Poster ID: 218
Paper ID: 8629
Oral: Friday 3:45pm

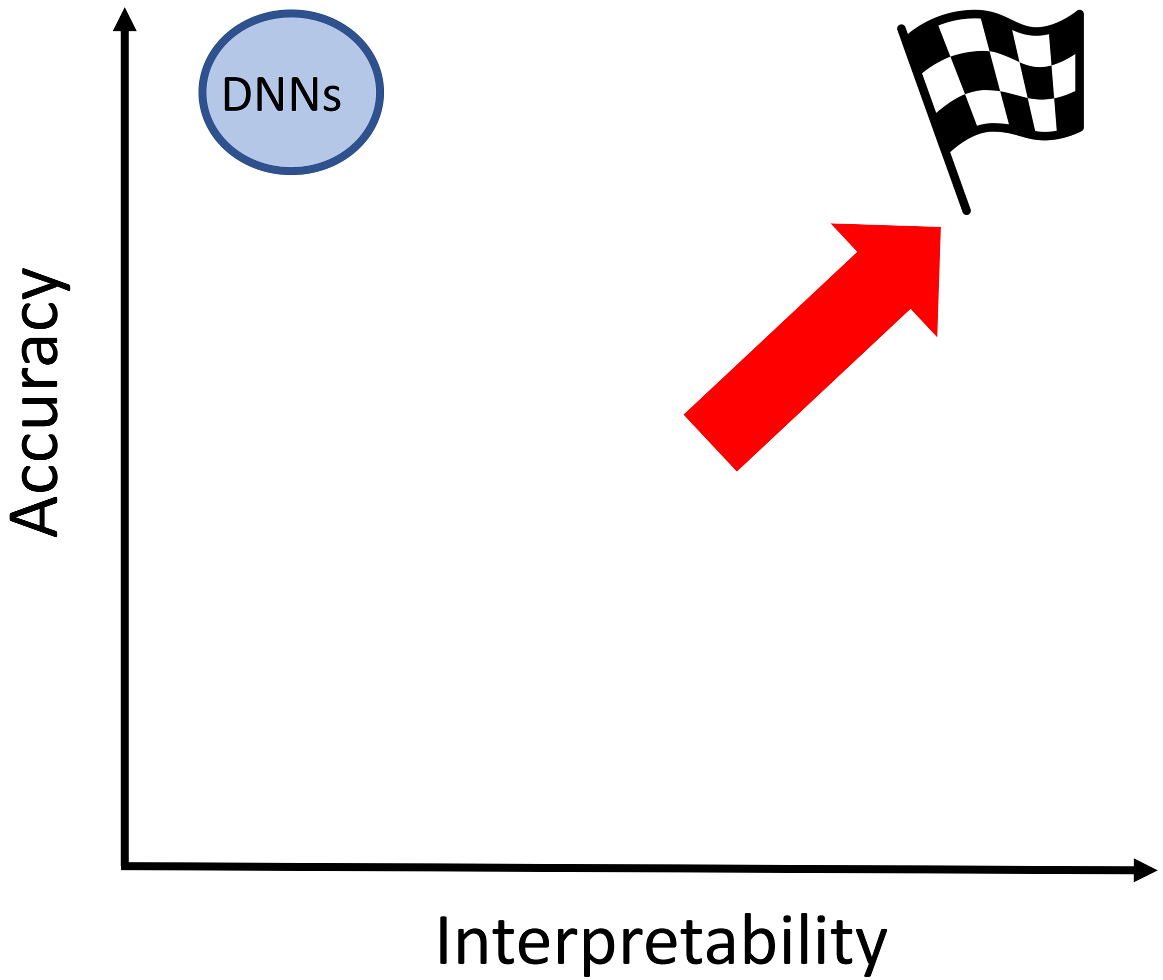
UNIVERSITY of VIRGINIA

Improving Interpretability via Explicit Word Interaction Graph Layer

Arshdeep Sekhon, Hanjie Chen, Aman Shrivastava, Zhe Wang, Yangfeng Ji, **Yanjun Qi**
University of Virginia, Charlottesville, USA




Goal: Improving Interpretability of NLP Models



Basic Intuition: Word Interactions are Ubiquitous

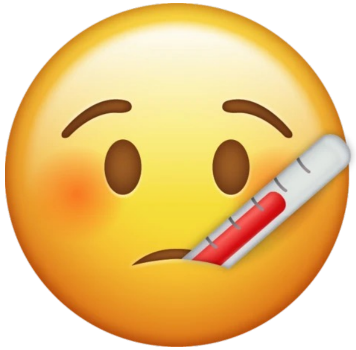
entertaining

satisfactory

well

engaging

fails



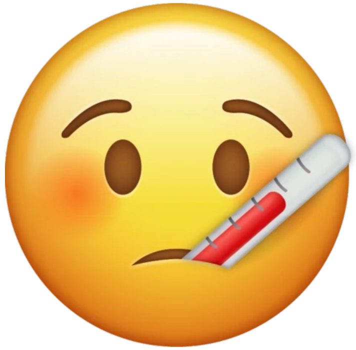
Basic Intuition: Word Interactions are Ubiquitous

Not entertaining

Satisfactory, but

Not well

Not engaging



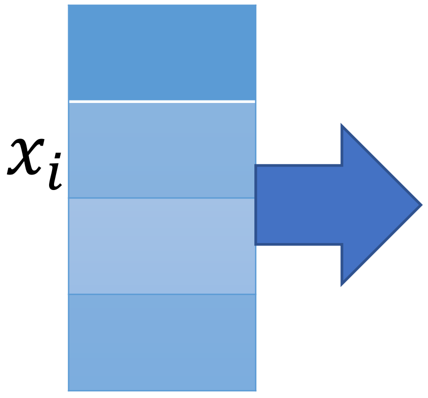
Never fails

Basic Intuition: Word Interactions are Ubiquitous

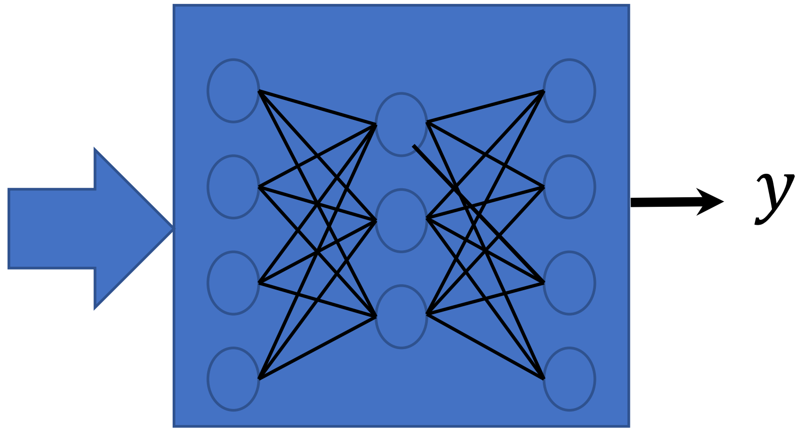
`“take care of my cat” offers a **refreshingly different** slice of asian cinema`

‘different’ highly relates to the word ‘refreshingly’, it will likely contribute substantially to the model’s sentiment prediction

Basic Idea: Word Interaction Graph Layer (WIGRAPH)



