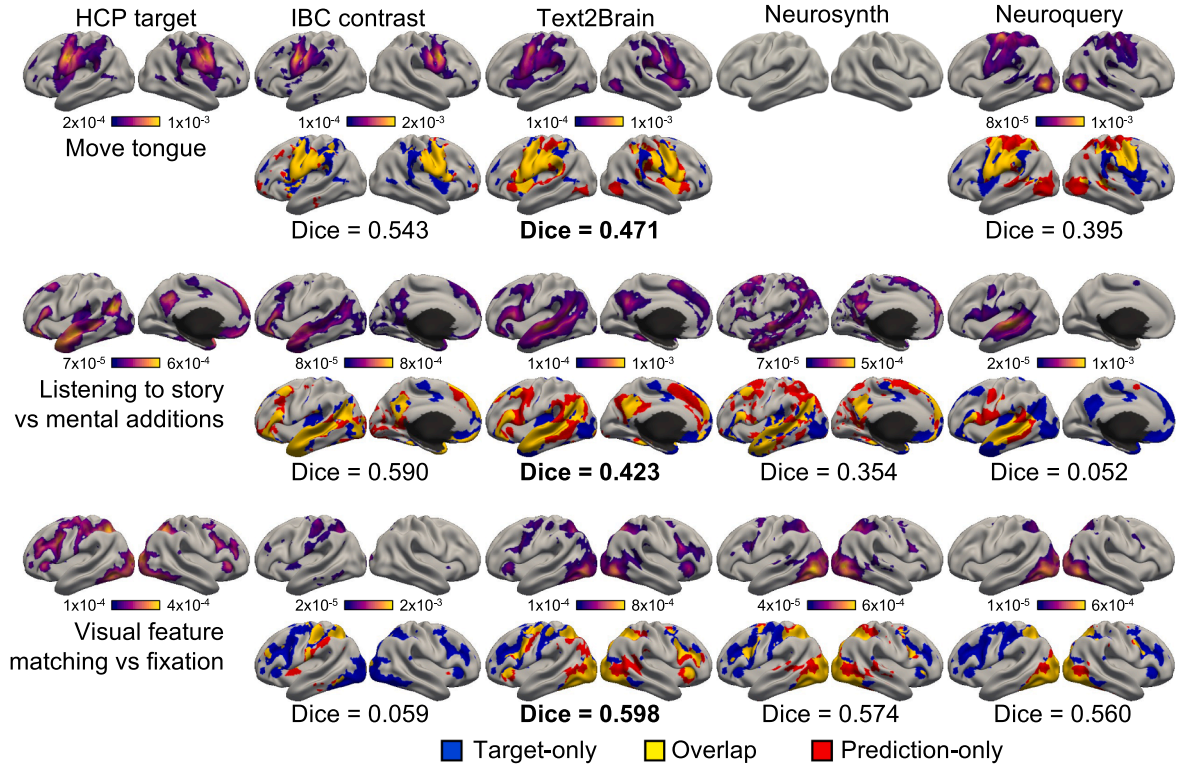


Fig. 5. Task activation maps predicted from contrasts' description. Each row shows both the thresholded maps of the top 25% most activated voxels (top) and the overlap between predicted and target binarized brain maps. Blue is activation in the target contrast, red is the predicted activation and yellow is the overlap. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

activation maps (Section 2.2). Its transformer-based encoder uses self-attention to encode a snippet of text input into vector representation (Vaswani et al., 2017; Devlin et al., 2018). Text2Brain's 3D convolutional decoder (CNN) then translates the vector representation into a 3D brain activation map. The Transformer is currently the most effective approach for modeling text since it can capture long-distance dependency between words and can learn efficiently through self-supervision from massive text corpora (Jawahar et al., 2019; Raffel et al., 2020). On the other hand, 3D CNNs are the most dominant architectural design in medical imaging (Milletari et al., 2016; Kamnitsas et al., 2017).

In our proposed approach, we first extract full text and activation coordinates from each research article to create data samples. Each sample consists of an input snippet from the full text and an output 3D activation map for each supported keyword, which are in turn linearly combined for queries. This approach can limit how many keywords are supported (Yarkoni et al., 2011; Dockès et al., 2020). On the other hand, Text2Brain stores the text and activation maps content in its parameters and can scale better to diverse input queries (Petroni et al., 2019). We use data augmentation to encourage Text2Brain to construct and store rich many-to-one mappings between textual description and activation maps (Section 2.4). This allows Text2Brain to better map semantically similar text queries to similar activation maps.

2.2. Implementation

Fig. 1 bottom left corner shows the Text2Brain model with its text encoder and 3D CNN image decoder. Text2Brain's text encoder is based on SciBERT, a BERT model that has been trained using scientific articles (Beltagy et al., 2019). BERT is a transformer-based model with bidirectional self-attention trained via self-supervision to learn semantic representations of textual input (Devlin et al., 2018). The text encoder outputs a vector representation of dimension 768. This vector is projected using a fully-connected layer and then reshaped to a 3D volume of dimension $4 \times 5 \times 4$ voxels with 64 channels at each voxel. The image decoder consists of 3 transposed 3D convolutional layers with 32, 16, 8 channels respectively. Text2Brain was trained using the Adam optimizer (Loshchilov and Hutter, 2018) and the mean-squared error with a batch size of 24 for 2000 epochs. The learning rate for the text encoder and image decoder are set at 10^{-5} and 3×10^{-2} respectively. The model's source code is available at <https://github.com/sabunculab/text2brain>.

