A Comprehensive Neural and Behavioral Task Taxonomy Method for Transfer Learning in NLP

Yunhao Zhang[1,2†], **Chong Li**[1,2†], **Xiaohan Zhang**[1,2], **Xinyi Dong**[3], **Shaonan Wang**[1,2]

[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University

{zhangyunhao2021, lichong2021}@ia.ac.cn; 202121061097@mail.bnu.edu.cn;

{shaonan.wang, xiaohan.zhang}@nlpr.ia.ac.cn

Abstract

Transfer learning is frequently utilized in scenarios with limited labeled examples, where a crucial step is to identify a related task to the target task. CogTaskonomy (Luo et al., 2022) was proposed to acquire a taxonomy of NLP tasks, specifically focusing on assessing the similarities between tasks. This method, inspired by cognitive processes, exhibits notable time efficiency. Nevertheless, it does not fully exploit the task-related information present in cognitive data and lacks a comprehensive evaluation of various types of cognitive data. To address these limitations, this paper proposes a comprehensive neural and behavioral method to investigate the relationship among NLP tasks. Our approach utilizes cognitive data, encompassing both neural data such as fMRI and EEG, as well as behavioral data including eye-tracking and semantic feature ratings. Each data modality is employed to establish a common representation space with Representation Similarity Analysis for projecting task-related representations. To fully leverage the cognitive information, we effectively extract the task-relevant information extracted from neural data through feature ranking. Experimental results on 12 NLP tasks demonstrate that our proposed method outperforms state-of-the-art methods on evaluating task similarity.

1 Introduction

Pre-trained Language Models (PLMs) achieve remarkable performance on downstream tasks through fine-tuning on abundant labeled data (Devlin et al., 2019; Radford et al., 2019; Peters et al., 2018). However, their performance tends to degrade when facing with limited labeled data (Chen et al., 2023; Hedderich et al., 2021). To overcome this challenge, researchers employ transfer learning by initially fine-tuning a PLM on a related task with ample labeled data, followed by fine-tuning on the target task (Dwivedi and Roig, 2019; Song et al., 2019). Nevertheless, devising an effective method to identify a suitable similar task remains a challenging endeavor (Ramirez et al., 2023).

To evaluate the relatedness between tasks, different methods have been employed. The first category of method is task embedding, which learns a dedicated high-dimensional representation for each task using a task encoder (James et al., 2018; Lan et al., 2019; Achille et al., 2019; Vu et al., 2020). Despite the low time complexity, modulating the model to adapt for a new task using this method is challenging. Another approach is Taskonomy method (Zamir et al., 2018), which fine-tunes a model on each task, and transfers each fine-tuned model to other tasks in a fully supervised manner. This process can effectively capture task similarity, while it is demanding and time-consuming. The third category encompasses the cognitively inspired CogTaskonomy proposed by Luo et al. (2022). This method projects task-related representations into a shared space based on fMRI and subsequently evaluates the similarity between tasks. It only needs to fine-tune a model for each task separately, which exhibits notable time efficiency. However, shared spaces in CogTaskonomy are not strict shared spaces, and this method does not effectively leverage task-related information from neural data. Our study falls under the third category and improves upon the existing approach by fully utilizing cognitive information to generate an enhanced shared space.

This paper proposes a method, called NBT (A Comprehensive **N**eural and **B**ehavioral Task **T**axonomy Method), integrating both neural data (i.e., fMRI, EEG) and behavioral data (i.e., Eye-tracking, Semantic feature ratings) to investigate the relationship among NLP tasks. We employ each data modality to establish a common representation space with Representation Similarity Analysis (RSA) (Kriegeskorte et al., 2008) for projecting

[†]Equal Contribution.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023, pages 233–241

task-related representations. Moreover, to fully exploit cognitive information, we employ feature ranking to effectively extract the task-relevant information extracted from neural data. Results on 12 NLP tasks show that NBT outperforms the previous cognitive-inspired method, and achieves comparable performance to state-of-the-art methods with lower computational time complexity on evaluating task similarity.

2 Methods

To evaluate task similarity, we propose NBT as demonstrated in Figure 1, which involves three steps: 1) calculating task-specific representations, 2) generating a shared space from cognitive data, and 3) mapping task-specific representations to cognitive data in a shared space to calculate the task similarity. Each step in Figure 1 is described as follows.

