Cognitive Information Bottleneck: Extracting Minimal Sufficient Cognitive Language Processing Signals

Yuto Harada, Yohei Oseki
The University of Tokyo
harada-yuto, oseki@g.ecc.u-tokyo.ac.jp

Abstract

In Reinforcement Learning from Human Feedback (RLHF), explicit human feedback, such as rankings, is employed to align Natural Language Processing (NLP) models with human preferences. In contrast, the potential of implicit human feedback, encompassing cognitive processing signals like eye-tracking and brain activity, remains underexplored. These signals capture unconscious human responses but are often marred by noise and redundancy, complicating their application to specific tasks. To address this issue, we introduce the Cognitive Information Bottleneck (CIB), a method that extracts only the task-relevant information from cognitive processing signals. Grounded in the principles of the information bottleneck, CIB aims to learn representations that maximize the mutual information between the representations and targets while minimizing the mutual information between inputs and representations. By employing CIB to filter out redundant information from cognitive processing signals, our goal is to provide representations that are both minimal and sufficient. This approach enables more efficient fitting of models to inputs. Our results show that the proposed method outperforms existing methods in efficiently compressing various cognitive processing signals and significantly enhances performance on downstream tasks. Evaluated on public datasets, our model surpasses contemporary state-of-the-art models. Furthermore, by analyzing these compressed representations, we offer insights into how cognitive processing signals can be leveraged to improve performance.

Keywords: Information Bottleneck, Cognitive Language Processing Signals

1. Introduction

The success of explicit human feedback in the form of ranking within Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), underscores the importance of aligning human preferences with the Natural Language Processing (NLP) model. While explicitly ranking preferences is an effective method, humans read text without being aware of each stage of the cognitive process. Unconscious cognitive processing signals, such as eye tracking and brain activity, may reflect more nuanced human responses. Toward leveraging cognitive processing signals as implicit human feedback, NLP research inspired by cognitive neuroscience provides the foundation. Research integrating cognitive processing signals into neural NLP models has shown improvements in model performance across various tasks, POS analysis (Murphy et al., 2022), dependency parsing (Strzyz et al., 2019), sentiment analysis (Barrett et al., 2018), Named Entity Recognition (Hollenstein and Zhang, 2019; Ren and Xiong, 2021), and relation extraction (Hollenstein et al., 2019). These findings indicate that the recording of human cognitive processes can be beneficial regardless of the task. However, previous works have encountered issues in utilizing cognitive process signals, which have hindered their integration into NLP models. We aim to solve the following three problems in the use of cognitive language processing signals.

Non-task-specificity Cognitive signals contain information about various types of processing for stimuli (Kutas and Federmeier, 2000). Not all of the signals may be useful for the target task, and redundant information may be noise to the target task. Distinct feature extractions may need to be performed for different target tasks, but previous work has primarily relied on heuristic-based feature engineering.

High dimensionality In particular, brain activity recordings are very high-dimensional and require appropriate dimensional reduction methods. Previous work averages features to a few dimensions for concatenation with word embeddings, but this may result in loss of valuable information.

Limited availability Due to the constraints of taking data on humans with expensive equipment, it is difficult to prepare data as large as required by general machine learning. Eye tracking, in particular, is expected to improve in the future, as it has become inexpensive to acquire in recent years, but current cognitive features are still considered low-resource.

Previous work has avoided these problems by averaging EEG data to reduce noise and dimension, but averaging methods lose a lot of information. Other fields have proposed feature extraction methods with deep neural networks, and it is necessary to develop a model suitable for NLP. In this paper, we propose a Cognitive Information Bottleneck

LREC-COLING 2024, pages 3480–3489
20-25 May, 2024.