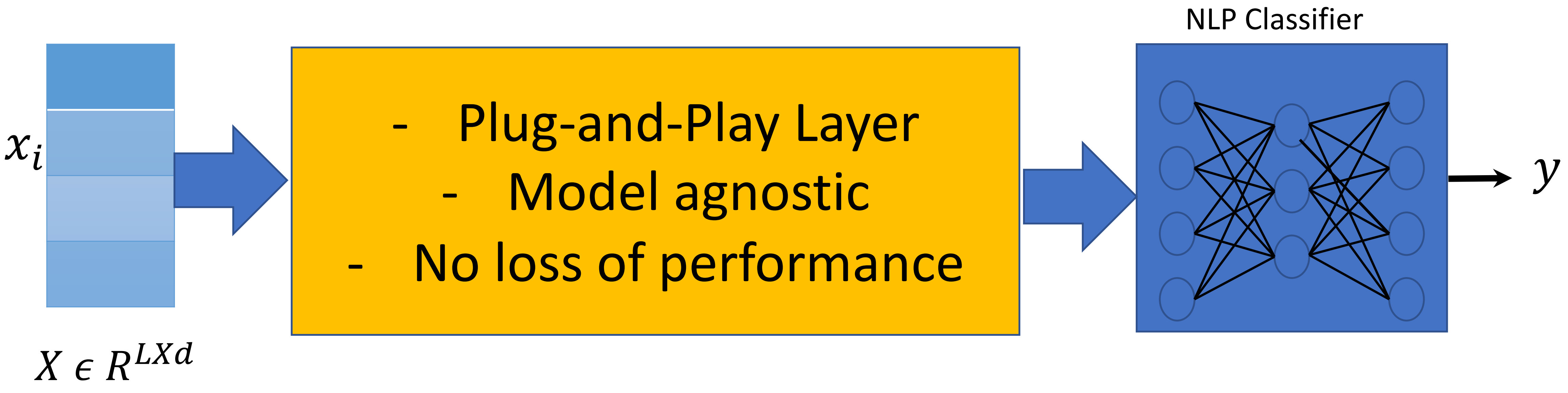
$X \in R^{LXd}$



Input Sentence of length L

A regular NLP Pipeline

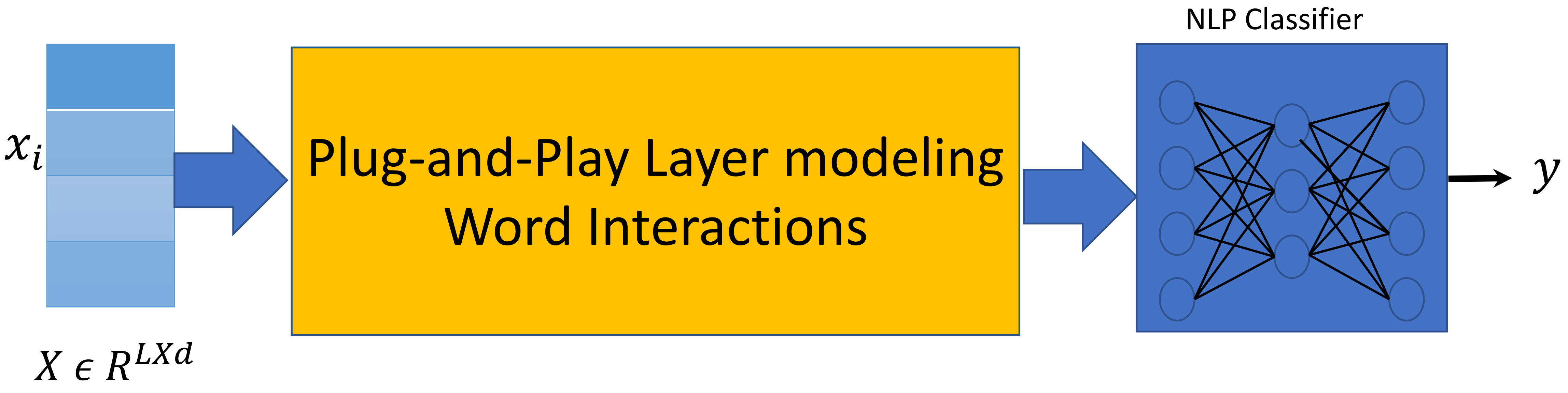
Basic Idea: Word Interaction Graph Layer(WIGRAPH)



Input Sentence of length L

A regular NLP Pipeline

Basic Idea: Word Interaction Graph Layer(WIGRAPH)

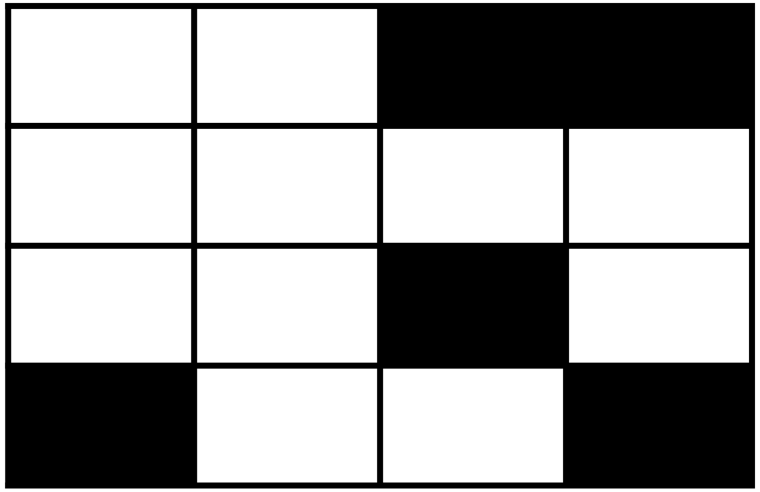


Input
Sentence of
length L

A regular NLP Pipeline

Basic Idea: Word Interaction Graph Layer (WIGRAPH)

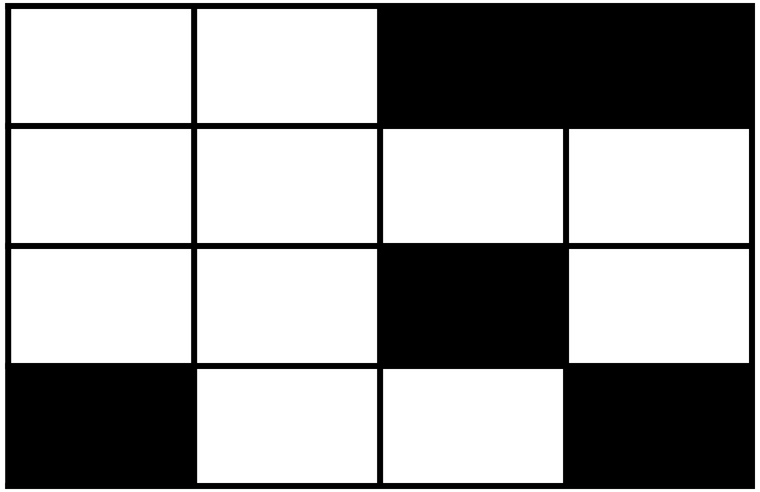
- A layer enhances a BASE model's decision-making process by providing explicit guidance on what words are more important using the information on those words they interact with.



an interaction graph (mask):

Basic Idea: Word Interaction Graph Layer (WIGRAPH)