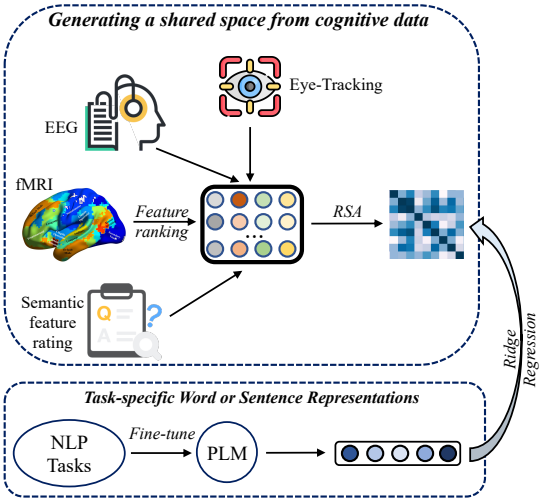

Figure 1: Architecture of NBT.

2.1 Extracting Task-Specific Representations

We take task-specific representations through the last layer of PLMs fine-tuned on various NLP tasks. For word-level stimuli, following the method in previous studies (Wang et al., 2022, 2023), we randomly sample a maximum of 1,000 sentences for each target word from Xinhua News corpus[1], and feed sentences into fine-tuned PLMs, with vectors extracted from the last layer. Subsequently, we average vectors of each target word to obtain a task-related representation. For sentence-level stimuli, the sentence representation is obtained by extracting the hidden state of [CLS] token from each sequence in the last layer.

[1]http://www.xinhuanet.com/whxw.htm

2.2 Generating a Shared Space from Cognitive Data

To fully extract task-relevant information from high-dimensional fMRI, we select a certain number of voxels by scoring them with their relevance to PLMs. Specifically, regression models are trained for each voxel to predict each dimension of representations of PLMs with this voxel and its adjacent three-dimensional neighbors. The correlation between the true and the predicted representations is regarded as the informative score of each voxel.

Subsequently, we utilize RSA, a widely employed technique for discerning correlations between neuronal responses derived from brain data and models. This facilitates the mapping of cognitive data originating from distinct subject-specific spaces into a unified representational space. It can also help to mitigate the inherent noise in cognitive data. To be more precise, we extract a group of word or sentence cognitive representations $E = \{e_1, e_2, ..., e_n\}$ from cognitive data. For each pair of representations (e_i, e_j), its similarity is measured by the Pearson correlation (ρ). Thus, we obtain cognitive data in the representational similarity space $M \in \mathbb{R}^{n \times n}$, where $M_{ij} = \rho(e_i, e_j)$.

2.3 Mapping Task-specific Representations to the Shared Space

Finally, we use ridge regression[2] to learn a mapping function between cognitive data M and task-specific representations $P_u \in \mathbb{R}^{n \times w}$ obtained by the PLM $f(\theta_u)$ fine-tuned on the u-th task, where n is the number of words or sentences as well as the dimensionality of M, and w is the dimensionality of representations of $f(\theta_u)$. The regression coefficients l, which is a n-dimension vector, and l_0 are learned by minimizing

$$\text{loss}(l, l_0) = \|P_u l + l_0 - m\|_2^2 + \lambda \|l\|_2^2 \quad (1)$$

for each column $m \in \mathbb{R}^{n \times 1}$ which is a single dimension of the M matrix. The regularization parameter λ for each dimension is set by the nested cross-validation. Each dimension of M and P_u is standardized across training stimuli.

After mapping, we obtain the final predict matrix $\hat{M}_u$ by averaging predict matrices over all participants for the u-th task, and calculate the correlation coefficients between $\hat{M}_u$ and M across each dimension to obtain task-specific representation CogP_u in the shared space, defined as follows:

[2]Compared to Multi-layer Perception, ridge regression has fewer parameters and lower time complexity.

$$\begin{aligned}\mathrm{CogP}_u = &[c\left(\hat{\boldsymbol{m}}_{u0}, \boldsymbol{m}_0\right), \ldots, c\left(\hat{\boldsymbol{m}}_{uh}, \boldsymbol{m}_h\right) \\ &\ldots, c\left(\hat{\boldsymbol{m}}_{uv}, \boldsymbol{m}_v\right)], 0 \leq h \leq v\end{aligned} \tag{2}$$

where $\hat{m}_{hu}$ and m_h are respectively the predicted and ground-truth vectors of the h-th dimension, n is the number of dimensions, and $c(\cdot)$ is the correlation function for vector pairs, e.g., ρ and the coefficient of determination (R^2).

