Towards a Deep and Unified Understanding of Deep Neural Models in NLP

Chaoyu Guan*[2], Xiting Wang*[2], **Quanshi Zhang[1]**, Runjin Chen[1], Di He[2], Xing Xie[2]

*Equal Contribution

[1]John Hopcroft Center and the MoE Key Lab of Artificial Intelligence, AI Institute, at the Shanghai Jiao Tong University, Shanghai, China

[2]Microsoft Research Asia, Beijing, China

Introduction

A key task in explainable AI is to associate latent representations with input units by

quantifying layerwise information discarding of inputs.

Most explanation methods (e.g., DNN visualization) have **coherency** & **generality** issues

- Coherency: requires that a method generates consistent explanations across different neurons, layers, and models.
- Generality: existing measures are usually defined under certain restrictions on model architectures or tasks.

Methods	Coherency			Generality
	Neuron	Layer	Model	
Gradient-based	✓	×	×	×
Inversion-based	✓	×	×	×
LRP	×	×	×	×
Ours	✓	✓	✓	✓

Our solution

Considering both coherency and generality

- **A unified information-based measure:** quantify the information of each input word that is encoded in an intermediate layer of a deep NLP model.
- **The information-based measure as a tool**
  - Evaluating different explanation methods.
  - Explaining different deep NLP models
- This measure enriches the capability of explaining DNNs.

Problem

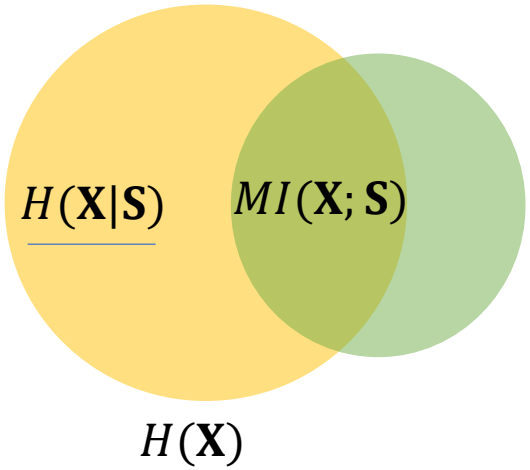
- **Quantification of sentence-level information discarding:** quantify the information of an entire sentence **x** that is encoded in **s**.
- **Quantification of word-level information discarding**: quantify the information of each specific word $\mathbf{x}_i$ that is encoded in **s**.
- **Fine-grained analysis of word attributes**: analyze the fine-grained reason why **s** uses the information of $\mathbf{x}_i$.

$\mathbf{x} = [\mathbf{x}_1^T, \dots, \mathbf{x}_n^T]^T \in \mathbf{X}$: Input sentence $\mathbf{x}_i^T$: word embedding

$\mathbf{s} = \Phi(\mathbf{x}) \in \mathbf{S}$: hidden state $\Phi(\cdot)$: function of the intermediate layer

Word Information Quantification

Multi-Level Quantification



$H(\mathbf{X}_i|\mathbf{s}=\Phi(\mathbf{x}))$ reflects how much information from word $\mathbf{x}_i$ is discarded by $\mathbf{s}$ during the forward propagation.

$$MI(\mathbf{X};\mathbf{S}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{S})$$

$$H(\mathbf{X}|\mathbf{S}) = \int_{\mathbf{s}\in\mathbf{S}} p(\mathbf{s}) H(\mathbf{X}|\mathbf{s}) d\mathbf{s}$$

Corpus level

$$H(\mathbf{x}) = -\int_{\mathbf{x}'\in\mathbf{X}} p(\mathbf{x}'|\mathbf{s}) \log p(\mathbf{x}'|\mathbf{s}) d\mathbf{x}'$$

Sentence level

$$H(\mathbf{X}|\mathbf{s}) =^{*} \sum_i H(\mathbf{X}_i|\mathbf{s})$$

$$H(\mathbf{X}_i|\mathbf{s}) = -\int_{\mathbf{x}'_i\in\mathbf{X}_i} p(\mathbf{x}'_i|\mathbf{s}) \log p(\mathbf{x}'_i|\mathbf{s})\, d\mathbf{x}'_i$$