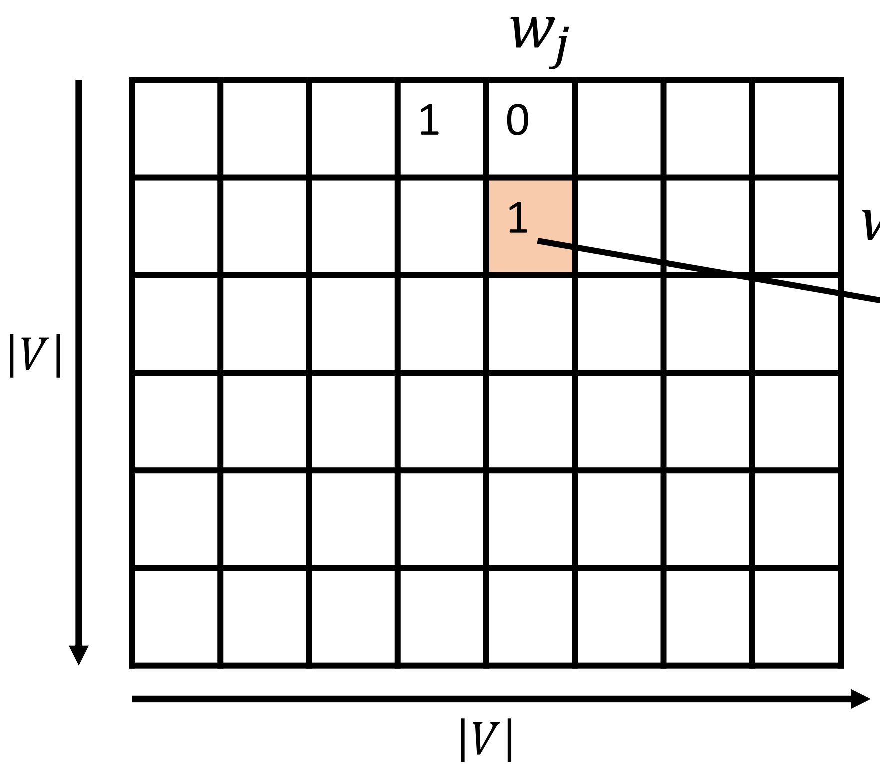
- A layer enhances a target model's decision-making process by providing explicit guidance on what words are more important using the information on those words they interact with.
- These task-level important word interactions need to be learnt



an interaction graph (mask): not all pairwise relations are informative

How to Model:

Word Interaction Graph A Through a Learnt Matrix Parameter

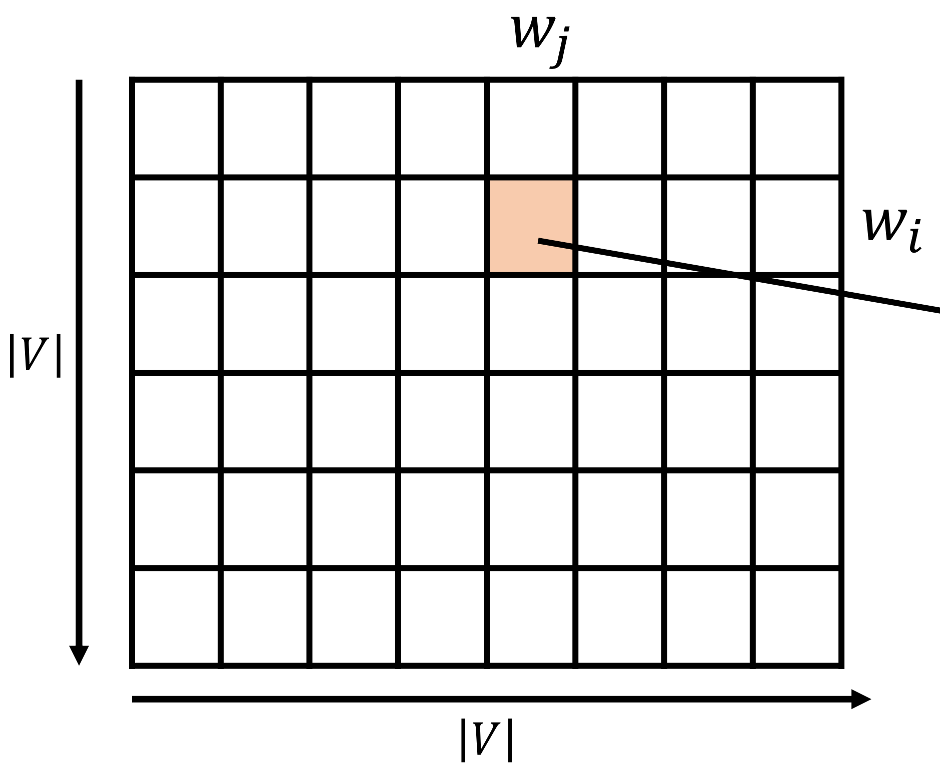


Word Interaction Graph A

A $_{ij}$ ~ Ber(Sigmoid(γ_{ij}))

Place a distribution over A
(stochastic edges)

Word Interaction Graph A Through a Learnable Matrix Parameter

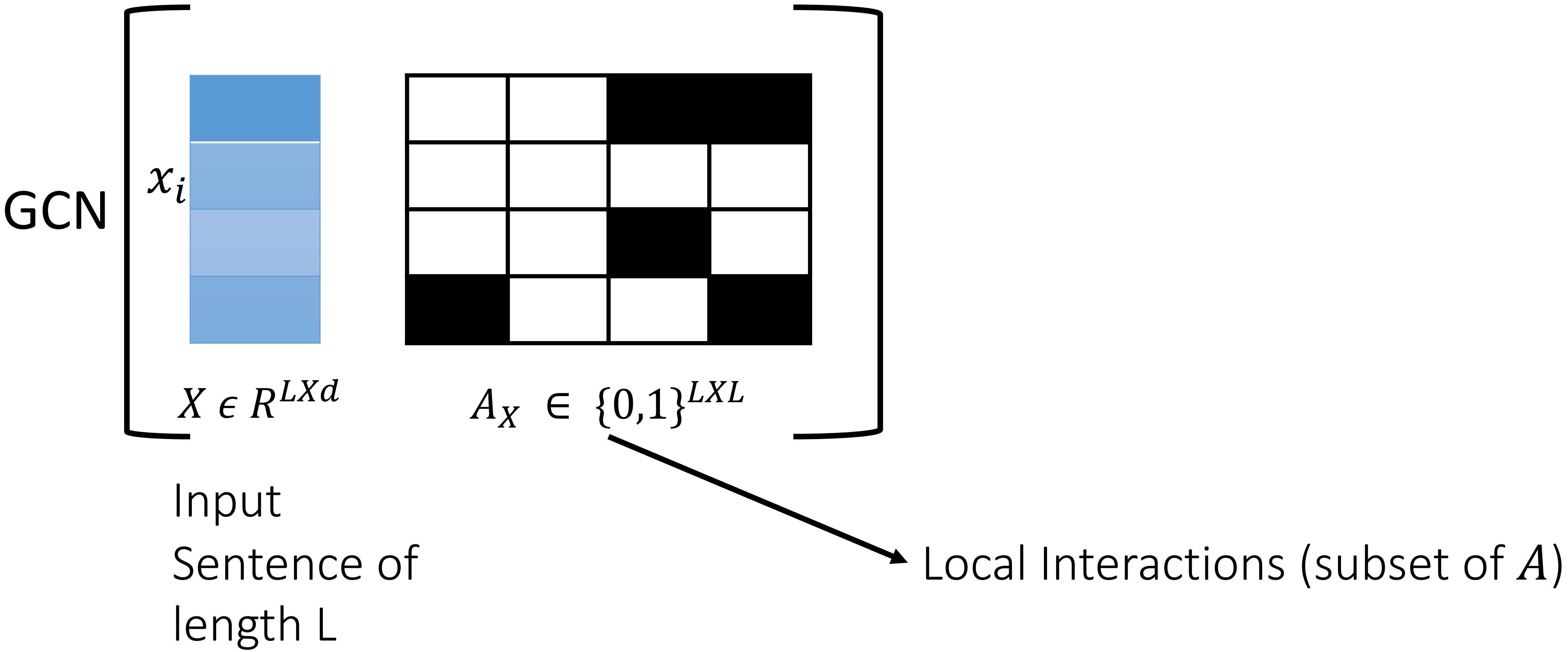


Probability of word i and word j interacting: parameterized Bernoulli distribution