2.3. Data preprocessing

We used the same set of 13,000 neuroimaging articles previously released in Dockès et al. (2020) in our experiments. Each article contains one or more tables of results that reported coordinates of peak activation in MNI152 coordinate system (Lancaster et al., 2007). The activation foci are also publicly released by Neuroquery (Dockès et al., 2020). Following the same procedure as Dockès et al. (2020), the set of activation foci associated with each table is used to generate an activation map by placing a Gaussian sphere with full width at half maximum (FWHM) of 9mm at each of the coordinates of peak activation. The chosen FWHM allows a fair comparison with Neuroquery (Dockès et al., 2020) in our experiments, and is consistent with previous work (Wager

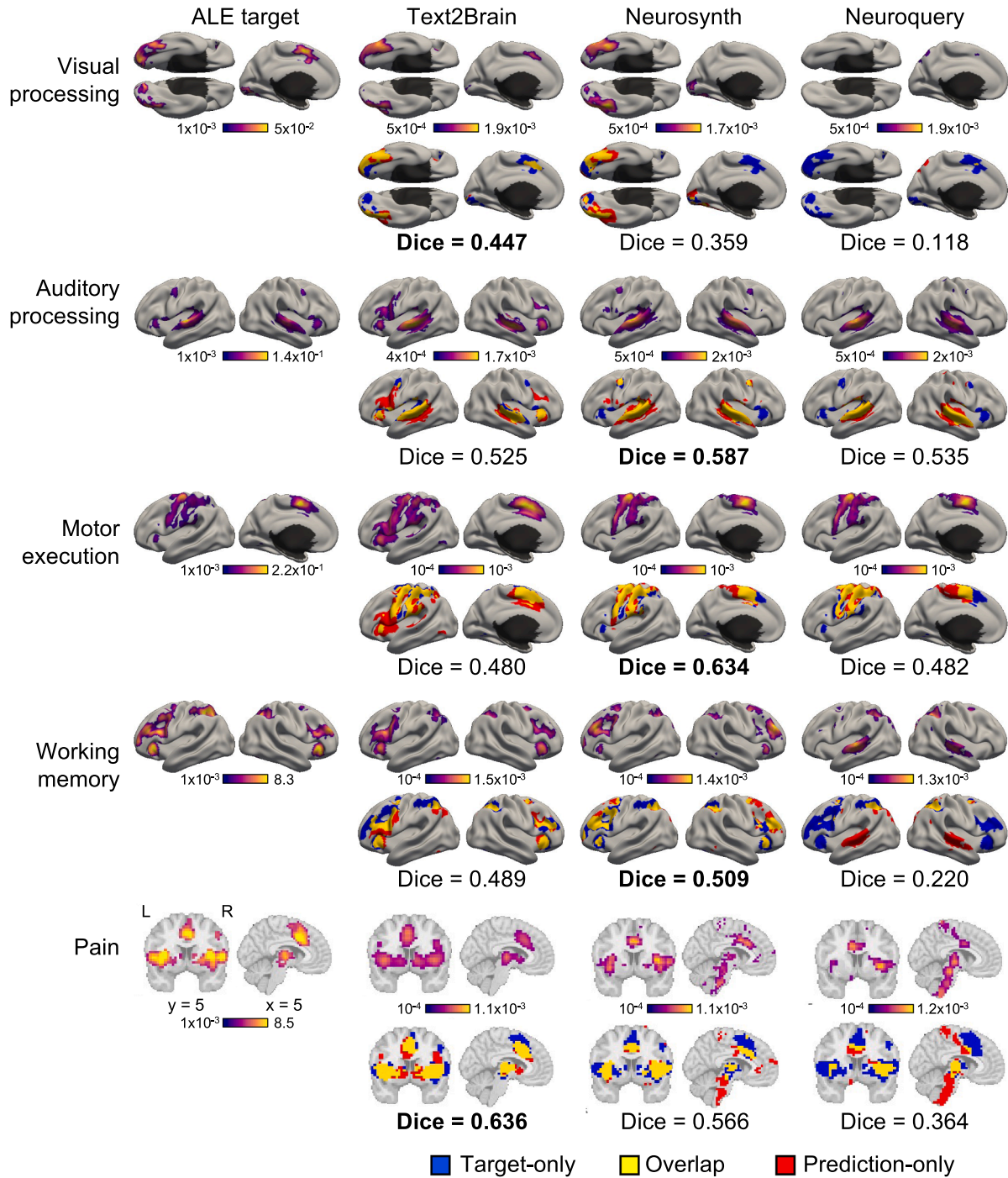


Fig. 6. Prediction of brain maps from meta-analytic studies of representative functional domains. The information of the investigated functional domains are listed in Table 2. Reference and predicted activation maps of the first 4 function domains are visualized on the brain surface. The last domain (“pain”) is visualized in the volume as most activation concentrates in the non-cortical parts of the brain. For all functional domains, Text2Brain generates reasonable activation maps and comparable with the baselines for the common functional domains.

et al., 2009; Yarkoni et al., 2011; Yeo et al., 2015). Supplemental section 6.6 shows an analysis of the effect of the Gaussian kernel’s FWHM used for preprocessing on Text2Brain’s predictive accuracy on an independent test set. This comparison confirms that the choice of the kernel’s FWHM is reasonable. An article-average activation map is also generated by averaging the activation maps of all the tables in the article. The

text associated with the activation maps are extracted from the articles’ full text. The articles’ full text are scraped using their PubMedID via the NCBI API³ and the Elsevier E-utilities API.⁴ As there may be multiple text snippets corresponding to the same activation map, the next section (Section 2.4) shows how the corresponding text of an activation map is

³ <https://www.ncbi.nlm.nih.gov/books/NBK25501/>

⁴ <https://dev.elsevier.com/>

Table 2

Meta-analytic studies of representative functional domains. The studies were selected from the ANIMA dataset (Reid et al., 2016) that have the most number of experiments and cover a diverse set of brain regions.

Functional domain	#Exp	Search queries
Visual processing (Heckner et al., 2021)	114	visual processing face monitor face discrimination film viewing fixation flashing checkerboard passive viewing visual object identification visual pursuit visual tracking visuospatial attention
Auditory processing (Heckner et al., 2021)	122	auditory processing divided auditory attention music comprehension oddball discrimination passive listening phonological discrimination pitch monitor pitch discrimination tone monitor tone discrimination
Motor execution (Heckner et al., 2021)	251	motor execution writing chewing swallowing drawing isometric force motor learning grasping finger tapping button press flexion extension
Working memory (Rottschy et al., 2012)	189	working memory short-term memory
Pain (Xu et al., 2020)	222	pain noxious nociception

selected.