We then utilize $s(\cdot)$ involving three different similarity functions (Cosine similarity (cos), R^2 and ρ) to calculate pairwise task similarity as follows:

$$\mathrm{Sim}_{uu'} = s\left(\mathrm{Cog\,P}_u, \mathrm{Cog\,P}_{u'}\right) \tag{3}$$

3 Experimental settings

3.1 Neural datasets

fMRI We utilize the fMRI dataset from Pereira et al. (2018), collecting functional activation data of 627 natural language sentences from 5 participants.
EEG We adopt the EEG dataset comprising 1100 sentences and 4384 words obtained from 12 participants, as published by Hollenstein et al. (2018).

3.2 Behavioral datasets

Semantic feature ratings We use the semantic dataset published by Binder et al. (2016), which includes 535 concepts with 65 semantic features. Each semantic feature of a word has a rating that is the average of annotations from 30 participants.
Eye-Tracking The Eye-Tracking dataset in our experiments comes from Hollenstein et al. (2018), which is collected with EEG simultaneously.

3.3 Transfer Learning Tasks

Twelve NLP tasks are involved in our experiments, covering sentence/token-level classification, information extraction, and passage ranking tasks (Tjong Kim Sang and De Meulder, 2003; Hendrickx et al., 2010; Nguyen et al., 2016; Rajpurkar et al., 2018; Wang et al., 2019). Task details are shown in Table 1.

3.4 Baseline Methods

Direct Similarity Estimation (DSE) DSE approximates the similarity of task pairs using the average similarity of sentence representations from PLMs fine-tuned on the corresponding task.

Analytic Hierarchy Process (AHP) On the other hand, the similarity of task pairs can be estimated from the pair-wise transfer learning results (Zamir et al., 2018). Given a target task, PLMs transferred from different source tasks are compared on a hold-out dataset to determine the transferability of the target task, which is further used to approximate the similarity between tasks.

Task	Dataset	#Train
Acceptability	CoLA	8,551
Natural Language Inference	MNLI	392,702
Paraphrase	QQP	363,846
Paraphrase	MRPC	3,668
Question Answering	QNLI	104,743
Sentiment Analysis	SST-2	67,349
Entailment	RTE	2,490
Textual Similarity	STS-B	5,749
Extractive Question Answering	SQuAD-2.0	129,941
Relation Extraction	Semeval-2010	8,000
Named Entity Recognition	CoNLL-2003	14,042
Passage Reranking	MS MARCO	3,213,835

Table 1: Statistic of tasks used.

Cognitive Representation Analytics (CRA) CRA first calculates the Representation Dissimilarity Matrix (RDM) by the dissimilarity of sentence representations, then approximates the similarity between tasks by the similarity between the corresponding RDMs (Luo et al., 2022).

Cognitive-Neural Mapping (CNM) CNM calculates the task similarity by mapping sentence representations from multiple fine-tuned PLMs to the same fMRI data (Luo et al., 2022).

3.5 Hyperparameters

Most of the hyperparameters used in our transfer-learning and baseline experiments are in line with the ones in Luo et al. (2022). The only exception is that the source model of TinyBERT is distilled from our fine-tuned BERT, rather than initialized from the public models of Jiao et al. (2020) whose download links are missing now.

The hyperparameter λ in ridge regression for each dimension is set utilizing nested cross-validation within the training set, respectively. Each voxel is normalized across training stimuli, as is the dimension of representations of PLMs. More formally, the nested cross-validation framework is applied to make sure that the data utilized for the regularization parameter tuning and the data employed to test the model is firmly independent. The interior 10-fold cross-validation is utilized to choose the optimal regularization parameter, and the extrinsic 10-fold nested cross-validation is applied to predict the values using the model with the optimal regularization parameter.