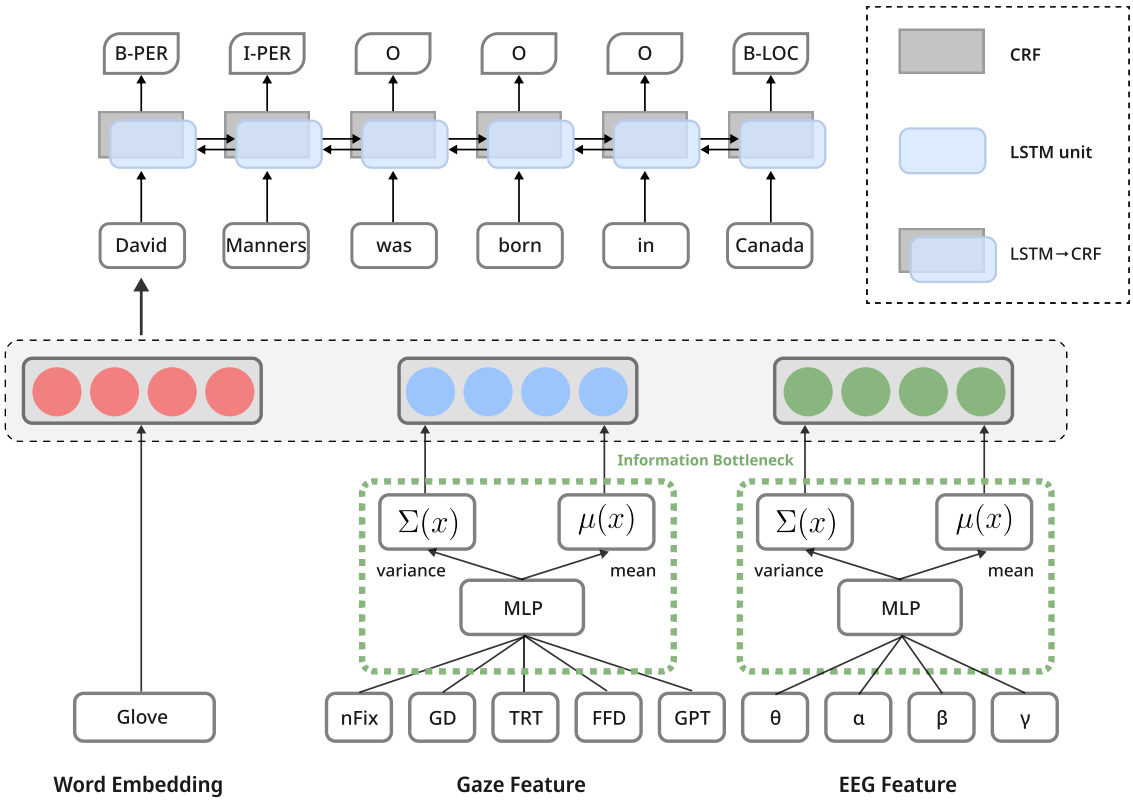

Figure 1: Overview of the Cognitive Information Bottleneck Method. The task example is NER, but can be applied to any task. Cognitive embeddings compressed by the information bottleneck method are concatted to word embeddings to make them inputs for the model.

(CIB) method to tackle these three issues: non-task-specificity, high dimension, and limited availability. Our CIB method is inspired by the information bottleneck principle (Tishby et al., 1999) and targets cognitive processing signals. The information bottleneck principle aims to minimize the mutual information between the input and its representation while extracting relevant information about the target from the input. By filtering out redundant information from the input, this approach generates the minimally sufficient representation that is optimal for the prediction task while ensuring task specificity. The information bottleneck method is a nonlinear feature extraction method that is robust to regularization in low-resource settings (Mahabadi et al., 2021) and has recently been employed in natural language processing. The CIB method can learn to complement pre-trained word embeddings with cognitive processing signals to generate stronger input representations. This method can be applied to any task. It integrates cognitive features more seamlessly into the NLP model to provide a robust interface between human responses and the neural network. In this paper, we evaluate our proposed method using a public dataset and compare it with existing feature extraction methods. The cognitive processing signals, compressed into an appropriate dimension by CIB, enhanced the model's performance in downstream tasks when used in conjunction with pretrained word embeddings. This indicates that proper feature extraction can potentially improve model performance.

In this work, we make the following main contributions:

- We propose the Cognitive Information Bottleneck (CIB), a novel method designed for efficient integration of cognitive processing signals into NLP models. CIB is very fast to train on a single GPU and improves pre-trained language models without fine-tuning.

- We evaluate the potential of CIB on public datasets, showing it compresses cognitive features more efficiently than existing methods, and improves the performance of downstream tasks, outperforming state-of-the-art models.

- We conducted a probing task for analysis. We showed that cognitive features have useful information for NLP and that their integration with pre-trained language models improves performance. Why and how cognitive features are useful is under-explored, and this analysis provides new insights.

ZuCo		
	Sentiment Reading Task	Normal Reading Task
number of sentences	400	300
domain	movie reviews	Wikipedia articles
task	rating the quality of the movie	natural reading

Table 1: Overview of the dataset used in this paper, in which subjects' cognitive language processing signals in two different domains were recorded for NER.

2. Related Works

2.1. Cognitive Processing Signals for NLP

In work integrating cognitive processing signals into the NLP model, two main types of cognitive process recordings have been used.