2.4. Training

Each training sample consists of a text-activation map pair and correspond to an neuroimaging article. The activation map is sampled uniformly at random from the union set of table-specific maps and article-average map. For each table-specific map, the first sentence of the corresponding table caption is chosen as the map's associated text. Our initial data exploration suggested that the first sentence to be the most relevant description of the activation map. For each article-average map, the associated text that describes the activation map is sampled uniformly at random from the following four sources: (1) the article's title; (2) one of the article's keywords; (3) the article's abstract; and (4) a randomly chosen subset of sentences from the discussion section of the article. This data augmentation strategy encourages Text2Brain to generalize over input texts of different lengths. Furthermore, matching the same activation pattern with multiple different text snippets encourages the model to recognize important words common across the snippets and to learn the association between different but synonymous words. Supplemental Fig. 12 shows our ablation study on the sampling strategy. The liberal (and likely noisy) construction of image-text pairs appears to perform better than more deliberate coupling of image-text snippets strategies (not reported) that we tried in our preliminary experiments. We surmise that simply presenting different text snippets to a target brain image is analogous to another augmentation strategy that allows the neural network to pool across samples and learn the relevant words and their weights with respect to the target brain maps. Training with the set up in Section 2.2 takes approximately 75 h on one Nvidia RTX GPU while one inference pass with an input query of up to 140 characters takes less than 1 s.

2.5. Baselines

We compare Text2Brain to 2 different baselines: Neurosynth (Yarkoni et al., 2011) and Neuroquery (Dockès et al., 2020). For a keyword, Neurosynth first finds all neuroimaging articles that mention that keyword. Then, one statistical test per voxel is performed across the activation maps corresponding to those studies to determine a significant association. Since Neurosynth was not formulated to handle multiple-word queries, for such query, we performed statistical test using activation maps from all articles that contain at least one of the keywords in the query.

Neuroquery extends Neurosynth's vocabulary of keywords by including more curated keywords from lexicons such as MeSH,

NeuroNames, and NIF (Lipscomb, 2000; Bowden and Martin, 1995; Gardner et al., 2008). The keyword encoding is obtained after performing non-negative matrix factorization of the articles' full text (as a bag of keywords) represented with term frequency - inverse document frequency (TF-IDF) features (Salton and Buckley, 1988). A ridge regression model was trained to map the text encoding to the activation. The inference of a keyword is smoothed by a weighed average of its most related keywords (in the TF-IDF space). For multiple-word queries, the predicted activation map is obtained by averaging the activation maps from all keywords in the input, weighed by the coefficients learnt during training.

2.6. Evaluation metrics

For thresholded target activation maps such as those computed by ALE (Eickhoff et al., 2009), the predicted brain maps are thresholded to retain the same number of most activated voxels as the target. For example, given an estimated activation map by ALE with statistically significant clusters of activation that cover 25% of the brain volume, the brain maps predicted by Text2Brain, Neuroquery, and Neurosynth are also thresholded to retain the top 25% most activated voxels in each map. The accuracy of prediction is measured by Dice score (Dice, 1945) which quantifies the extent of overlap between the predicted and target brain maps (details are in Supplemental Section 6.2).

Furthermore, we use Dice scores at different thresholds to estimate the similarity between predicted and target activation maps at different levels of detail (Ngo et al., 2022). This evaluation procedure is similar to that used in Dockès et al. (2020) for a thresholded target map, but we apply the same thresholding to both the target and predicted map. For example, at 5% threshold (considering the 5% most activated voxels), the Dice score measures the correspondence of the fine-grained details between the target and predicted activation maps. At higher thresholds (e.g. 25%), the score captures the gross agreement between activation clusters. We also estimated the area under the Dice curve (AUC) as a summary measure using approximated integration of Dice scores across all thresholds from 5% up to 30%. Supplemental Fig. 9 shows the Dice curve for an example pair of target-predicted activation maps. Note that the range of thresholds in the x-axis also conveys the maximum percentage of the gray matter mask that has an activation in the target brain map. For example, if only a proportion of gray matter mask has activation, such as the case of Neuroquery prediction that mostly extends up to 30% of the gray matter mask or a sparse target activation pattern from the coordinate-based meta-analysis, the x-axis range will not be extended up to 1.

In our experiments, all evaluation is performed in the MNI152 volumetric space, which is the original space of all predicted maps. For visualization, with activation maps that mostly concentrate in the cerebral cortex, the original volumetric images are transformed from MNI152 space to fs_LR surface space using Connetome Workbench (Van Essen et al., 2013) via the FreeSurfer surface space (Buckner et al., 2011; Fischl, 2012), with isolated surface clusters of less than 20 vertices being removed (Wu et al., 2018). Activation maps with significant activation in the non-cortical parts of the brain are visualized by cross-sectional slices with significant activation using Nilearn (Abraham et al., 2014).

2.7. High-level model comparison

Text2Brain can better handle input text than prior approaches because its vocabulary is not limited to a fixed pre-defined set of words (Table 1). In contrast, Neurosynth and Neuroquery rely on fixed word vocabularies and cannot predict for queries consisting of out-of-vocabulary words. Besides, Neurosynth's and Neuroquery's vocabularies are not sufficiently extensive, covering only a fraction (under 10%) (Dockès et al., 2020) of terms in relevant neuroimaging lexicons such as Cognitive Atlas (Poldrack and Yarkoni, 2016) and NeuroNames (Bowden and Martin, 1995). Text2Brain's usage of byte-pair encoding