3.6 Evaluation Metric

To empirically evaluate the similarity between NLP tasks, we conducted pair-wise transfer learning experiments for each task in Table 1, and quantify it with the actual transfer learning performance (Results are reported in Appendix B). In other words, the more similar the source task is, the better performance on the target task the model gets.

Task Ranking Score (TRS) is used to assess the distance between the task similarity estimated from different methods mentioned before, e.g., DSE, and the empirical task similarity in the transfer learning experiments. Specifically, the task ranking score is obtained from the average ranking of the best source task estimated by the method in the real transfer learning experiment for all target tasks. Then, the random guess leads to $\frac{N}{2}$ in the task ranking score, while the best one always gets 1 under the N tasks setting.

4 Results and Analysis

Method	Complexity	TRS ↓	
		TinyBERT	BERT
Baselines			
Random	$O(1)$	6.0	6.0
DSE	$O(n)$	4.9	4.8
CRA	$O(n)$	5.0	4.4
CNM	$O(n)$	4.1	4.6
AHP*	$O(n^2)$	**1.4**	**2.5**
Ours			
NBT$_{eeg}$	$O(n)$	4.0	4.2
NBT$_{eye}$	$O(n)$	3.6	3.8
NBT$_{sem}$	$O(n)$	3.3	3.9
NBT$_{fmri}$	$O(n)$	**2.4**	**2.9**

Table 2: Task ranking scores (TRS) for different task similarity estimation methods. Results denoted by * come from Luo et al. (2022).

Main result Table 2 displays the task ranking scores obtained from various methods. It is evident that NBT outperforms CogTaskonmoy (CNM and CRA), with NBT$_{fmri}$ exhibiting an average improvement of 41.13% over them. Additionally, NBT$_{eeg}$, NBT$_{eye}$ and NBT$_{sem}$ separately exceed them by 8.93%, 24.40%, and 20.02%. These results suggest that NBT is superior to CogTaskonomy in capturing the relation among NLP tasks by utilizing distinct neural and behavioral data. Although NBT has the same time complexity with CNM, NBT is more efficient than CNM in practice[3]. Moreover, NBT has comparable performance to AHP with lower computational time and less memory.

Corr. Coef.	Task Sim.	Method	TRS ↓	
			TinyBERT	BERT
cos	R^2	CNM	4.3	4.8
		NBT$_{fmri}$	2.7	3.3
	ρ	CNM	4.1	4.6
		NBT$_{fmri}$	2.9	3.0
	cos	CNM	4.2	4.4
		NBT$_{fmri}$	2.9	3.2
ρ	R^2	CNM	4.4	4.4
		NBT$_{fmri}$	2.6	**2.7**
	ρ	CNM	4.2	4.6
		NBT$_{fmri}$	**2.4**	2.9
	cos	CNM	4.2	4.8
		NBT$_{fmri}$	3.0	3.0

Table 3: TRS of NBT$_{fmri}$ and CNM for BERT and TinyBERT with different measures of task similarity and correlation coefficients.

Evaluating NBT with different similarity measure combinations There are multiple options to calculate correlation coefficients of mapping performance (e.g., ρ, R^2) and task similarity (e.g., cos, R^2, ρ). To demonstrate the robustness of the proposed method, we evaluate the performance of NBT across various measure combinations. As the CNM utilizes fMRI to measure task relations, we compare it with NBT using identical cognitive data. It can be seen from Table 3 that NBT$_{fmri}$ has better performance than CNM in all cases. Specifically, NBT$_{fmri}$ outperforms CNM by 35.39% on average. In addition, TinyBERT and BERT show minimal performance difference using NBT$_{fmri}$, while TinyBERT is more resilient to different cases compared to BERT with CNM, which indicates that the proposed method has greater generalization for different models.

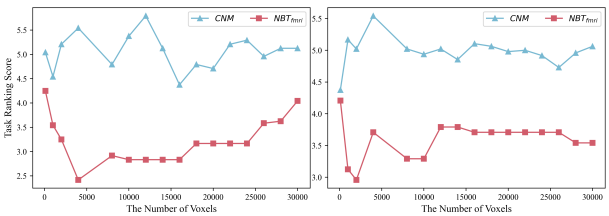

Figure 2: TRS of NBT$_{fmri}$ and CNM with TinyBERT (left) and BERT (right) predicting different numbers of voxels.