The first is eye tracking data. Eye movement data recording has been used in a variety of fields, including NLP and computer vision. Eye trackers can record very accurately, down to the millisecond, the point and time a human was looking while reading text. In recent years, eye tracking technology using mobile devices has improved significantly. It is expected that in the future it will be possible to obtain accurate data even in environments where special equipment is not available. Before its use in NLP, eye tracking was an important data for understanding human language processing. Humans do not pay attention to every word when reading, but take into account many linguistic factors (Demberg and Keller, 2008). This has been supported by observing the number of fixations and fixation times. For example, word length and frequency (Rayner, 1977) and predictability from context (Juhasz and Rayner, 2003) have been found to be related to fixation. Due to the relative ease of obtaining data, attempts to integrate cognitive processing signals into NLP models began with eye tracking. Early research used it for tasks such as part-of-speech tagging (Barrett et al., 2016) and prediction of multiword expressions (Rohanian et al., 2017). Later, neural models used eye tracking for sentiment analysis (Barrett et al., 2018) and NER (Hollenstein and Zhang, 2019; Ren and Xiong, 2021). Recently, ScanTextGAN (Khurana et al., 2023) has also been proposed to generate eye tracking in situations where gaze data is not available. Many works have discovered the advantages of using eye tracking data.

Second, EEG data. Although the use of EEG data in NLP tasks has not been explored as much as eye-tracking data, there are advantages to utilizing this source for NLP tasks. For example, Dambacher and Kliegl (2007) found that the amplitude of N400 increases with fixation time, indicating that EEG can complement eye-tracking data with each other. It has also been observed to contain other information about language processing, such as differences in processing verbs and nouns, concrete and abstract nouns, common and proper nouns (Weiss and Mueller, 2003), and decoding of POS information (Murphy et al., 2022). It is expected that EEG encodes a wealth of information that eye tracking does not. However, it is therefore difficult to integrate EEG data into NLP tasks. For the first time, Hollenstein et al. (2019) investigated the effectiveness of using EEG data in multiple tasks, including named entity recognition, relation extraction, and sentiment analysis. While Hollenstein et al. (2019) directly concatenated EEG features for word embedding, Ren and Xiong (2021) proposed CogAlign to account for the difference between the two modalities, textual and cognitive. However, in these previous works, they averaged the values obtained from 105 electrodes for normalization. Although the averaging method reduces the dimension of the EEG features, it has the problem of losing various information that the EEG has. In other fields, deep neural networks have been employed for feature extraction from EEG in machine learning (Zhang et al., 2021; Dai et al., 2020), and methods suitable for NLP need to be developed.

2.2. Information Bottleneck

The Information Bottleneck (IB) principle (Tishby et al., 2000) originated in information theory and was proposed for signal processing. IB aims to find a compressed representation of a signal while preserving the maximum information of the signal. Recently, the variational information bottleneck (VIB) method (Alemi et al., 2016) was proposed; VIB enables the application of IB principles to deep learning by approximating IB constraints. VIB has been employed in a variety of fields, including computer vision (Peng et al., 2018) and reinforcement learning (Goyal et al., 2019). In NLP, VIB can improve model performance by using compressed embeddings rather than original word embeddings (Li and Eisner, 2019). Finding a minimal yet sufficient representation for the target task is useful both for improving performance and for making the model more interpretable. This paper optimizes VIB for multiple types of cognitive features. The model solves the NLP task using not only word embeddings but also VIB-compressed representa-

tions. Upon receiving a cognitive feature, the VIB module learns to preserve information that the word embedding does not have in order to maximize task performance. Since our approach aims to integrate multiple cognitive modalities, our work is similar to the multimodal information bottleneck (MIB) (Mai et al., 2022) in multimodal learning. Although the motivation and architecture are different, our application of the IB method is convincing in the context of related research.

3. Cognitive Information Bottleneck

We define the cognitive feature corresponding to the i-th word as $x_i^{cog} \in X^{cog}$. These cognitive features represent signals derived from human cognitive processes, such as eye-tracking data and EEG data. Next, to compress these cognitive features, we introduce the Cognitive Information Bottleneck (CIB). The CIB generates a compressed representation of the cognitive features, $z_i^{cog} \in Z$. This compressed representation eliminates the redundancy of the cognitive features and retains only the information relevant to the task. We denote the predicted label for the target task of the i-th word as $y_i \in Y$. The objective of the CIB is to maximize the mutual information $I(X^{cog}, Z)$ between the cognitive features and their compressed representation, while minimizing the mutual information $I(Z, Y)$ between the compressed representation and the predicted labels. This is expressed by the following equation:

$$L_{IB} = \beta I(X^{cog}, Z) - I(Z, Y) \tag{1}$$

Here, β is a trade-off parameter that adjusts the balance between mutual information and redundancy. We apply this CIB within the framework of the Variational Information Bottleneck (VIB). The objective function of the VIB is as follows:

$$\begin{aligned} L_{CIB} = E_{q(z|x)}[-\log p(y|z)] \\ + \beta D_{KL}(q(z|x)||p(z)) \end{aligned} \tag{2}$$