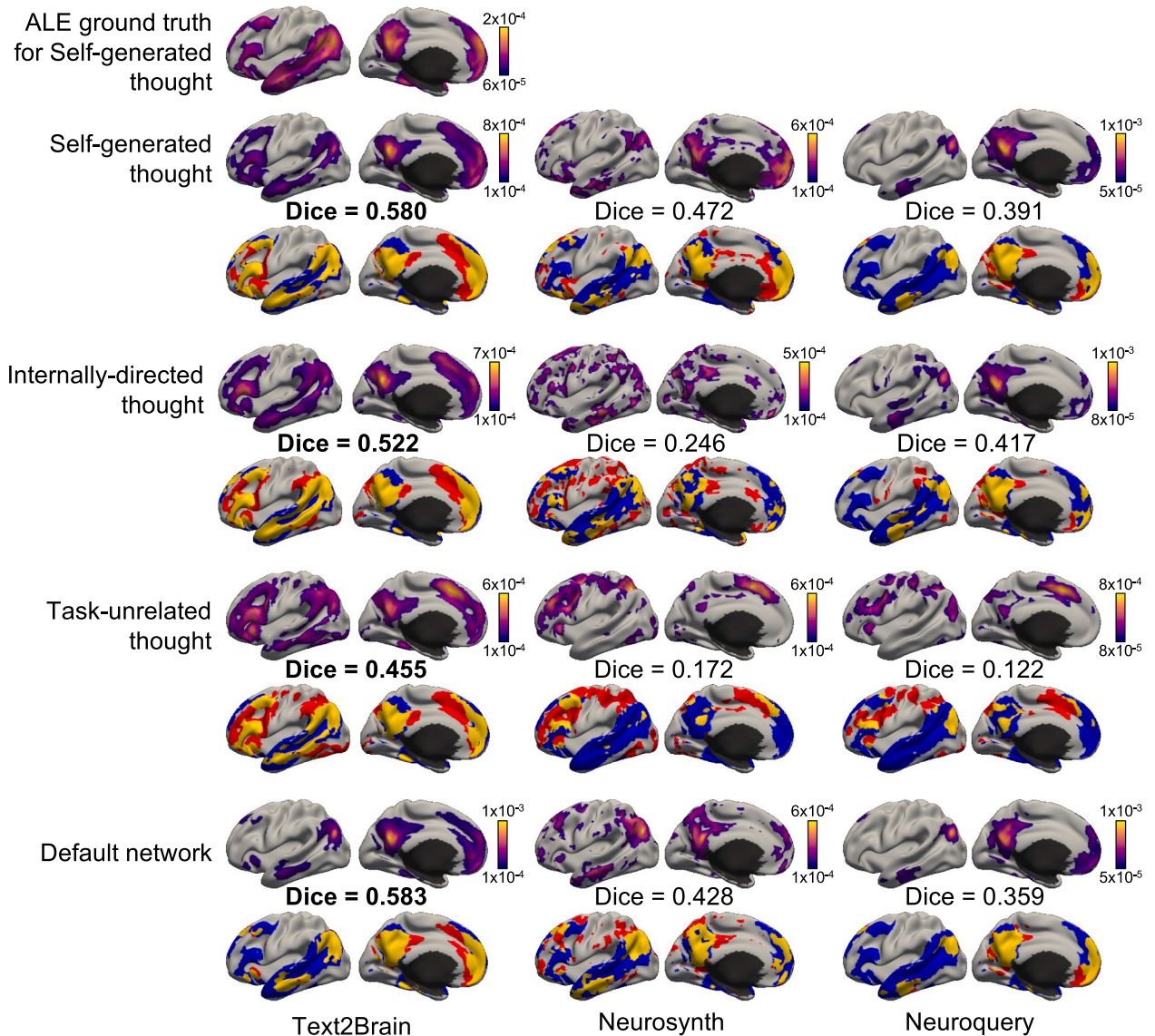


Fig. 7. Prediction of self-generated thought activation map using synonymous queries. While Text2Brain generates consistent prediction across the similar queries, Neurosynth and Neuroquery's prediction deteriorate on the "internally-directed thought" and "task-unrelated thought" queries.

enables the model to handle infrequent and out-of-vocabulary words more gracefully, by breaking down those words into digestible sub-word tokens (Sennrich et al., 2016). Hence, Text2Brain's vocabulary is open ended and can scale with training data to be unlimited in theory. Besides, Text2Brain's training is not limited to only training set data. Text2Brain can leverage self-supervised learning from non-neuroimaging scientific articles, as well as neuroimaging articles that do not report activation coordinates to learn a better text-to-activation-map transformation. By finetuning a SciBERT text encoder pretrained on the larger dataset of scientific articles (including non-neuroimaging articles), Text2Brain seems to converge on an optimum with a more useful representational space of the input text. Supplemental Section 6.4 shows the comparison between the Text2Brain model that uses pretrained SciBERT text encoder versus a randomly initialized text encoder. Evaluation on predicting article-average activation maps from both sets of test articles in the Neuroquery dataset (similar to Section 3.1) suggests that pretraining benefits the Text2Brain's performance. Furthermore, Text2Brain uses contextualized text embeddings to model semantic relationship between words so it can

deal with nuanced queries more effectively. Methods such as Neurosynth and Neuroquery may have difficulty dealing with complex expressions. By simply averaging the keywords' activation maps to arrive at the prediction for a complex query, these methods may fail to account for relationship between words in the query, such as order and semantic. Lastly, while the predictive approach of Neuroquery constructs the predicted activation map by modelling voxels' activation independently, Text2Brain generates the whole-brain activation with a 3D convolutional decoder that takes in the text encoding produced by the language model. By upsampling and computing the whole-brain activation from a bottleneck, Text2Brain can better model both the short and long-distance relationship between voxels.

3. Experimental setup

3.1. Predict activation maps from article title

Two test sets were created from the Neuroquery dataset of 13,000 studies. The first test set consists of 1000 randomly sampled articles. The

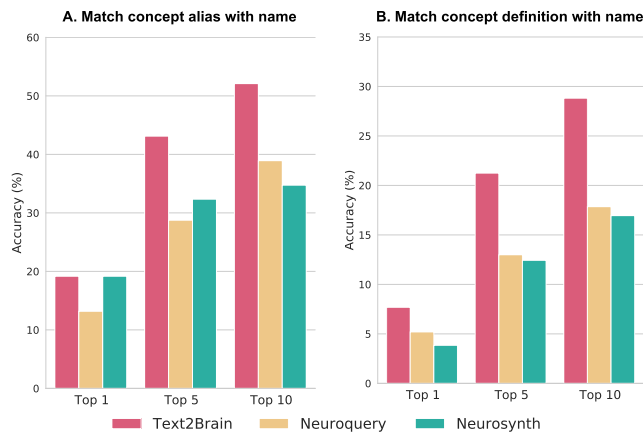


Fig. 8. Accuracy of matching Cognitive Atlas concept names with their description and aliases using models' predicted brain maps.

second test set also consists of 1000 articles but was randomly sampled such that the keywords (defined by the articles' authors) do not appear in the training and validation articles. The two test sets are labeled as easy and hard test sets respectively. Of the remaining articles, 1000 are randomly held out as a validation set for parameters tuning. For each article, the article-average activation map is predicted from its title using Text2Brain, as well as the Neurosynth and Neuroquery baselines. Both Text2Brain and Neuroquery were trained on the 10,000 articles in the training set. The Text2Brain model is trained using both the articles' titles and samples from the full-text, while Neuroquery is trained on the articles' full-text. We use predictions from the publicly available Neurosynth model at <https://neurosynth.org>, which was trained on the articles' abstracts. Note that Neurosynth is not a predictive model meant for out-of-sample prediction, but for performing automated statistical testing of associations between terms and brain locations.