To learn A is to learn its Matrix Parameter γ

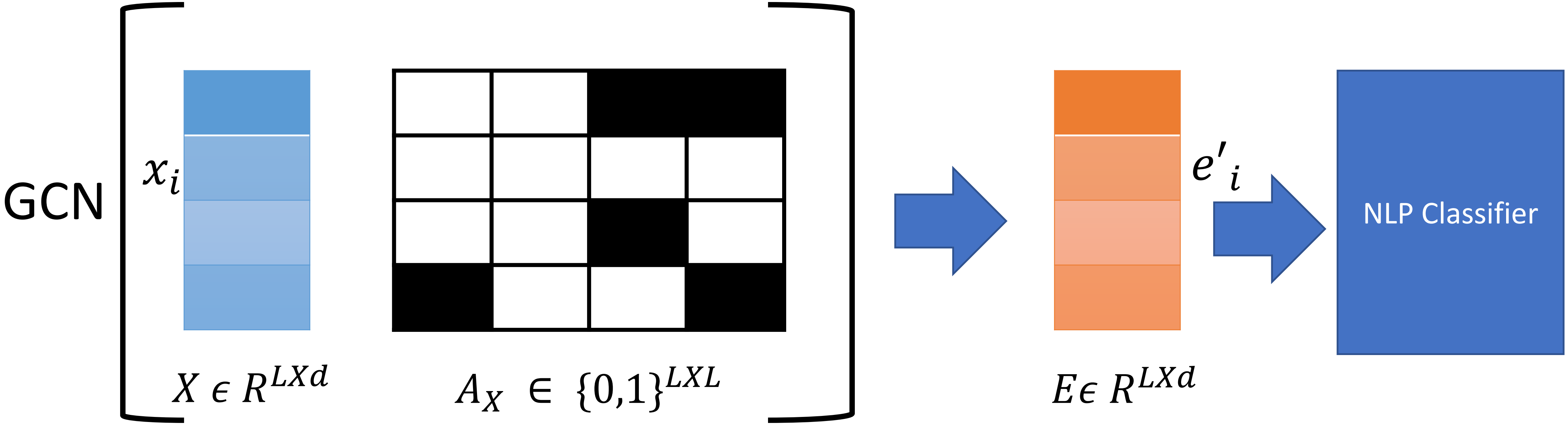
How to Use:

WIGRAPH derived Embeddings: Local Interactions



How to Use:

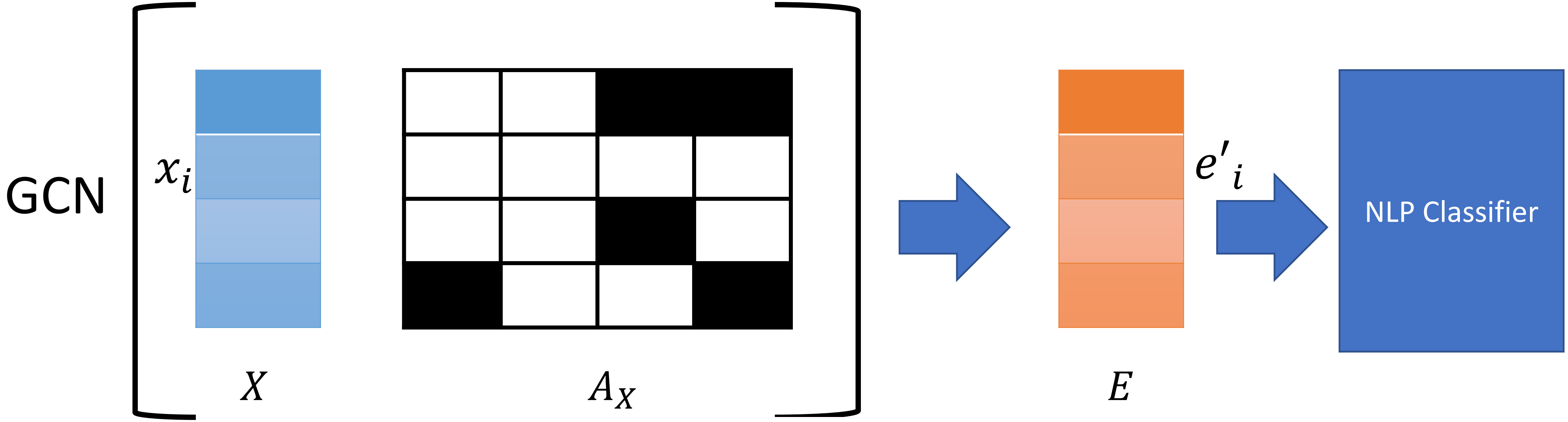
WIGRAPH derived Embeddings: Interaction Based Embeddings



Input Sentence of length L

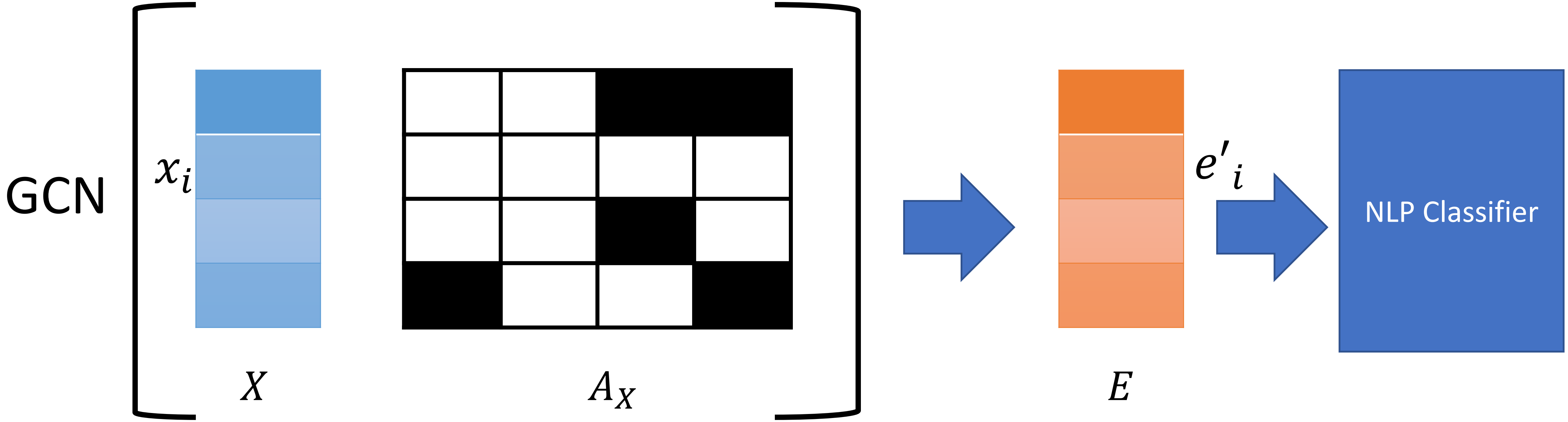
Updated Embeddings based on Interactions