Here, D_{KL} represents the Kullback-Leibler divergence, $q(z|x)$ is the probability distribution that generates the compressed representation z from the input x, and $p(y|z)$ is the probability distribution that generates the predicted label y from the compressed representation z. Finally, we concatenate the compressed cognitive feature z_i^{cog} and the word embedding $x_i^{word} \in X^{word}$. We denote this concatenated representation as x_{input}, which serves as the input for the downstream tasks. This provides the model with a representation that combines cognitive features and semantic information of words.

4. Experiments

4.1. Dataset

We evaluated the Cognitive Information Bottleneck (CIB) model on the Zurich Cognitive Language Processing Corpus (ZuCo)[1](Hollenstein et al., 2018). ZuCo is a unique dataset in which both eye-tracking and EEG data were recorded. The full corpus contains 1100 English sentences read by 12 adult native speakers, but it was 700 of these sentences that were provided with the NER task, which was used in the experiment.

Subjects performed several reading tasks, and the corpus recorded cognitive processing signals as they read different texts. The 400 sentences were a Sentiment Reading task, in which subjects read 400 movie review sentences extracted from the Stanford Sentiment Treebank (Socher et al., 2013) and were tasked with estimating movie scores. 300 sentences were a Normal reading task, in which subjects naturally read 300 sentences about great historical people contained in Wikipedia. Three types of entities were labeled as PERSON, ORGANIZATION, and LOCATION, and 1179 of the total 15237 tokens were Named Entities. An overview of the information inside the dataset is given in 1.

ZuCo also provides sentiment analysis and relation extraction tasks, which are consistent with the subject's task in reading. To verify that the proposed method can extract information independent of the subject's task, we tested on the NER task.

4.2. Cognitive Features

Gaze features ZuCo provides five features for eye tracking: number of fixations (nFIX): total number of fixations landing on a word; first fixation duration (FFD): first fixation duration on the current word; total reading time (TRT): sum of all fixation durations on the current word; gaze duration (GD): in first-pass reading, the sum of all fixation times on the current word before the gaze moves out of the word; go-past time (GPT): the sum of all fixation times before moving right from the current word (including backward to the previous word, starting from the current word). To increase the robustness of the signal, eye-tracking features were averaged across all subjects. We use these five features as they are and handle them as 5-dimensional Gaze features.

EEG features ZuCo provides word-level EEG features because eye tracking and EEG are recorded at the same time; that is, the EEG corresponding to a given fixation duration can be identified. EEG is

[1]Data is available at: https://osf.io/q3zws/

Model	Signals dim	Performance			
		P(%)	R(%)	F(%)	Δ F(%)
Baseline (Glove)	w/ CharEmb	85.00	81.45	83.10	-
	w/o CharEmb	80.03	77.19	78.50	-4.6
Gaze	Raw	86.18	82.10	83.91	0.81
	Noise	84.26	82.76	83.41	0.31
	PCA	86.25	83.21	84.17	1.07
	MLP	86.51	84.82	85.65	2.55
	CIB	**89.72**	**84.95**	**86.21**	**3.11**
EEG	Raw	83.1	78.67	80.82	-2.28
	Noise	78.7	74.62	76.51	-6.59
	PCA	84.61	80.34	82.16	-0.94
	MLP	86.88	84.32	85.58	2.48
	CIB	**86.93**	**84.56**	**85.83**	**2.73**
Gaze+EEG	Raw	83.57	79.57	81.13	-1.97
	Noise	79.53	72.29	75.77	-7.33
	PCA	84.69	81.75	84.13	1.03
	MLP	85.67	84.83	85.25	2.15
	CIB	**87.42**	**86.25**	**86.54**	**3.44**

Table 2: Compression performance results for the cognitive bottleneck method. The best score for each experimental setting is shown in bold. "Signals dim" means the dimension of a cognitive feature.