3.2. Predict activation maps from contrast descriptions

3.2.1. Individual brain charting (IBC) task contrasts

The Individual Brain Charting (IBC) project (Pinho et al., 2020) estimates an extensive functional atlas of the human brain via fMRI data of subjects measured under a large number of task conditions. In particular, the IBC dataset consists of 180 task contrasts measured on 12 subjects. We use the activation maps provided by the IBC project to measure the predictive accuracy of Text2Brain and the two baselines over a wide range of functional domains, given the contrast descriptions from IBC.

3.2.2. Human connectome project (HCP) task contrasts

While the IBC dataset offers a large number of reference brain maps, the small number of subjects might make some results less reliable. We also utilized the Human Connectome Project (HCP) data both for reference and a measure of reliability of target maps. The HCP dataset consists of neuroimaging data from over 1200 subjects, including task fMRI (tfMRI) of 86 task contrasts from 7 domains (Barch et al., 2013), which overlap with 43 contrasts under the IBC dataset. We evaluate the model prediction of HCP task contrasts from their descriptions. While HCP provides detailed descriptions of task contrasts, we opt for the more concise contrast descriptions provided by the Individual Brain Charting (IBC) as they are more succinct and thus more favorable to the baselines. The IBC contrast descriptions are extracted from the metadata of the activation maps released on Neurovault <https://neurovault.org/images/360528>. The list of all IBC description of HCP contrasts are included in Supplemental Table 6.1. On the other hand, the target (ground-truth) activation maps are the HCP group-average contrast maps, as the large number of subjects provides more reliable estimates of the contrast

maps. In the analyses of this experiment, we use the agreement between the IBC and HCP maps as a measure of reliability. Despite using similar protocols, there are subtle differences between the IBC and HCP experiments. For instance, the original HCP language task was conducted in English but the corresponding language task in the IBC project was conducted in French.

3.3. Predict representative meta-analytic brain maps

The automated approach to brain map generation of Text2Brain and the 2 baselines are compared against published brain maps created from a manually curated set of meta-analyses. In particular, 5 cognitive concepts and their corresponding activation maps of 5 representative meta-analytic studies from ANIMA database (Reid et al., 2016) were selected. The 5 meta-analytic studies were selected for having the most number of experiments and their different coverage of the human brain. The cognitive processes of interest are visual processing, auditory processing, motor execution (Heckner et al., 2021), working memory (Rottschy et al., 2012), and pain (Xu et al., 2020). Each study searches for published neuroimaging studies that contain a set of texts queries relevant to the cognitive concept of interest. For example, in Rottschy et al. (2012), the phrases to search for working memory-related studies are "working memory" and "short-term memory". The same text queries for discovering relevant studies in the original meta-analysis were used as input to Neurosynth, Neuroquery, and Text2Brain. Table 2 shows the search queries and the number of experiments included in the original meta-analysis of the 5 chosen cognitive concepts. Activation maps generated from all text input queries corresponding to each cognitive concept are averaged to yield a single brain map for each model. The reference brain images for comparison are the activation maps released by the studies and made publicly available on ANIMA. The reference activation maps are produced by Activation Likelihood Estimation (ALE) (Turkeltaub et al., 2002; Laird et al., 2005; Eickhoff et al., 2009) and thresholded to retain only the statistically significant clusters of activation. For all reference ALE maps, the cluster-level forming threshold at voxel-level is $p < 0.001$ and cluster-level corrected threshold is set at $p < 0.05$ by the original authors (Eickhoff et al., 2012). For comparison, the generated brain maps are thresholded to keep the same number of survived voxels as those in the reference activation maps. The accuracy of each model's generated brain map is evaluated as the Dice score between the (thresholded) generated brain map and the target (thresholded) brain map (see Section 2.6).

3.4. Evaluate robustness of model prediction to semantically-equivalent queries

With the continual improvement of our understanding of the human brain and mind, neuroscientific knowledge is also an ever evolving repertoire. Several neuroimaging concepts have also been changing, adapting and broadening over time. Thus, we were interested in examining if our approach is robust to semantically equivalent queries. For example, "self-generated thought", one of the most intensively examined cognitive domains in neuroscience, has had its definition refined and assigned different denominations over the years. As a cognitive paradigm, different names have been used to refer to the set of inward-oriented psychological processes, such as "self-generated thought" (Smallwood, 2013), or "task-unrelated thought" (Andrews-Hanna et al., 2014). Both terms are associated with "default network" (Buckner et al., 2008), the set of brain regions with elevated activation when subjects are not subjected to any external stimulus.

To assess models' prediction of synonymous queries, we utilized the ontology from the Cognitive Atlas (Poldrack et al., 2011; Bilder et al., 2009). The Cognitive Atlas is a collaborative knowledge base for neuroscience with content such as cognitive concepts, their description and synonyms (aliases) contributed by the project's voluntary participants (Miller et al., 2010). At the time of our experiments, Cognitive

Atlas includes 885 concepts with definition, 108 of which have at least one alias. We considered a model to be robust with respect to a specific cognitive concept's definition if the activation map predicted from the description matches the predicted map from the concept's name. In particular, given a model's predicted brain maps from all 885 Cognitive Atlas concept names and their description, we assess if the model's brain map predicted from a concept's definition is one of the k maps (out of 886 possible maps) most similar to the model's brain map predicted from the concept's name. In our experiments, top-1, top-5 and top-10 matching accuracy were evaluated using Dice AUC metrics. The different values of k 's account for the uncertainty of the concepts' natural language text, e.g., different contributors might use different names to refer to the same concept. Similarly, models' robustness with respect to a cognitive concept's alias is measured by the accuracy of matching the activation maps predicted from the text of a concept's alias and its name.

4. Results

4.1. Validation of activation maps predicted from article title

Fig. 2 shows the quality of activation maps predicted from the titles of 1000 articles in each of the two test sets (Section 3.1). In the easy test set (the test articles' keywords can overlap with the training articles'), the proposed Text2Brain model (mean Dice AUC = 0.0636) outperforms Neuroquery (mean Dice AUC = 0.0523) and Neurosynth (mean Dice AUC = 0.0453). In the hard test set (the test articles' keywords are not present in the training set), the Text2Brain model (mean Dice AUC = 0.0609) also performs better than Neuroquery (mean AUC = 0.0499) and Neurosynth (mean AUC = 0.0457). Paired-sample t -tests show that the performance differences in both test sets are statistically very significant. The p -values when comparing Neuroquery and Neurosynth are $p = 5.25 \times 10^{-27}$ and $p = 2.40 \times 10^{-12}$. Fig. 2 also indicates how the different models handle out-of-sample input text. Text2Brain can make a prediction for all input texts, evident with positive Dice AUCs for all samples. On the other hand, Neurosynth fails to make prediction for some article titles in both test sets, resulting in zero Dice AUCs for such samples. Similarly, Neuroquery fails to make prediction for some samples in the hard test set. These failure cases are caused by the limited vocabularies of Neurosynth and Neuroquery that cannot cover the words in the test input queries. On the other hand, the language model of Text2Brain is finetuned from SciBert, which has been pretrained on a broader lexicon and utilizes sub-word tokens to extend the vocabulary to unseen words (more details in Section 2.7).