[3]To map a fine-tuned model to fMRI data, NBT only costs 23.3s (for BERT), which is 2.7% of time spent with CNM (879.3s) when the voxel number is 30K

Evaluating NBT with numbers of voxels The number of voxels is a critical parameter in the proposed method. In this part, we further evaluate the stability of our proposed method by comparing it with CNM across voxel numbers. Figure 2 shows that NBT$_{fmri}$ exhibits superior performance to CNM across various numbers of voxels. Moreover, it's hard to predict the number of voxels when achieving the best ranking score using CNM, while NBT$_{fmri}$ can achieve relatively optimal results in the 2000-10000 voxel range.

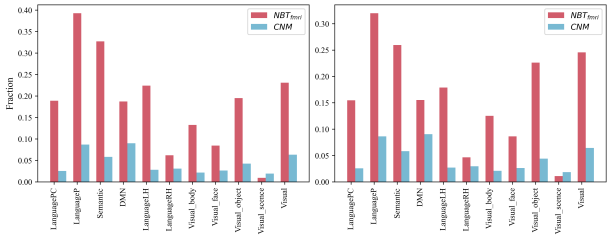

Figure 3: Distribution of informative voxels of NBT$_{fmri}$ and CNM with TinyBERT (left) and BERT (right) across the brain (averaged over 5 participants).

Analysis on the spatial distribution of informative voxels In this section, we explore the distribution of voxels selected by NBT$_{fmri}$ and CNM across brain networks associated with semantic processing. As can be obtained from Figure 3, NBT$_{fmri}$ selects a obviously larger proportion of voxels within areas relating to semantic processing compared to CNM, including languagePC, languageP, semantic and languageLH, which suggests the proposed method can effectively extract the task-relevant information from neural signal. Moreover, voxels selected using NBT$_{fmri}$ also distribute in areas relating to visual semantics, indicating that the method can effectively utilize semantic information from each brain functional area to accurately separate task presentations from neural data.

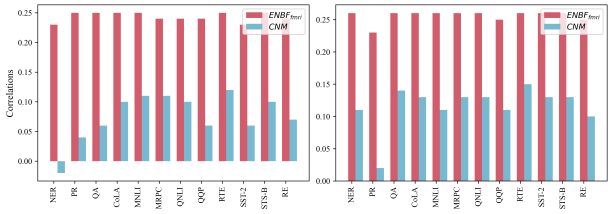

Figure 4: Voxel prediction results of NBT$_{fmri}$ and CNM with TinyBERT (left) and BERT (right) (averaged over 5 participants and 30K voxels).

Analysis on the voxel prediction evaluation In this part, we compare results of the proposed method and CNM on voxel prediction. It can be noticed from Figure 4 that NBT$_{fmri}$ obtains higher correlation than CNM, suggesting that the proposed method can better establish the connection between neural signals and task-representations of fine-tuned PLMs, and can better isolate task presentations from neural signals.

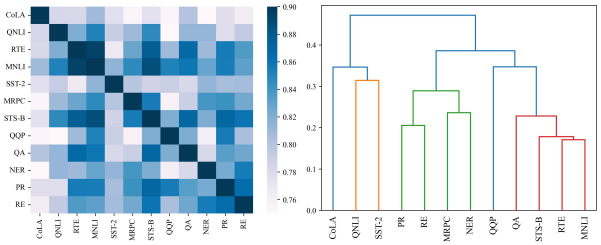

Figure 5: Task similarity matrix from the results of NBT$_{fmri}$ (left) and taxonomy tree (right).

Analysis on taxonomy tree of 12 NLP tasks In this section, we explore the task similarity matrix and the taxonomy tree of 12 NLP tasks from NBT$_{fmri}$. Compared with the taxonomy tree from CogTaskonomy, "QA, STS-B" and "QNLI, SST-2" are both found to be put in one cluster in two taxonomies. Furthermore, the proposed method clusters three tasks that need to infer semantic relationships (MNLI, STS-B and RTE) in one cluster, while CogTaskonomy divides STS-B into other clusters.