Model	Embeddings dim		Signals	Performance			
	Glove	Character		P(%)	R(%)	F(%)	Δ F(%)
Hollenstein et al. (2019)	100	50	Baseline	84.52	81.66	82.92	-
			Gaze	86.19	**84.28**	**85.12**	**2.2**
			EEG	**86.7**	81.5	83.9	0.38
			Gaze+EEG	85.1	83.2	84.0	0.39
CogAlign (Ren and Xiong, 2021)	300	60	Baseline	89.34	78.60	83.48	-
			Gaze	90.76	82.52	86.41	2.93
			EEG	89.87	**83.08**	86.21	2.73
			Gaze+EEG	**91.28**	83.02	**86.79**	**3.31**
CIB	100	50	Baseline	85.00	81.45	83.10	-
			Gaze (Raw)	**89.72**	84.95	86.21	3.11*
			EEG (Raw)	86.93	84.56	85.83	2.73*
			Gaze+EEG (Raw)	87.42	**86.25**	**86.54**	**3.44***

Table 3: Performance comparison between the proposed method and previous work." Gaze" are manually extracted features and "EEG" are features reduced in dimension by averaging. Significance is indicated with the asterisks: * = p<0.01.

provided in 4 frequency bands, divided into Theta (4 to 8 Hz), Alpha (8.5 to 13 Hz), Beta (13.5 to 30 Hz), gamma (30.5 to 49.5 Hz). Frequency bands are said to reflect different functions of cognitive processing in the brain(Meyer, 2018). EEG values are recorded for 105 electrodes, so EEG features have 420 dimensions.

4.3. Baseline

We compared our results to several baselines to test the effectiveness of the cognitive features and the proposed model.[2]

[2]The source code is available at: https://github.com/osekilab/CIB

Cognitive Features **Noise** is the baseline for examining the effect of inputting cognitive features into the model. We used random noise with the same dimension as the cognitive features.

Dimensional Reduction Algorithm **PCA** is simply a baseline that uses Principal Component Analysis to reduce dimension. This is not task-specific. **MLP** is another baseline that uses a simple multilayer perceptron to reduce dimension in a nonlinear fashion. The size of the model is aligned with the CIB module.

NER Performance We compare our model to previous methods on ZuCo dataset. We show how

it differs from their settings. Hollenstein et al. (2019) is the first baseline. They process Raw-Gaze features (5 dimensions) into 17-Gaze features (17 dimensions) that can account for contextual influences. They also transformed the 17-Gaze features to 24 quantiles to normalize them. For our Raw-Gaze features, we do not perform these preprocessing steps and let the CIB module do all the feature extraction. They also averaged all the values for 4 frequency bands, whereas the Raw-EEG features had 420 dimensions. We do not perform such preprocessing to avoid missing information due to drastic averaging. Ren and Xiong (2021) is the next baseline, a strong state-of-the-art model. They follow Hollenstein et al. (2019) and use the same Gaze and EEG features. However, they improved performance by applying the CogAlign method to those cognitive features. Most of the experimental settings are also common to Hollenstein et al. (2019), but the dimension of the word embeddings is different.

4.4. Settings

In our experiments, we adhere to the following settings, which are consistent with previous work: We use precision, recall, and F1 score as evaluation metrics for Named Entity Recognition (NER). Our base model is the BiLSTM-CRF model proposed by Lample et al. (2016), which employs a single layer for both the forward and backward LSTM. We utilize 100-dimensional pre-trained Glove embeddings (Pennington et al., 2014). For character-based embeddings, a bidirectional LSTM is set to 25 dimensions and trained on the ZuCo corpus. We conduct our experiments using 10-fold cross-validation and train with a dropout rate of 0.5. We set the number of hidden LSTM units in the BiLSTM-CRF model to 100.

We also experimented with pre-trained BERT (Devlin et al., 2018). Contextual Transoformer-based word embedding has not been employed in previous work, and we will investigate whether it can be improved with cognitive features against stronger baselines. For the CIB module, we set the compression ratio beta to 0.0001 and the learning rate to 0.01. Regarding the compression dimensions, Gaze features are compressed from 5 dimensions to 3, while EEG features are compressed from 420 dimensions to 10, 20, 30, 50, 100, 200 dimensions.

5. Results

5.1. Exploring the Dimension

The figure2 shows the results of the search for the dimension of feature extraction with CIB. Since the compression ratio is fixed, changing the number of dimensions determines the optimal compression. EEG performs best when compressing from 420 to 10 dimensions, and EEG+Gaze performs best when compressing from 425 to 30 dimensions. EEG+Gaze also performs better with optimal compression. It seems that Gaze and EEG can be improved more when used together than separately. This supports the results of previous work (Hollenstein et al., 2019; Ren and Xiong, 2021).

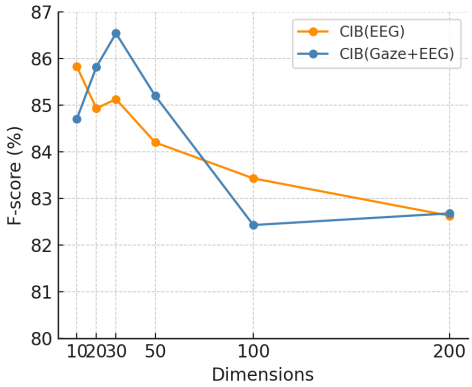

Figure 2: Exploring the dimension of feature extraction with CIB.