WIGRAPH derived Embeddings: Graph Convolution Message Passing



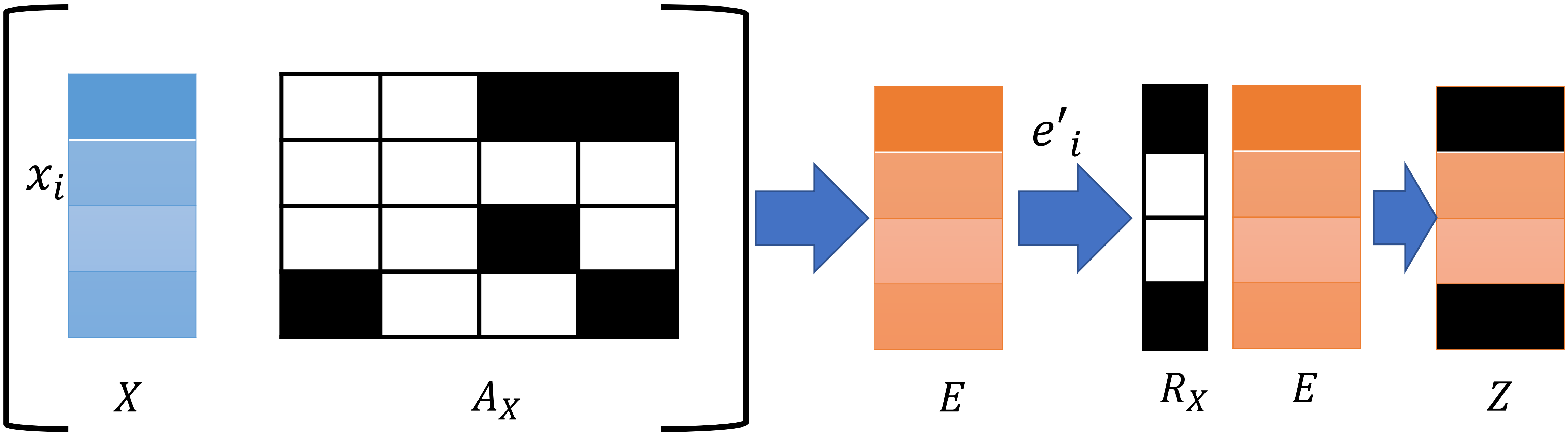
$$\mathbf{e}'_i = \mathbf{x}_i + \sigma\left(\frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)} \mathbf{x}_j\right)$$

WIGRAPH Embeddings: An Interaction Mask



- an interaction mask: not all interactions are informative
- A needs to be learnt

WIGRAPH Variation: Masking Words also



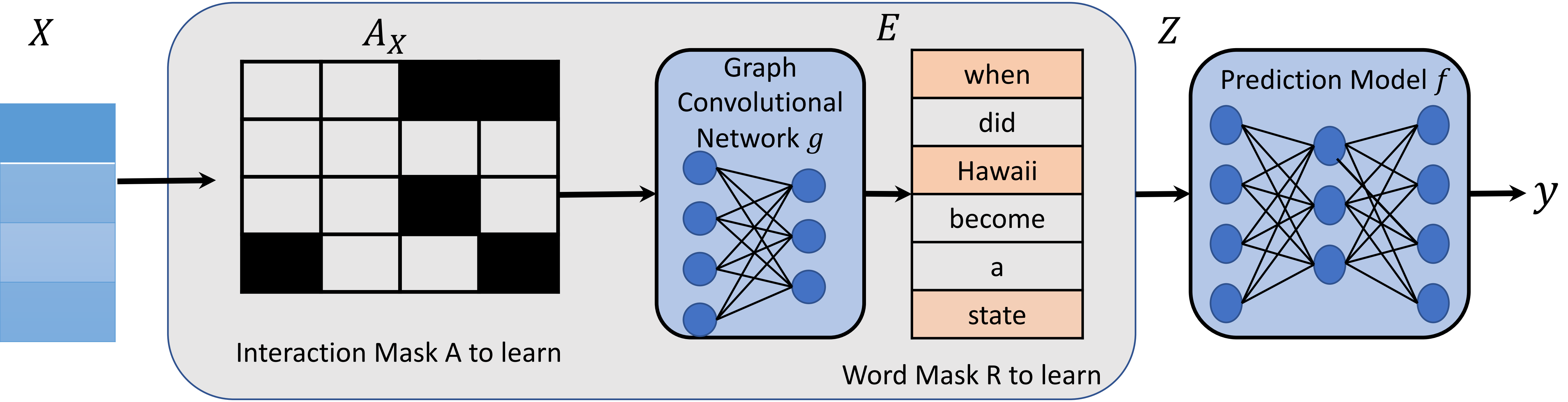
GCN

Word Mask: not all words are informative for a specific task

$$\mathbf{z}_i = \mathbf{R}_{\mathbf{x}_i} \mathbf{e}'_i$$

How to Use:

WIGRAPH Layer: Plug-and-Play and NLP Model Agnostic

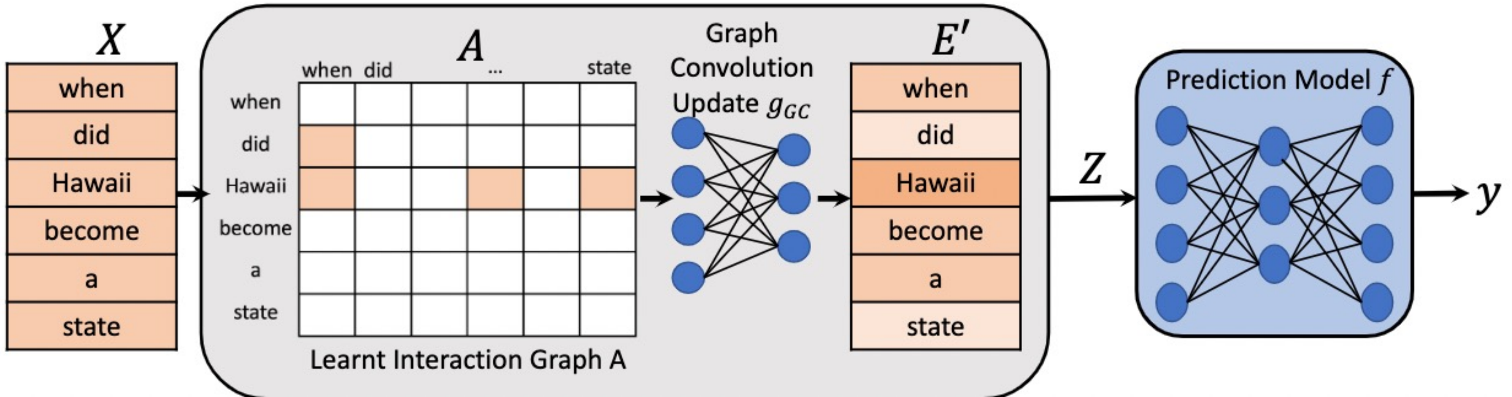


How to Learn WIGRAPH: Variational Information Bottleneck Loss

$$max_{\mathbf{A},\mathbf{R},\{\mathbf{W}\}}\{I(\mathbf{Z};\mathbf{Y}) - \beta I(\mathbf{Z};\mathbf{X})\}$$