5 Conclusion

We propose a comprehensive neural and behavioral method to investigate the similarity between various NLP tasks. This method can fully extract task-relevant information from neural data thus capture task taxonomy and effectively guide transfer learning across diverse NLP tasks, which also can be beneficial for other cross-task learning paradigms, including multi-task learning (Zhang et al., 2021; Chen et al., 2021), meta learning (Yin, 2020) and lifelong learning (Biesialska et al., 2020). Results on 12 tasks show that the proposed method outperforms the previous cognitive-inspired method, and reaches comparable performance to the state-of-the-art method with $O(n)$ computational time complexity on evaluating task similarity.

Limitations

The proposed method utilizes cognitive data, including neural data such as fMRI and EEG, along with behavioral data such as eye-tracking and semantic feature ratings, to efficiently capture the inter-task relationships in NLP with reduced computational time complexity. Although existing

work has proven that human brain contains information about NLP tasks (Oota et al., 2022), it is unclear what kind of task-relevant information human brain contains (i.e., sentence/token-level classification, information extraction, and passage ranking tasks). Moreover, pre-trained models are fine-tuned on downstream tasks in a fully supervised manner, which is different from how human learn and understand new knowledge (Kühl et al., 2022). Therefore, the task similarity based on cognitive data may show different pattern from the real task similarity. Finally, since the proposed method currently exhibits effective performance on NLP, it will be extended to vision and multi-modal domains in our future work.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. 2019. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6430–6439.

Magdalena Biesialska, Katarzyna Biesialska, and Marta R Costa-Jussa. 2020. Continual lifelong learning in natural language processing: A survey. *arXiv preprint arXiv:2012.09823*.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130–174.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Shijie Chen, Yu Zhang, and Qiang Yang. 2021. Multi-task learning in natural language processing: An overview. *arXiv preprint arXiv:2109.09138*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kshitij Dwivedi and Gemma Roig. 2019. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12387–12396.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.

Stephen James, Michael Bloesch, and Andrew J Davison. 2018. Task-embedded control networks for few-shot imitation learning. In *Conference on robot learning*, pages 783–795. PMLR.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Niklas Kühl, Marc Goutier, Lucas Baier, Clemens Wolff, and Dominik Martin. 2022. Human vs. supervised machine learning: Who learns patterns faster? *Cognitive Systems Research*, 76:78–92.

Lin Lan, Zhenguo Li, Xiaohong Guan, and Pinghui Wang. 2019. Meta reinforcement learning with task embedding and shared policy. *arXiv preprint arXiv:1905.06527*.

Yifei Luo, Minghui Xu, and Deyi Xiong. 2022. CogTaskonomy: Cognitively inspired task taxonomy is beneficial to transfer learning in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 904–920, Dublin, Ireland. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *the Workshop on Cognitive Computation: Integrating neural and symbolic*

approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016, volume 1773 of CEUR Workshop Proceedings. CEUR-WS.org.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. 2022. Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3220–3237, Seattle, United States. 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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pierluigi Zama Ramirez, Adriano Cardace, Luca De Luigi, Alessio Tonioni, Samuele Salti, and Luigi Di Stefano. 2023. Learning good features to transfer across tasks and domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jie Song, Yixin Chen, Xinchao Wang, Chengchao Shen, and Mingli Song. 2019. Deep model transferability from attribution maps. *Advances in Neural Information Processing Systems*, 32.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across nlp tasks. *arXiv preprint arXiv:2005.00770*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Shaonan Wang, Yunhao Zhang, Weiting Shi, Guangyao Zhang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2023. A large dataset of semantic ratings and its computational extension. *Scientific Data*, 10(1):106.

Shaonan Wang, Yunhao Zhang, Xiaohan Zhang, Jingyuan Sun, Nan Lin, Jiajun Zhang, and Chengqing Zong. 2022. An fmri dataset for concept representation with semantic feature annotations. *Scientific Data*, 9(1):721.

Wenpeng Yin. 2020. Meta-learning for few-shot natural language processing: A survey. *arXiv preprint arXiv:2007.09604*.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722.

Yunhao Zhang, Jiajun Yu, Xinyi Dong, and Ping Zhong. 2021. Multi-task support vector machine with pinball loss. *Engineering Applications of Artificial Intelligence*, 106:104458.

A Correlation selection in NBT$_{fmri}$

There are three options for us to calculate the correlation coefficient (R^2, ρ, cos) on all voxels between predicted values and ground-truth values. We calculate task ranking scores using these options for NBT$_{fmri}$ with TinyBERT and BERT, as shown in Figure 6. Our findings suggest that ρ outperforms R^2 in almost all cases.