Model	Signals	Performance F(%)	Performance Δ F(%)
Baseline (Glove)	-	83.10	-2.14
Baseline (BERT)	-	85.24	-
Gaze	Raw	85.02	-0.22
	Noise	85.70	0.36
	PCA	86.05	0.81
	CIB	**86.60**	**1.36**
EEG	Raw	83.61	-1.63
	Noise	79.87	-5.37
	PCA	84.30	-0.94
	CIB	**86.09**	**0.85**
Gaze + EEG	Raw	83.39	-1.85
	Noise	80.35	-4.89
	PCA	85.69	0.45
	CIB	**87.01**	**1.77**

Table 4: Results of experiments using BERT, a contextual embedding. The results for the best performing cognitive features in terms of compressed dimension are included.

5.2. Effectiveness of CIB compression

The results of the CIB compression performance evaluation are summarized in Table 2, where the Baseline model inputs Glove and Character Embedding into BiLSTM-CRF. All hidden units of the

LSTM in Table 2 are fixed at 100, which means that all settings are the same as in Hollenstein et al. (2019). “EEG” compresses to 10 dimensions, and “EEG +Gaze” compresses to 30 dimensions.

First, we see that the Raw setting does not improve the baseline much or rather worsens it for all cognitive features. The deterioration is marked for EEG and EEG+Gaze. This may be because EEG features with high signal-to-noise ratios have 420 large dimensions, which prevent the model from learning.

For the Noise setting, we used random noise of the same size as the cognitive features in the Raw setting. Random noise produced larger deterioration, indicating that the Raw setting worked better than the random noise.

In the PCA setting, principal component analysis compresses cognitive features. PCA slightly improves the Raw setting for all cognitive features, indicating the effectiveness of dimensional reduction while considering principal components.

In the MLP setting, the model compresses cognitive features nonlinearly, improving task performance. It achieves almost the same performance as the CIB model, indicating that the deep learning model is particularly effective in extracting features from the EEG.

In the CIB setting, the proposed method compresses cognitive features. It significantly improves the PCA setting for all cognitive features, achieving the best scores in all four settings. EEG alone scores lower than Gaze alone, but by combining EEG and Gaze and then compressing them in the CIB setting, the performance is further improved. The improvement is comparable to that achieved using character embedding, and the effect is as good as that of general learning techniques in NER.

5.3. Comparison with previous works

Table 3 compares the performance of the CIB with previous works. Baseline shows the results of training the model without cognitive features in each paper. Note that Hollenstein et al. (2019) and this paper have all the same Baseline setup, but Ren and Xiong (2021) is somewhat different. They use a strong baseline with 300 Glove embedding dimensions and 60 character embedding dimensions, which may result in slightly higher scores overall.

Our CIB model achieved greater improvement than the state-of-the-art model in the Gaze and Gaze+EEG settings. This indicates that CIB extracts useful information from cognitive features. The main difference from previous works is that we use raw cognitive features. For example, for Gaze, CIB automatically retrieves important representations from the 5-dimension raw features without manual selection as in previous works. For EEG, CIB automatically filters out noise from 420 dimensions and extracts only task-specific representations, without drastically averaging the dimensions as in previous works.Our method seems to be particularly effective in the Gaze+EEG setting, suggesting that Gaze can assist in finding richer representations from EEG.

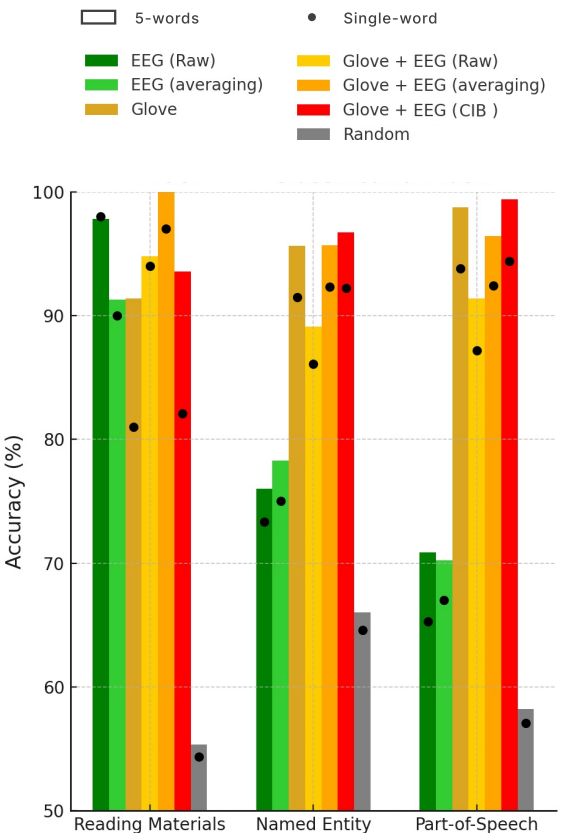

Figure 3: Results of classification task using EEG data and word embedding.“ Random” means chance level, which is a random input feature with the same dimension as the EEG.