4.2. Prediction of task contrast maps from description

Fig. 3 shows the Dice AUC scores for the prediction of Text2Brain, Neuroquery and Neurosynth against the IBC group-average task contrast maps. Text2Brain (mean Dice AUC = 0.0507) improves upon both Neuroquery (mean Dice AUC = 0.0457, $p = 4.11 \times 10^{-4}$), and Neurosynth (mean Dice AUC = 0.0404, $p = 1.58 \times 10^{-9}$). The p -values are measured by 2-tail paired-sample t -test between Text2Brain and the two baselines.

Fig. 4 shows the AUC scores for the prediction of the three models and the IBC average contrasts, against the HCP target maps. The 22 contrasts with above-average HCP-IBC's AUC scores, considered to be the reliable contrasts, are shown. Across all 43 HCP contrasts, Text2Brain (mean AUC = 0.082) performs better than the baselines, i.e. Neuroquery (mean AUC = 0.0755, $p = 0.08$), Neurosynth (mean AUC = 0.047, $p = 1.5 \times 10^{-5}$), where p -values are computed from the paired t -test between Text2Brain's and the baselines' prediction. As reference, IBC contrasts yield a mean AUC = 0.094 when compared to the corresponding HCP maps (Statistical comparison with Text2Brain, $p = 0.077$).

Fig. 5 shows the prediction for three contrasts correspond to different HCP task groups, namely "MOTOR", "LANGUAGE", "RELATIONAL" thresholded at the top 25% most activated voxels. The three task groups were chosen to show results for a range of target images with different levels of reliability. The two task groups "MOTOR" and "LANGUAGE" are the two most reliable task (having the highest average HCP-IBC AUC across all contrasts), while "RELATIONAL" has the lowest average HCP-IBC AUC. Text2Brain's prediction improves over the baselines for the three contrasts. Neurosynth was not able to generate activation maps for one of the contrast descriptions ("Move tongue"). On the other hand, for the "Move tongue" contrast, Neuroquery predicts activation in the primary cortex, but the peak is in the wrong location, shifted more toward the hand region of the homunculus. Additionally, there is a false positive prediction in the occipital cortex, which might be an artifact from modeling brain activation coupled with visual stimuli-related words describing the motor experiments.

4.3. Prediction of brain maps from representative meta-analytic studies

Fig. 6 shows the prediction of activation maps for 5 representative meta-analytic studies with the most number of experiments from ANIMA (Reid et al., 2016). Among the three models, Neuroquery has the lowest Dice score on average, with prediction on "Visual processing", "Working memory", and "Pain" that significantly deviates from the target maps. On the other hand, Neurosynth-derived brain maps consistently match well against the target maps. The high accuracy of Neurosynth prediction is expected since the five chosen cognitive concepts are among the most commonly studied concepts with the most number of experiments reporting activation coordinates in the literature. Given high number of available experiments and the input queries mostly exist in Neurosynth's predefined keyword set, the activation coordinates scraped by automated method by Neurosynth would be very similar to the manually curated data in the original meta-analysis. Lastly, Text2Brain also predicts consistently reasonable brain maps for all five cognitive concepts, and matches the target maps better than Neurosynth for "Visual Processing" and "Pain". Results in Fig. 6 shows that Text2Brain could learn appropriate relationship between common search phrases and the activation pattern of a diverse set of functional domains.

5. Robustness of models to input queries

5.1. Example of "self-generated thought" synonyms

We examine the prediction for "self-generated thought", which is one of the most extensively investigated functional domains, due to its involvement in a wide range of cognitive processes that do not require external stimuli (Andrews-Hanna et al., 2014), and is associated with the default network (Buckner et al., 2008). The ground-truth map for self-generated thought, taken from (Ngo et al., 2019), is estimated using activation likelihood estimation (ALE) (Eickhoff et al., 2009) applied on activation foci across 167 imaging studies of 7 tasks selected based on strict criteria (Spreng et al., 2009; Mar, 2011; Sevinc and Spreng, 2014). The resulting ALE map is thresholded with the cluster-level forming threshold at voxel-level $p < 0.001$, and cluster-level corrected threshold $p < 0.05$ (Eickhoff et al., 2012).

Fig. 7 shows the prediction of self-generated thought activation map using four different query terms, thresholded to retain the same number of activated voxels as the target map.

Across all four queries, Text2Brain's prediction best matches the ground-truth activation map compared to the baselines. For the "self-generated thought" and "default network" queries, all approaches generate activation maps that are consistent with the ground-truth, which includes the precuneus, the medial prefrontal cortex, the temporo-parietal junction, and the temporal pole. Text2Brain and Neuroquery both make reasonable prediction from the "internally-

directed thought” query while Neurosynth’s prediction is largely scattered and does not match the target map. Lastly, Text2Brain can also replicate a similar activation pattern to the target from the query “task-unrelated thought”, evident by only a slight drop in the Dice score. However, Neuroquery and Neurosynth both generate activation maps that differ from the typical default network’s regions, such as activation in the prefrontal cortex, and also result in a large drop of the Dice scores.

5.2. Prediction of cognitive atlas concepts from synonymous queries

Fig. 8 shows the accuracy of matching cognitive concept names from the Cognitive Atlas (Poldrack et al., 2011) with their definitions and atlases using the different models’ predicted brain maps. Prediction by Text2Brain is more robust than both Neuroquery and Neurosynth with respect to the concept definition and alias. In particular, Text2Brain has the same top-1 accuracy of matching the brain map predicted from a concept’s alias with the prediction from the concept name compared to Neurosynth. This result is expected given that Neurosynth can yield accurate brain map for keywords that are included in their vocabulary. In contrast, Text2Brain improves over Neurosynth for top-1 accuracy of matching concept name with the longer text of concept definition. Text2Brain is more robust than both Neurosynth and Neuroquery baselines in terms of top-5 and top-10 matching accuracies for both concept aliases and definitions. Fig. 8 indicates that Text2Brain prediction is robust to natural language text queries of different length and complexity.