Maximize Information between Z and Y:
Z maximally predictive of Y

Minimize Information between Z and X:
Remove non-relevant features



WIGRAPH: Variational Information Bottleneck Loss for Word Interaction

$$max_{\mathbf{A},\mathbf{R},\{\mathbf{W}\}}\{I(\mathbf{Z};\mathbf{Y})-\beta I(\mathbf{Z};\mathbf{X})\}$$

Maximize Information between Z and Y:
Z maximally predictive of Y

Minimize Information between Z and X:
Remove non-informative interactions and words

$$\begin{aligned} max_{\mathbf{A},\mathbf{R},\{\mathbf{W}\}}\{&\mathbb{E}_{q(\mathbf{Z}|\mathbf{x}^m)}log(p(\mathbf{y}^m|\mathbf{x}^m;\mathbf{A},\mathbf{R},\{\mathbf{W}\})) \\ &-\beta_i KL(q(\mathbf{R}|\mathbf{x}^m)||p_{r0}(\mathbf{R})) \\ &-\beta_g KL(q(\mathbf{A}|\mathbf{x}^m)||p_{a0}(\mathbf{A}))\} \end{aligned}$$

Ber(0.5)

WIGRAPH: Variational Information Bottleneck Loss for Word Interaction

$$max_{\mathbf{A},\mathbf{R},\{\mathbf{W}\}}\{I(\mathbf{Z};\mathbf{Y}) - \beta I(\mathbf{Z};\mathbf{X})\}$$

Maximize Information between Z and Y:
Z maximally predictive of Y

Minimize Information between Z and X:
Remove non-relevant features

$$-(\mathbb{E}_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}^m;\mathbf{A},\mathbf{R},\{\mathbf{W}\}) + \beta_i H_q(\mathbf{R}_{\mathbf{x}}|\mathbf{x}^m) + \beta_g H_{\mathbf{q}}(\mathbf{A}_{\mathbf{x}}|\mathbf{x}^m)) + \beta_{sparse}||\mathbf{A}_{\mathbf{x}}||_1$$

Deep Variational Information Bottleneck
Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, Kevin Murphy

Gumbel softmax to sample discrete A, while keeping the training differentiable end to end.

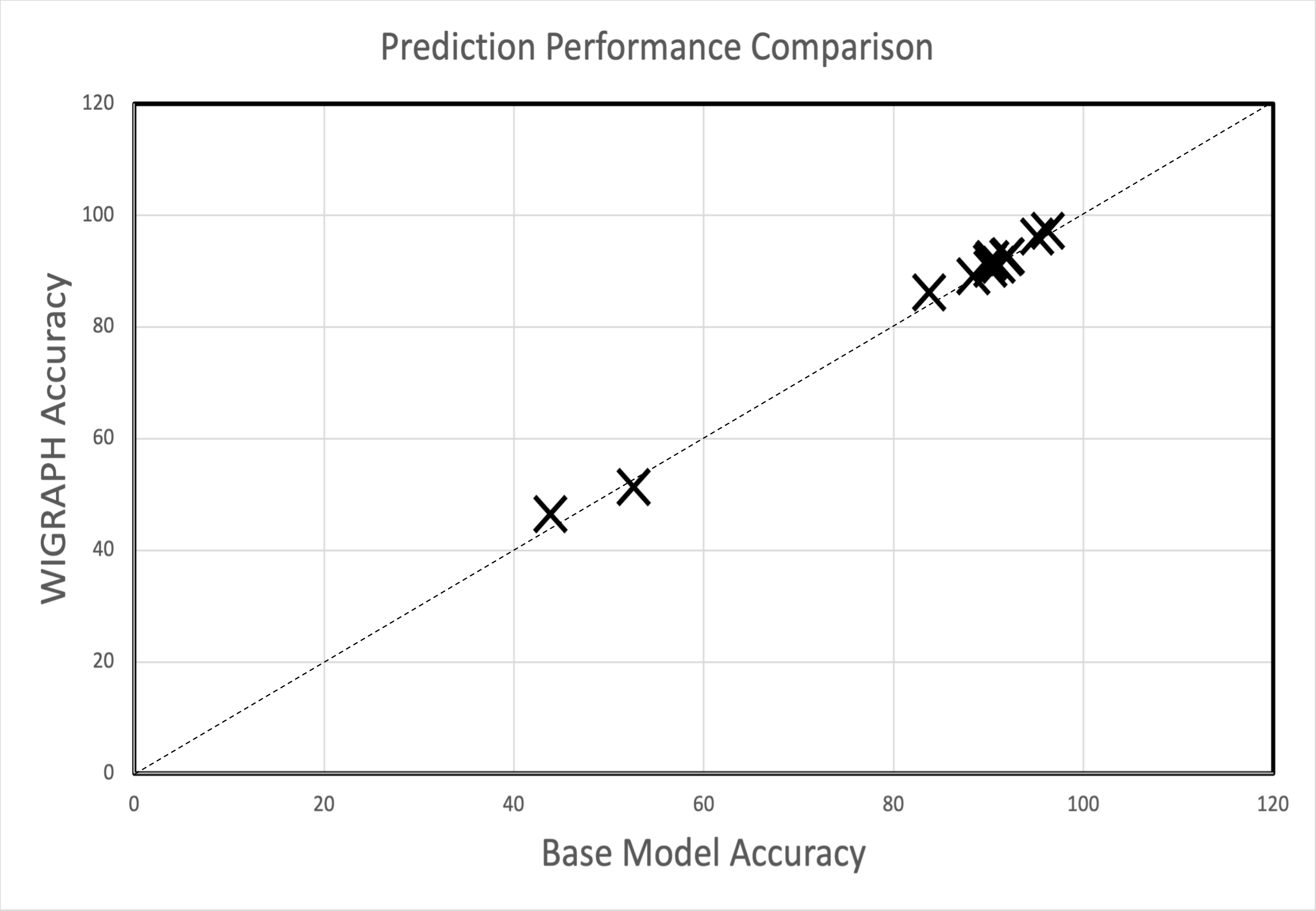
Related Work

- Post-Hoc Explanation Methods:
  - LIME, SampleShapley: *post hoc* word level importance scores
  - Can’t improve models’ intrinsic interpretability
- Inherently interpretable model
  - (Alvarez-Melis and Jaakkola 2018a; Rudin 2019)
  - Intensive engineering efforts
- User-specified priors as domain knowledge to guide model
  - (Cam- buru et al. 2018; Du et al. 2019; Chen and Ji 2019; Erion et al. 2019; Molnar, Casalicchio, and Bischl 2019)
  - Information priors not be available in many tasks.
- Special layer to improve models:
  - VMASK (Chen and Ji 2020) : a special case of our method

Evaluation: Improves Prediction Performance

Dataset	Train/Dev/Test	C	V	L
sst1	8544/1101/2210	5	17838	50
sst2	6920/872/1821	2	16190	50
imdb	20K/5K/25K	2	29571	250
AG News	114K/6K/7.6K	4	21838	50
TREC	5000/452/500	6	8026	15
Subj	8000/1000/1000	2	9965	25