5.4. Evaluation with a powerful baseline

Although previous works have only focused on static embeddings such as Glove, it has not been investigated whether contextual embeddings can also be improved with cognitive features. Cognitive features may not be effective for richer embeddings. We performed a similar NER task using 768 dimensional pre-trained BERT embeddings (Devlin et al., 2018) averaged across layers.

The results for the BERT Baseline with 150 LSTM hidden units are shown in Table 4. Baseline improved the F-score by 2.14 compared to Glove. At this powerful baseline, the improvement is modest, but CIB consistently shows improvement. The F-score of the CIB model in the Gaze+EEG setting achieves SoTA for this dataset. The BERT embedding also appears to be less affected by noise due to its 768 dimensions and size.

6. Further Analysis

Our CIB method extracts only useful task-specific information from noisy EEG data to improve performance in downstream tasks. What information did the EEG contain and how did it contribute to the NLP task? Did CIB integrate EEG with word embedding, resulting in richer representations for NLP? We conducted a Probing task and decoded the information that the original EEG and the integrated representations contained.

6.1. Probing Tasks

We conducted a Decoding task inspired by Murphy et al. (2022), in which three different tasks illustrate the representation. **Reading Materials** is a binary classification of which document the subject was reading during the EEG recording. As mentioned in Section 4, the dataset consists of two types of materials: movie reviews and Wikipedia. While not obvious, distinguishing between them may be useful for the NER task. For example, the movie reviews are all PER labels. **Named Entity** is also a binary classification of whether or not it is a Named Entity for a given noun. This is not sequential labeling, but a simple task to measure understanding of Named Entity. **Part-of-Specch** is a three-valued classification of nouns, verbs, and prepositions, a frequent POS. Syntactic understanding is critical to the model in order to extract entire noun phrases as Named entities in a sentence.

Following Murphy et al. (2022), a small Transformer performs the classification task. Almost all parameters are the same as in their model, but since the dataset size is different, we set the encoder layers to 2, mlp size to 128, and qkv size to 64. The Decoding model is trained, validated and tested on the same data as the NER model. Note that the chance level for binary classification is therefore not 50%. In **Single-Word**, only one word is used as input, while in **5-Words**, five words of the same class are used as input in sequence.

To distinguish between CIB integration and mere concat effects, multiple conditions are implemented. In EEG (Raw), the 420 dimensions of the word are taken as input; in EEG (averaging), the 420 dimensions are averaged and reduced to 8 dimensions, the same way as in Hollenstein et al. (2019). Glove+EEG (Raw) simply concatenates the Glove vector and the 420 dimensional EEG, and Glove+EEG (averaging) simply concatenates the Glove vector and the 8 dimensional averaged EEG. Glove+EEG (CIB) reduces the EEG to 10 dimensions using the CIB method. It is expected that EEG (CIB) is a minimal and sufficient representation to complement Glove. We ran 10 experiments with different seed and evaluated the results by averaging the accuracies.

6.2. Evaluation

The figure3 shows the results of the probing task. First, we find that the EEG data outperform the overall chance level (Random) and have extensive information about NLP and subjects. In particular, Reaing Materials shows almost no loss of accuracy even in the Single word condition. On the contrary, it seems to be poor at decoding NLP information. In the other two tasks, it is surprising that the averaged EEG performs better than the original data. Averaging is a simple but effective strategy and confirms the results of previous work.

Glove embedding is generally more accurate. Reading Materials as information about the subject is comparable to EEG, while Named Entity and POS perform much better than EEG. In Glove+EEG (Raw) and Glove+EEG (averaging), simple concat is not working very well. However, in Reading Materials, simple concatenation with EEG performs the best, which may be due to the fact that EEG is superior to Glove in this task. In other tasks, EEG has a rather negative influence on Glove, which can be prevented to some extent by averaging.