B Oracle Task Ranking

After pair-wise transfer learning, we evaluate the performance of models on the validation set of target tasks and report them in Table 4 and Table 5 for BERT and TinyBERT, respectively.

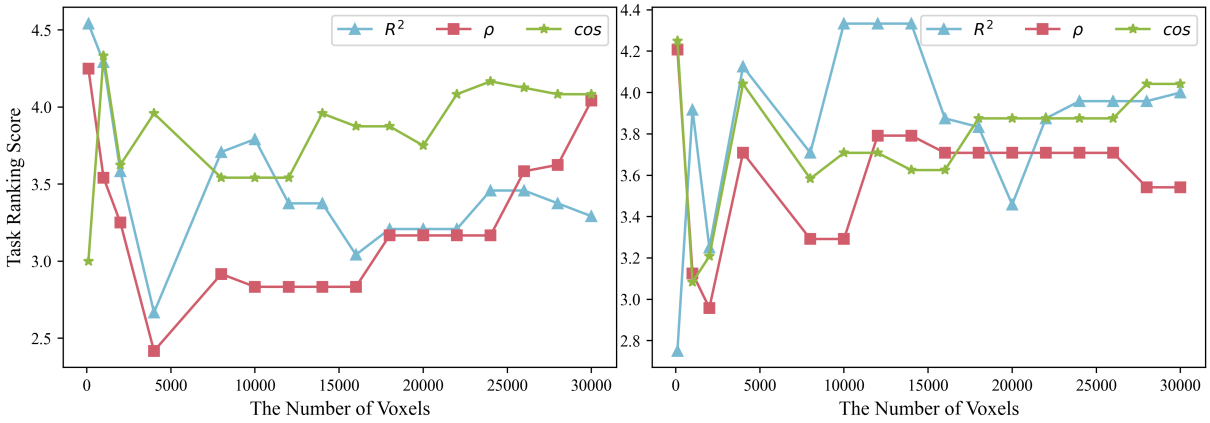

Figure 6: Task ranking scores with different correlation coefficients in NBT$_{fmri}$ based on TinyBERT (left) and BERT (right).

Source Task	Target Task											
	CoLA Mcc	**QNLI** Acc	**RTE** Acc	**MNLI** Acc	**SST-2** Acc	**MRPC** F1	**STS-B** r^s	**QQP** F1	**NER** F1	**RE** F1	**QA** F1	**PR** MRR@10
CoLA	-	91.40(4)	64.98(6)	83.95(8)	92.55(5)	90.82(1)	87.94(7)	88.04(1)	93.70(3)	90.73(4)	76.02(6)	65.84(2)
QNLI	55.57(5)	-	70.40(2)	84.37(5)	92.78(3)	90.16(3)	88.71(2)	87.25(10)	93.53(6)	89.98(8)	77.20(1)	63.21(9)
RTE	55.47(6)	91.73(1)	-	84.03(7)	92.43(6)	88.04(8)	88.65(4)	87.66(7)	93.48(7)	90.94(1)	76.39(3)	63.47(7)
MNLI	54.84(7)	91.12(6)	77.26(1)	-	93.12(1)	89.64(6)	88.65(4)	87.74(4)	93.44(8)	89.20(10)	75.88(7)	64.47(4)
SST-2	56.76(2)	91.07(7)	61.73(10)	84.39(4)	-	87.78(9)	87.90(8)	87.64(8)	93.60(4)	90.32(5)	76.24(5)	62.75(10)
MRPC	56.77(1)	91.58(2)	65.34(4)	84.51(3)	92.78(3)	-	88.18(6)	87.96(2)	93.79(2)	90.84(2)	75.20(9)	66.66(1)
STS-B	56.25(4)	90.39(10)	65.34(4)	83.74(10)	92.89(2)	90.24(2)	-	87.59(9)	93.60(4)	90.83(3)	75.04(10)	61.20(11)
QQP	54.69(8)	90.55(9)	63.90(7)	84.66(1)	92.43(6)	89.01(7)	88.70(3)	-	93.22(10)	89.59(9)	76.93(2)	64.44(5)
NER	56.50(3)	91.36(5)	63.54(8)	84.32(6)	92.43(6)	89.93(4)	87.73(9)	87.72(6)	-	90.02(7)	74.04(11)	63.81(6)
RE	53.93(9)	91.07(7)	62.09(9)	83.92(9)	91.97(9)	87.46(10)	87.17(11)	87.82(3)	93.85(1)	-	76.35(4)	63.31(8)
QA	52.59(10)	91.47(3)	67.15(3)	84.65(2)	91.86(10)	89.72(5)	89.18(1)	87.74(4)	93.41(9)	90.11(6)	-	65.44(3)
PR	5.20(11)	89.44(11)	61.01(11)	82.49(11)	91.40(11)	82.98(11)	87.44(10)	87.13(11)	90.53(11)	86.75(11)	75.68(8)	-

