Towards Transparent and Explainable Attention Models

Akash Kumar Mohankumar, Preksha Nema, Sharan Narasimhan,
Mitesh M. Khapra, Balaji Vasan Srinivasan, Balaraman Ravindran

Motivation

- Attention mechanisms have become an indispensable component of modern-day neural networks e.g., LSTM-based models, Transformers.

Great service, atmosphere and food!

loved the gluten free options

food was terrible !

- Apart from providing improvements in predictive performance, they are often used to understand the internal workings of a model

Motivation

- Attention mechanisms have become an indispensable component of modern-day neural networks e.g., **LSTM-based models**, Transformers.

Great service, atmosphere and food!

loved the gluten free options

food was terrible !

- Apart from providing improvements in predictive performance, they are often used to understand the internal workings of a model

But, Does Attention Offer Interpretability?

- Recent works Serrano and Smith (2019)[1], Jain and Wallace (2019)[2] show that high attention weights need not necessarily correspond to a higher impact on the model's predictions.

- And hence attention distributions do not provide a *faithful* explanation for the model's predictions.

1. Sofia Serrano and Noah A. Smith. 2019. **Is attention interpretable?** In ACL
2. Sarthak Jain and Byron C. Wallace. 2019. **Attention is not explanation**. In NAACL-HLT

But, Does Attention Offer Interpretability?

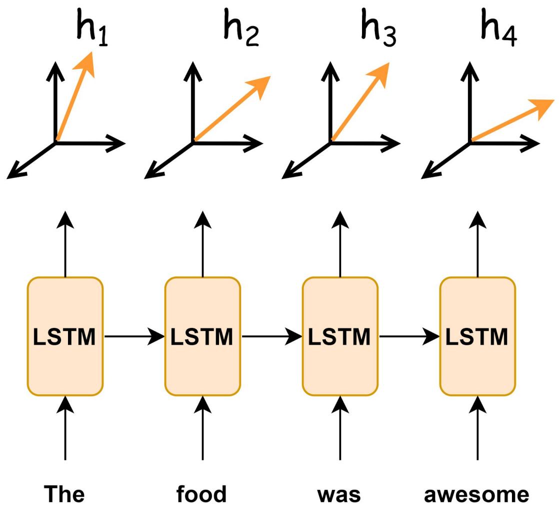
- Recent works Serrano and Smith (2019)[1], Jain and Wallace (2019)[2] show that high attention weights need not necessarily correspond to a higher impact on the model's predictions.

- And hence attention distributions do not provide a *faithful* explanation for the model's predictions.

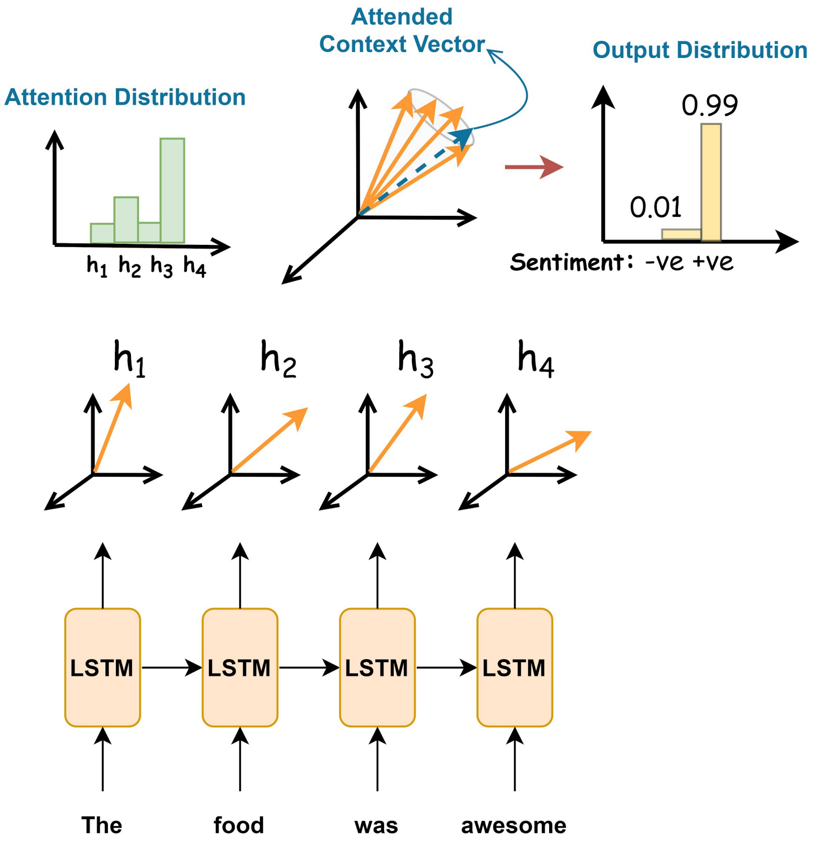
- On the other hand, Wiegreffe and Pinter (2019)[3] argue that there is still a possibility that attention distributions may provide a *plausible* explanation which can be understood by a human even if it is not faithful to how the model works.