Word level

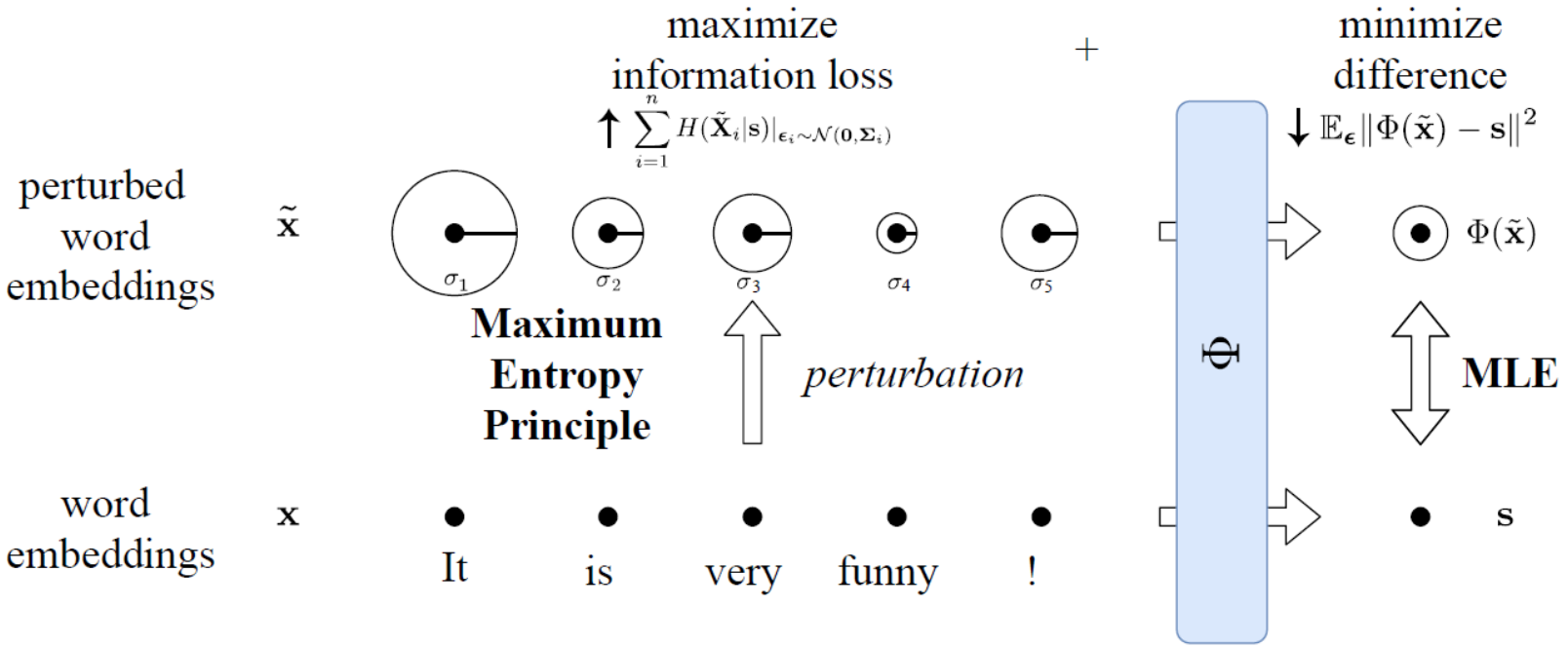
* Suppose the words are independent in one sentence.

Word Information Quantification

Perturbation-based Approximation

We use $H(\widetilde{\mathbf{X}}_i|\mathbf{s})$ to approximate $H(\mathbf{X}_i|\mathbf{s})$ by minimizing the following loss:

$$L(\boldsymbol{\sigma}) = \mathbb{E}_{\boldsymbol{\epsilon}}\left|\left|\Phi(\tilde{\mathbf{x}}) - \mathbf{s}\right|\right|^2 - \lambda \sum_{i=1}^{n} H(\widetilde{\mathbf{X}}_i|\mathbf{s}) \Big|_{\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma_i^2 \mathbf{I})}$$



Fine-Grained Analysis of Word Attributes

Disentangle the information of a common concept $\mathbf{c}$ away from each word $\mathbf{x}_i$

$$A_i = \log p(\mathbf{x}_i|\mathbf{s}) - \mathbb{E}_{\mathbf{x}'_i \in \mathbf{X}_i} \log p(\mathbf{x}'_i|\mathbf{s})$$

Importance of the i-th word concerning random words

$$A_{\mathbf{c}} = \mathbb{E}_{\mathbf{x}'_i \in \mathbf{X}_{\mathbf{c}}} \log p(\mathbf{x}'_i|\mathbf{s}) - \mathbb{E}_{\mathbf{x}'_i \in \mathbf{X}_i} \log p(\mathbf{x}'_i|\mathbf{s})$$

Importance of the common concept $\mathbf{c}$ w.r.t. random words

$r_{i,\mathbf{c}} = A_i - A_{\mathbf{c}}$ indicates the remaining information of the word $\mathbf{x}_i$ when we remove the information of the common attribute $\mathbf{c}$ from the word.

Comparative Study

- Three baselines: LRP, gradient-based, perturbation
- Conclusion: our method provides the most faithful explanations for
  - Across timestamp analysis
  - Across layer analysis
  - Across model analysis

Our method clearly shows that the model gradually focuses on the most important parts of the sentence.