Table 4: Transfer learning results of BERT. Mcc denotes the Matthews correlation coefficient, r^s is the Spearman's rank correlation coefficient, and MRR@10 denotes the Mean Reciprocal Rank for the top 10. The ranking for the source task to the target task is denoted in the right parenthesis.

Source Task	Target Task											
	CoLA	**QNLI**	**RTE**	**MNLI**	**SST-2**	**MRPC**	**STS-B**	**QQP**	**NER**	**RE**	**QA**	**PR**
	Mcc	Acc	Acc	Acc	Acc	F1	r^s	F1	F1	F1	F1	MRR@10
CoLA	-	78.89(9)	55.23(9)	74.65(9)	86.12(9)	76.4(11)	25.15(9)	82.73(9)	74.98(10)	85.46(8)	47.31(9)	58.57(9)
QNLI	16.57(8)	-	61.01(4)	77.74(8)	85.67(10)	87.80(4)	84.44(7)	84.94(8)	85.71(8)	86.04(6)	68.44(3)	62.70(8)
RTE	16.11(9)	87.55(4)	-	79.30(5)	87.84(5)	87.80(4)	85.96(3)	85.40(6)	87.63(3)	87.13(2)	64.25(7)	63.94(6)
MNLI	28.19(1)	89.27(2)	74.01(1)	-	91.28(1)	90.36(1)	87.51(2)	86.67(1)	87.18(5)	88.19(1)	70.24(1)	63.99(5)
SST-2	9.91(10)	75.93(10)	55.23(9)	72.57(10)	-	80.45(8)	14.99(11)	81.78(10)	70.07(11)	78.29(10)	47.03(10)	55.55(10)
MRPC	19.69(7)	87.46(5)	62.82(3)	79.67(4)	88.42(4)	-	85.57(4)	85.80(3)	87.20(4)	85.18(9)	63.86(8)	63.31(7)
STS-B	25.86(2)	85.54(8)	59.57(6)	78.93(7)	86.93(7)	89.45(2)	-	85.65(4)	87.06(6)	86.78(3)	64.64(6)	64.66(2)
QQP	22.29(3)	86.82(6)	59.57(6)	79.92(3)	87.16(6)	87.83(3)	85.42(6)	-	84.81(9)	86.08(5)	67.01(5)	64.13(4)
NER	5.66(11)	63.04(11)	58.12(8)	68.93(11)	81.65(11)	79.60(10)	19.01(10)	79.82(11)	-	76.14(11)	42.93(11)	55.34(11)
RE	21.79(4)	86.45(7)	55.23(9)	82.95(1)	89.56(3)	80.14(9)	83.19(8)	85.49(5)	88.86(1)	-	68.14(4)	65.30(1)
QA	19.81(6)	88.72(3)	64.26(2)	79.01(6)	86.93(7)	86.72(7)	85.53(5)	85.32(7)	86.41(7)	85.79(7)	-	64.60(3)
PR	20.31(5)	89.77(1)	61.01(4)	82.72(2)	90.83(2)	87.69(6)	87.90(1)	86.29(2)	87.89(2)	86.70(4)	70.19(2)	-

Table 5: Transfer learning results of TinyBERT. Mcc denotes the Matthews correlation coefficient, r^s is the Spearman's rank correlation coefficient, and MRR@10 denotes the Mean Reciprocal Rank for the top 10. The ranking for the source task to the target task is denoted in the right parenthesis.