In Glove+EEG (CIB), the integration of EEG and Glove seems to work well: for two tasks, the integration with EEG successfully improves performance. Reading Materials does not show such an improvement, and it may not have been necessary to supplement the EEG data for the NER task. These results suggest that CIB can extract only the information necessary for a task from EEG data with various information and complement existing word embedding. Conversely, Glove + EEG (CIB) has the best performance for Named Entity and POS decoding. It outperforms simply averaged EEG results, indicating that the CIB module preserves the respective information from the EEG data well.

7. Conclusion

We proposed the Cognitive Information Bottleneck method for extracting task-specific representations from cognitive language processing signals such as eye tracking and EEG data, and evaluated it on public datasets. The proposed method compresses information more efficiently than the existing method, and achieves an improvement over state-of-the-art models in downstream tasks using cognitive processing signals; experiments with BERT, a contextual embedding, confirm its effectiveness, and the results achieve a new state-of-the-art. This approach can be used for any task other than NER. It facilitates the retrieval of information for specific purposes from cognitive processing signals and promotes the integration of human responses into NLP models. Future work should improve the architecture for efficient training when used in conjunction with pre-trained large-scale language models.

Ethical considerations

This study uses an existing dataset, ZuCo (Hollenstein et al., 2018), which has been sufficiently designed to ensure that there are no ethical concerns. The main uses of these data are for training in machine learning and natural language processing, and for analyzing the human reading process. Our objectives are consistent with these.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This work was supported by JST PRESTO Grant Number JPMJPR21C2, Japan.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning*, pages 302–312.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 579, page 584.

Guanghai Dai, Jun Zhou, Jiahui Huang, and Ning Wang. 2020. Hs-cnn: a cnn with hybrid convolution scale for eeg motor imagery classification. *Journal of neural engineering*, 17(1):016025.

Michael Dambacher and Reinhold Kliegl. 2007. Synchronizing timelines: Relations between fixation durations and n400 amplitudes during sentence reading. *Brain research*, 1155:147–162.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

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*.

Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Matthew Botvinick, Hugo Larochelle, Yoshua Bengio, and Sergey Levine. 2019. Infobot: Transfer and exploration via the information bottleneck. *arXiv preprint arXiv:1901.10902*.

Nora Hollenstein, Maria Barrett, Marius Troendle, Francesco Bigiolli, Nicolas Langer, and Ce Zhang. 2019. Advancing nlp with cognitive language processing signals. *arXiv preprint arXiv:1904.02682*.

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.

Nora Hollenstein and Ce Zhang. 2019. Entity recognition at first sight: Improving ner with eye movement information. *arXiv preprint arXiv:1902.10068*.

Barbara J Juhasz and Keith Rayner. 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1312.

Varun Khurana, Yaman Kumar Singla, Nora Hollenstein, Rajesh Kumar, and Balaji Krishnamurthy. 2023. Synthesizing human gaze feedback for improved nlp performance. *arXiv preprint arXiv:2302.05721*.

Marta Kutas and Kara D Federmeier. 2000. Electrophysiology reveals semantic memory use in language comprehension. *Trends in cognitive sciences*, 4(12):463–470.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Xiang Lisa Li and Jason Eisner. 2019. Specializing word embeddings (for parsing) by information bottleneck. *arXiv preprint arXiv:1910.00163*.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2021. Variational information bottleneck for effective low-resource fine-tuning. *arXiv preprint arXiv:2106.05469*.

Sijie Mai, Ying Zeng, and Haifeng Hu. 2022. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*.

Lars Meyer. 2018. The neural oscillations of speech processing and language comprehension: state

of the art and emerging mechanisms. *European Journal of Neuroscience*, 48(7):2609–2621.

Alex Murphy, Bernd Bohnet, Ryan McDonald, and Uta Noppeney. 2022. Decoding part-of-speech from human EEG signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2201–2210, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. 2018. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Keith Rayner. 1977. Visual attention in reading: Eye movements reflect cognitive processes. *Memory & cognition*, 5(4):443–448.

Yuqi Ren and Deyi Xiong. 2021. Cogalign: Learning to align textual neural representations to cognitive language processing signals. *arXiv preprint arXiv:2106.05544*.

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha. 2017. Using gaze data to predict multiword expressions.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Michalina Strzyz, David Vilares, and Carlos Gómez-Rodríguez. 2019. Towards making a dependency parser see. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1500–1506, Hong Kong, China. Association for Computational Linguistics.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Sabine Weiss and Horst M Mueller. 2003. The contribution of eeg coherence to the investigation of language. *Brain and language*, 85(2):325–343.

Ce Zhang, Young-Keun Kim, and Azim Eskandarian. 2021. Eeg-inception: an accurate and robust end-to-end neural network for eeg-based motor imagery classification. *Journal of Neural Engineering*, 18(4):046014.