Models: LSTM, BERT, RoBERTa, distilBERT



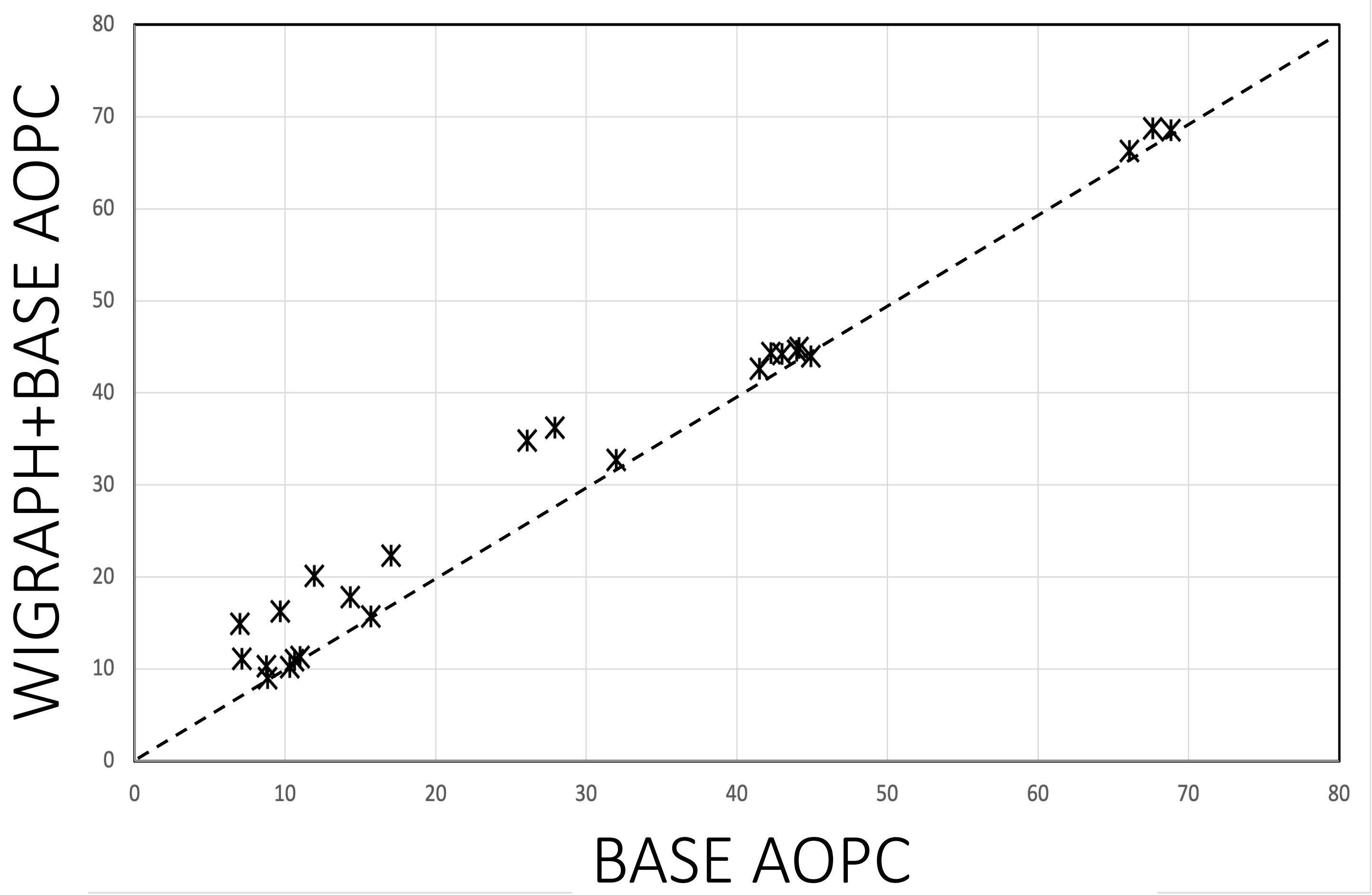
Evaluation: Local Interpretability

- Local Interpretability Faithfulness: Area Over Perturbed Curve(AOPC)
  - average change of prediction probability on the predicted class over a test dataset by deleting top k words in explanations ((Nguyen 2018; Samek et al. 2016))

$$AOPC = \frac{1}{K+1} \sum_{k=1}^{K} < f(\mathbf{x}) - f(\mathbf{x}_{\backslash 1,\ldots,k}) >_{p(\mathbf{x})}$$

  - Generate local explanations using LIME/SampleShapley and compare AOPC.

Improves Local Interpretability



Higher AOPC scores reflect better interpretation faithfulness.

Improves Local Interpretability (better word attribution)

Model	Explanation
BASE	still , this thing feels flimsy and ephemeral
WIGRAPH	still , this thing feels flimsy and ephemeral
BASE	so young , so smart , such talent , such a wise
WIGRAPH	so young , so smart , such talent , such a wise
BASE	it is risky , intelligent , romantic and rapturous from start to finish
WIGRAPH	it is risky , intelligent , romantic and rapturous from start to finish
BASE	take care of my cat offers a refreshingly different slice of asian cinema
WIGRAPH	take care of my cat offers a refreshingly different slice of asian cinema

Evaluation: Interaction Interpretability

- Interaction Occlusion Score(IoS)
  - sort entries of A and filter out the top K global interaction scores, denoted by $\boldsymbol{A}_{ij}^{k}$
  - calculate the average accuracy of the model after only using these top k interactions

IoS measures the interaction interpretability faithfulness of a target model on its learnt interactions.

27

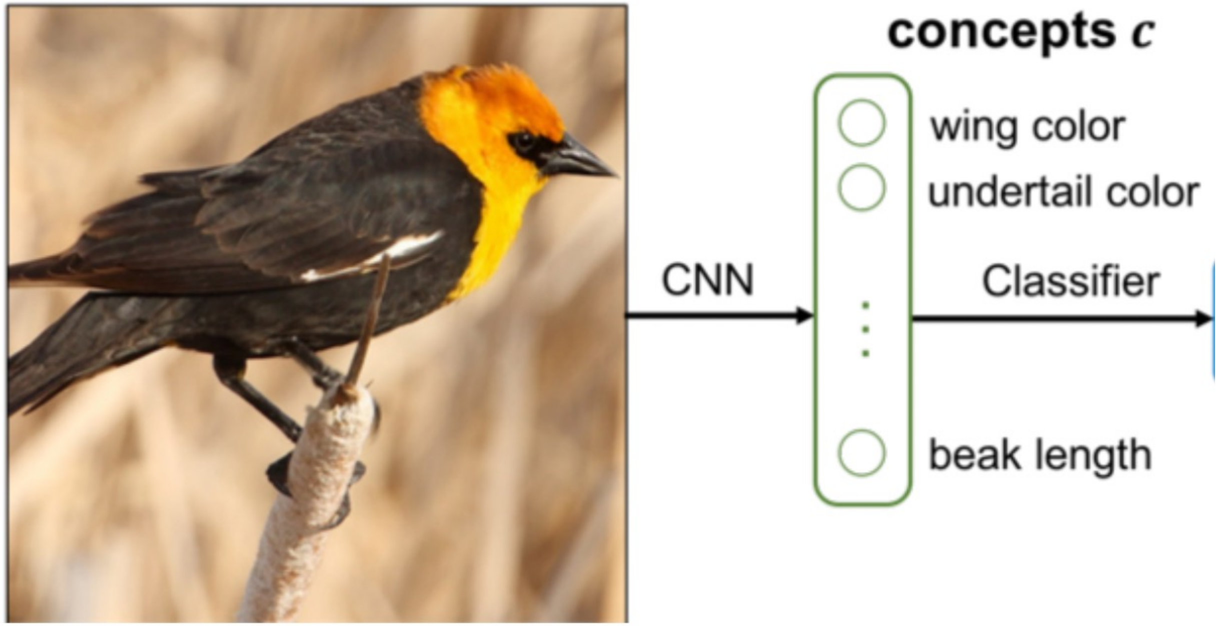
Evaluation: Interaction Interpretability