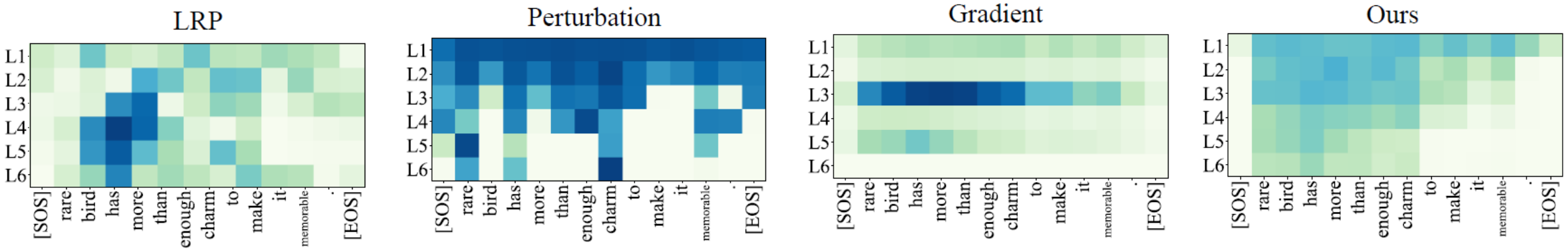


Figure 3. Saliency maps of different **layers** comparing with three baselines. Our method shows how information decreases through layers.

Understanding Neural Models in NLP

We explain four NLP models (BERT, Transformer, LSTM, and CNN):

- What information is leveraged for prediction?
- How does the information flow through layers?
- How do the models evolve during training?

	SST-2 (Acc)	CoLA (MCC)	QQP (Acc)
BERT	**0.9323**	**0.6110**	**0.9129**
Transformer	0.8245	0.1560	0.7637
LSTM	0.8486	0.1296	0.8658
CNN	0.8200	0.0985	0.8099

Understanding Neural Models in NLP

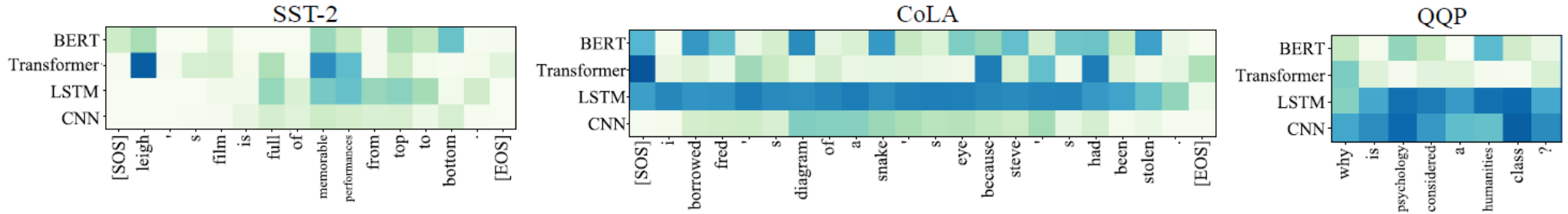


Figure 5. Words that different models use for prediction. For QQP, we only show the first question from the question pair.

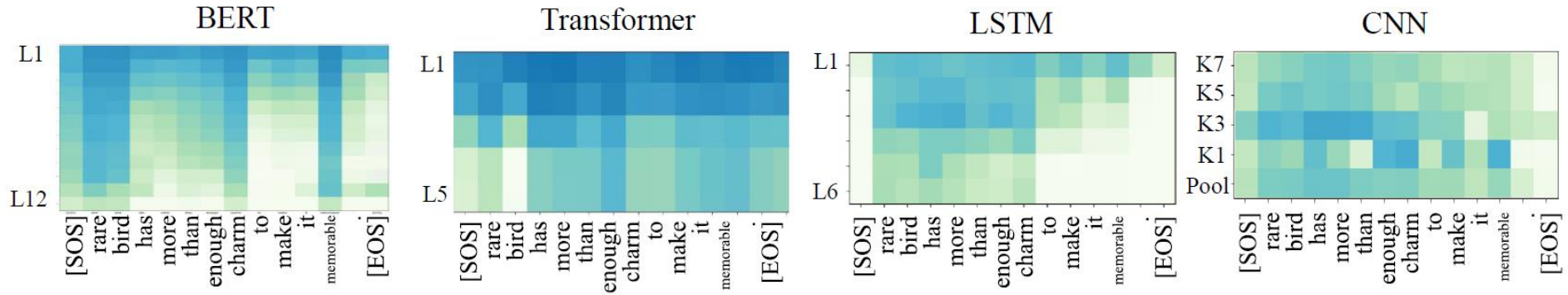


Figure 7. Layerwise analysis of word information. For all models other than CNN, the information gradually decreases through layers

- **Bert and Transformer use words for prediction, while LSTM and CNN use subsequences of sentences for prediction.**
- **Different models process the input sentence in different manners.**

Towards A Deep and Unified Understanding of Deep Neural Models in NLP

Please visit our poster at **#62**!