6. Conclusion

In this work, we present a model named Text2Brain for generating activation maps from free-form text query. By finetuning a high-capacity SciBERT-based text encoder to predict coordinate-based meta-analytic maps, Text2Brain captures the rich relationship in the language representational space, allowing the model to generalize its prediction for synonymous queries. This is evident in the better performance of Text2Brain in predicting the self-generated thought activation map using different descriptions of the functional domain. Text2Brain’s capability to implicitly learn relationships between textual terms and images ensures the model can remain relevant and useful even as neuroimaging literature continues to evolve with new discoveries and rephrasing of existing concepts. We also show that Text2Brain accurately predicts most of the task contrasts included in the IBC and HCP dataset, validating its capability to make prediction for longer, arbitrary queries. Text2Brain also preempts failure cases in Neurosynth and Neuroquery, where they cannot predict input queries undefined in the vocabulary list, even though these queries are relevant to neuroscience research (e.g. title of an article). On the other hand, we also observed that Text2Brain had difficulties handling queries that involve logical reasoning, such as the direction of a contrast. For example, while queries such as “A vs B” and “B vs A” can be inferred by human to correspond with inverted activation maps, Text2Brain sometimes treats one direction to be the same as the other. We suspect that this type of error is likely due to the model’s inability to generalize “vs” as an “subtractive” operator. Resolving such limitation will likely require modifications to the language model. Furthermore, in the future, we plan to enhance the interpretability of our approach, such as to attribute regions of activations in the generated map to specific words in the input query, as well as to efficiently match activation maps and scientific descriptions most relevant to the synthesized images.

We believe that the flexibility of Text2Brain can significantly lower the barrier for researchers at all stages of their careers to synthesize brain activation maps needed for their research. For example, the ability of Text2Brain to generate meaningful neural activation patterns of synonymous queries for a functional domain can improve the accuracy of delineating region-of-interests (ROIs) relevant to the functional process, as well as to assess the reliability of each ROI. Discovery of these

ROIs is useful for several applications such as meta-analytic connectivity modeling (MACM) (Laird et al., 2013). We look forward to such application of Text2Brain in aiding future neuroscientific research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by NIH grants R01LM012719, R01AG053949, the NSF NeuroNex grant 1707312, the NSF CAREER 1748377 grant and Jacobs Scholar Fellowship.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2022.102540

References

- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., 2014. Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* 8, 14.
- Andrews-Hanna, J.R., Smallwood, J., Spreng, R.N., 2014. The default network and self-generated thought: component processes, dynamic control, and clinical relevance. *Ann. N. Y. Acad. Sci.* 1316 (1), 29.
- Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., et al., 2013. Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage* 80, 169–189.
- Beltagy, I., Lo, K., Cohan, A., 2019. SciBERT: a pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.
- Bilder, R.M., Sabb, F.W., Parker, D.S., Kalar, D., Chu, W.W., Fox, J., Freimer, N.B., Poldrack, R.A., 2009. Cognitive ontologies for neuropsychiatric phenomics research. *Cogn. Neuropsychiatry* 14 (4–5), 419–450.
- Bowden, D.M., Martin, R.F., 1995. Neuronames brain hierarchy. *NeuroImage* 2 (1), 63–83.
- Buckner, R. L., Andrews-Hanna, J. R., Schacter, D. L., 2008. The brain’s default network: anatomy, function, and relevance to disease.
- Buckner, R.L., Krienen, F.M., Castellanos, A., Diaz, J.C., Yeo, B.T., 2011. The organization of the human cerebellum estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106 (5), 2322–2345.
- Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14 (5), 365–376.
- Carp, J., 2012. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage* 63 (1), 289–300.
- Church, J.A., Petersen, S.E., Schlaggar, B.L., 2010. The “task B problem” and other considerations in developmental functional neuroimaging. *Hum. Brain Mapp.* 31 (6), 852–862.
- Costafreda, S.G., Brammer, M.J., David, A.S., Fu, C.H., 2008. Predictors of amygdala activation during the processing of emotional stimuli: a meta-analysis of 385 pet and fmri studies. *Brain Res. Rev.* 58 (1), 57–70.
- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dice, L.R., 1945. Measures of the amount of ecologic association between species. *Ecology* 26 (3), 297–302.
- Dockès, J., Poldrack, R.A., Primet, R., Gözükan, H., Yarkoni, T., Suchanek, F., Thirion, B., Varoquaux, G., 2020. NeuroQuery, comprehensive meta-analysis of human brain mapping. *Elife* 9, e53385.
- Eickhoff, S.B., Bzdok, D., Laird, A.R., Kurth, F., Fox, P.T., 2012. Activation likelihood estimation meta-analysis revisited. *NeuroImage* 59 (3), 2349–2361.
- Eickhoff, S.B., Laird, A.R., Grefkes, C., Wang, L.E., Zilles, K., Fox, P.T., 2009. Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* 30 (9), 2907–2926.
- Fischl, B., 2012. Freesurfer. *NeuroImage* 62 (2), 774–781.
- Gardner, D., Akil, H., Ascoli, G.A., Bowden, D.M., Bug, W., Donohue, D.E., Goldberg, D. H., Grafstein, B., Grethe, J.S., Gupta, A., et al., 2008. The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics* 6 (3), 149–160.
- Heckner, M.K., Cieslik, E.C., Küppers, V., Fox, P.T., Eickhoff, S.B., Langner, R., 2021. Delineating visual, auditory and motor regions in the human brain with functional neuroimaging: a brainmap-based meta-analytic synthesis. *Sci. Rep.* 11 (1), 1–11.