Methods	Models	IMDB	SST-1	SST-2	AG News	TREC	Subj
LSTM	WIGRAPH-NOA	88.53	45.70	83.96	**91.07**	91.00	90.30
	WIGRAPH-topA	**88.84**	**45.82**	**84.48**	90.91	**91.27**	**90.58**
BERT	WIGRAPH-NOA	85.62	51.31	**89.18**	**90.79**	96.40	95.90
	WIGRAPH-topA	**85.67**	**51.52**	89.02	90.47	**97.04**	**95.97**
RoBERTa	WIGRAPH-NOA	89.02	52.10	91.52	90.13	95.2	95.50
	WIGRAPH-topA	**90.02**	**53.84**	**92.51**	**91.50**	**96.00**	**96.20**
distilBERT	WIGRAPH-NOA	85.08	47.10	86.82	90.08	95.00	95.20
	WIGRAPH-topA	**86.32**	**47.25**	**87.00**	**90.10**	**96.40**	**96.00**

Average Post Hoc Interaction Occlusion Score when using top K scoring interactions for prediction

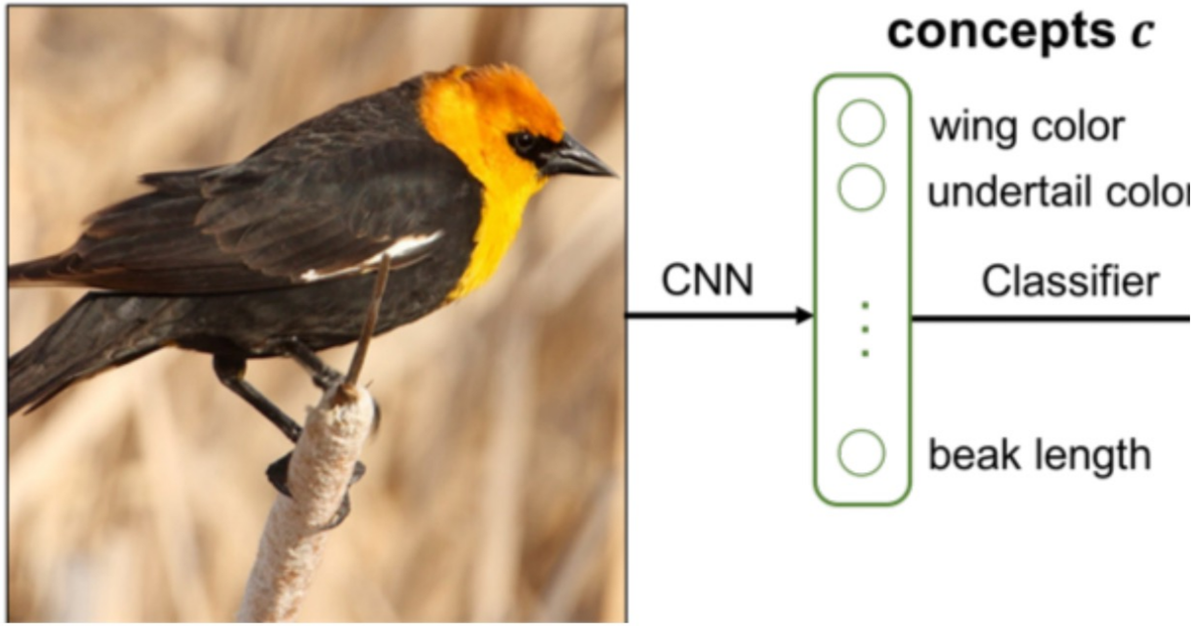
Here "WIGRAPH-NOA" uses an identity A

Extending WIGRAPH to Concept Interaction in Vision Tasks



- Concepts describe high level attributes of an image.
- Interactions between concepts can affect prediction.
- Concept interactions are unknown.

Extending WIGRAPH to Concept Interaction in Vision Tasks

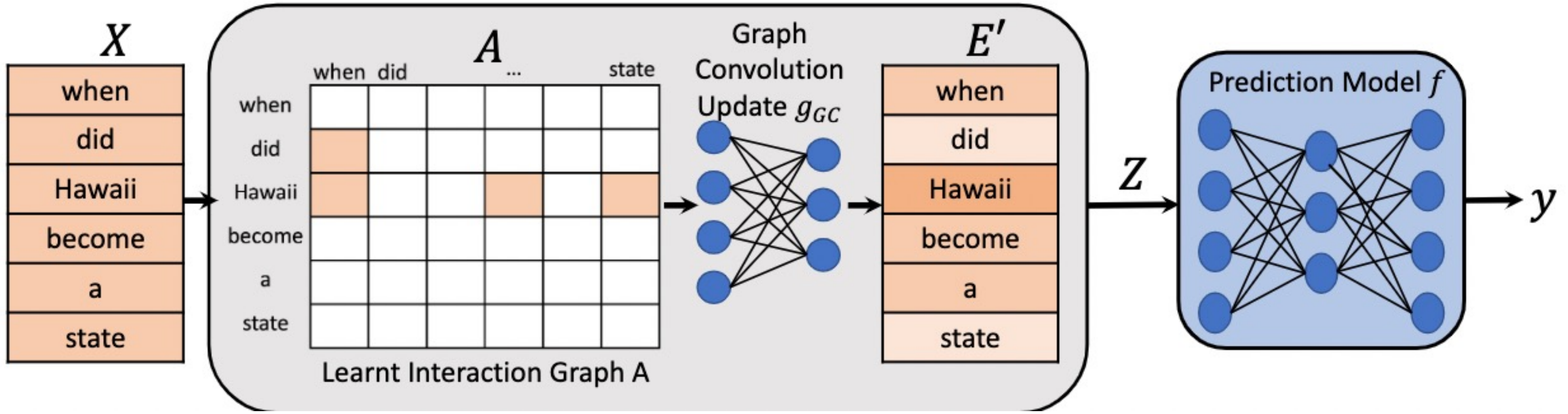


Model	Accuracy
Concept Bottleneck	89.41
Concept Bottleneck + WIGRAPH	95.74

- Concepts describe high level attributes of an image.
- Interactions between concepts can affect prediction.
- Concept interactions are unknown.

WIGRAPH Summary

a special layer in the form of discovering global word-word interactions



- Improve model's intrinsic interpretability
- Plug-and-Play Layer
- Model agnostic
- No loss of prediction performance

Thank You

Poster ID: 218

Possible Future Work

- 1. Higher-order interactions (not only pairwise)
- 2. More efficient parametrization in a scalable way
- 3. Extend to other tasks, like NLI, QA, Multi-modal, …