- Jawahar, G., Sagot, B., Seddah, D., 2019. What does bert learn about the structure of language? *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.
- Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78.
- Laird, A.R., Eickhoff, S.B., Rottschy, C., Bzdok, D., Ray, K.L., Fox, P.T., 2013. Networks of task co-activations. *NeuroImage* 80, 505–514.
- Laird, A.R., Fox, P.M., Price, C.J., Glahn, D.C., Uecker, A.M., Lancaster, J.L., Turkeltaub, P.E., Kochunov, P., Fox, P.T., 2005. ALE meta-analysis: controlling the false discovery rate and performing statistical contrasts. *Hum. Brain Mapp.* 25 (1), 155–164.
- Lancaster, J.L., Tordesillas-Gutiérrez, D., Martínez, M., Salinas, F., Evans, A., Zilles, K., Mazziotta, J.C., Fox, P.T., 2007. Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. *Hum. Brain Mapp.* 28 (11), 1194–1205.
- Lipscomb, C.E., 2000. Medical subject headings (mesh). *Bull. Med. Libr. Assoc.* 88 (3), 265.
- Loshchilov, I., Hutter, F., 2018. Decoupled weight decay regularization. *Proceedings of ICLR*.
- Mar, R.A., 2011. The neural bases of social cognition and story comprehension. *Annu. Rev. Psychol.* 62, 103–134.
- Miller, E., Seppa, C., Kittur, A., Sabb, F., Poldrack, R., 2010. The cognitive atlas: employing interaction design processes to facilitate collaborative ontology creation. *Nat. Preced.*, 1 <https://doi.org/10.1038/npre.2010.4532.1>.
- Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.
- Minzenberg, M.J., Laird, A.R., Thelen, S., Carter, C.S., Glahn, D.C., 2009. Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. *Arch. Gen. Psychiatry* 66 (8), 811–822.
- Ngo, G.H., Eickhoff, S.B., Nguyen, M., Sevinc, G., Fox, P.T., Spreng, R.N., Yeo, B.T., 2019. Beyond consensus: embracing heterogeneity in curated neuroimaging meta-analysis. *NeuroImage* 200, 142–158.
- Ngo, G.H., Khosla, M., Jamison, K., Kuceyeski, A., Sabuncu, M.R., 2022. Predicting individual task contrasts from resting-state functional connectivity using a surface-based convolutional network. *NeuroImage* 248, 118849.
- Ngo, G.H., Nguyen, M., Chen, N.F., Sabuncu, M.R., 2021. Text2brain: Synthesis of brain activation maps from free-form text query. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 605–614.
- Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., Miller, A., 2019. Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473.
- Pinho, A.L., Amadon, A., Gauthier, B., Clairis, N., Knops, A., Genon, S., Dohmatob, E., Torre, J.J., Ginisty, C., Becuwe-Desmidt, S., et al., 2020. Individual brain charting dataset extension, second release of high-resolution fMRI data for cognitive mapping. *Sci. Data* 7 (1), 1–16.
- Poldrack, R.A., Kittur, A., Kalar, D., Miller, E., Seppa, C., Gil, Y., Parker, D.S., Sabb, F.W., Bilder, R.M., 2011. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. *Front. Neuroinform.* 5, 17.
- Poldrack, R.A., Yarkoni, T., 2016. From brain maps to cognitive ontologies: informatics and the search for mental structure. *Annu. Rev. Psychol.* 67, 587–612.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1–67.
- Reid, A.T., Bzdok, D., Genon, S., Langner, R., Müller, V.I., Eickhoff, C.R., Hoffstaedter, F., Cieslik, E.-C., Fox, P.T., Laird, A.R., et al., 2016. Anima: a data-sharing initiative for neuroimaging meta-analyses. *NeuroImage* 124, 1245–1253.
- Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A.R., Schulz, J.B., Fox, P.T., Eickhoff, S.B., 2012. Modelling neural correlates of working memory: a coordinate-based meta-analysis. *NeuroImage* 60 (1), 830–846.
- Rubin, T.N., Koyejo, O., Gorgolewski, K.J., Jones, M.N., Poldrack, R.A., Yarkoni, T., 2017. Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. *PLoS Comput. Biol.* 13 (10), e1005649.
- Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* 24 (5), 513–523.
- Sennrich, R., Haddow, B., Birch, A., 2016. Neural machine translation of rare words with subword units. *Proceedings of ACL*. <https://doi.org/10.18653/v1/P16-1162>.
- Sevinc, G., Spreng, R.N., 2014. Contextual and perceptual brain processes underlying moral cognition: a quantitative meta-analysis of moral reasoning and moral emotions. *PLoS One* 9 (2), e87427.
- Shackman, A.J., Salomons, T.V., Slagter, H.A., Fox, A.S., Winter, J.J., Davidson, R.J., 2011. The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nat. Rev. Neurosci.* 12 (3), 154–167.
- Smallwood, J., 2013. Distinguishing how from why the mind wanders: a process-occurrence framework for self-generated mental activity. *Psychol. Bull.* 139 (3), 519.
- Spreng, R.N., Mar, R.A., Kim, A.S., 2009. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *J. Cogn. Neurosci.* 21 (3), 489–510.
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al., 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.
- Turkeltaub, P.E., Eden, G.F., Jones, K.M., Zeffiro, T.A., 2002. Meta-analysis of the functional neuroanatomy of single-word reading: method and validation. *NeuroImage* 16 (3), 765–780.
- Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E., Yacoub, E., Ugurbil, K., Consortium, W.-M.H., et al., 2013. The WU-minn human connectome project: an overview. *NeuroImage* 80, 62–79.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.
- Wager, T.D., Lindquist, M.A., Nichols, T.E., Kober, H., Van Snellenberg, J.X., 2009. Evaluating the consistency and specificity of neuroimaging data using meta-analysis. *NeuroImage* 45 (1), S210–S221.
- Wu, J., Ngo, G.H., Greve, D., Li, J., He, T., Fischl, B., Eickhoff, S.B., Yeo, B.T., 2018. Accurate nonlinear mapping between MNI volumetric and freesurfer surface coordinate systems. *Hum. Brain Mapp.* 39 (9), 3793–3808.
- Xu, A., Larsen, B., Baller, E.B., Scott, J.C., Sharma, V., Adebimpe, A., Basbaum, A.I., Dworkin, R.H., Edwards, R.R., Woolf, C.J., et al., 2020. Convergent neural representations of experimentally-induced acute pain in healthy volunteers: a large-scale fMRI meta-analysis. *Neurosci. Biobehav. Rev.* 112, 300–323.
- Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* 8 (8), 665–670.
- Yeo, B.T., Krienen, F.M., Eickhoff, S.B., Yaakub, S.N., Fox, P.T., Buckner, R.L., Asplund, C.L., Chee, M.W., 2015. Functional specialization and flexibility in human association cortex. *Cereb. Cortex* 25 (10), 3654–3672.