1. Sofia Serrano and Noah A. Smith. 2019. **Is attention interpretable?** In ACL
2. Sarthak Jain and Byron C. Wallace. 2019. **Attention is not explanation**. In NAACL-HLT
3. Sarah Wiegreffe and Yuval Pinter. 2019. **Attention is not not explanation**. In EMNLP

Quick Recap: LSTM based model

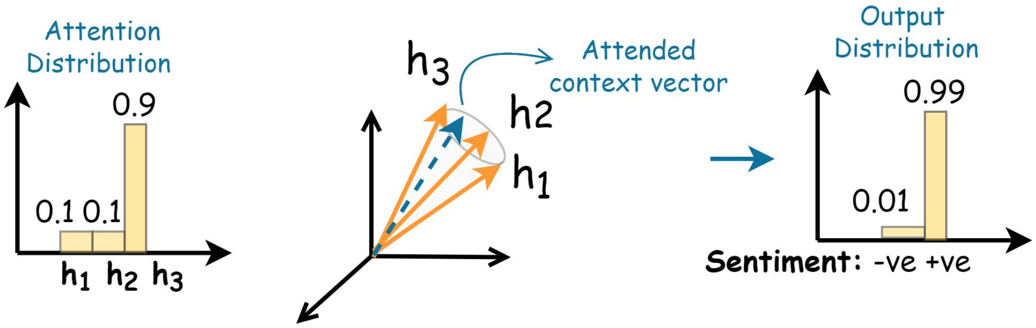


Quick Recap: LSTM based model



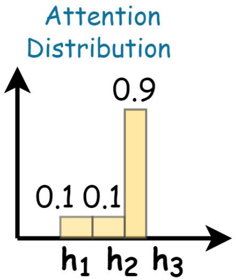
Do Attention distributions provide a faithful explanation?

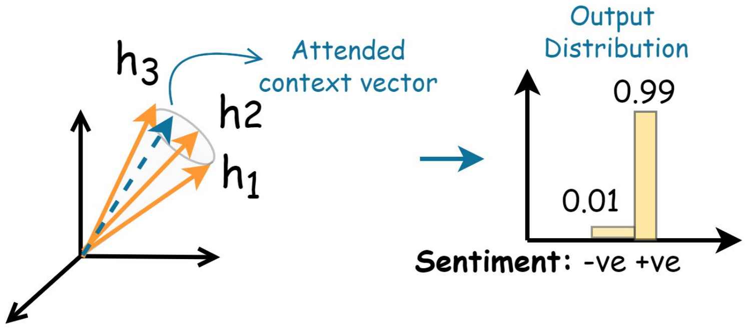
Case 1: High similarity in input representations

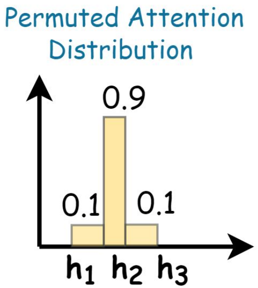


Do Attention distributions provide a faithful explanation?

Case 1: High similarity in input representations

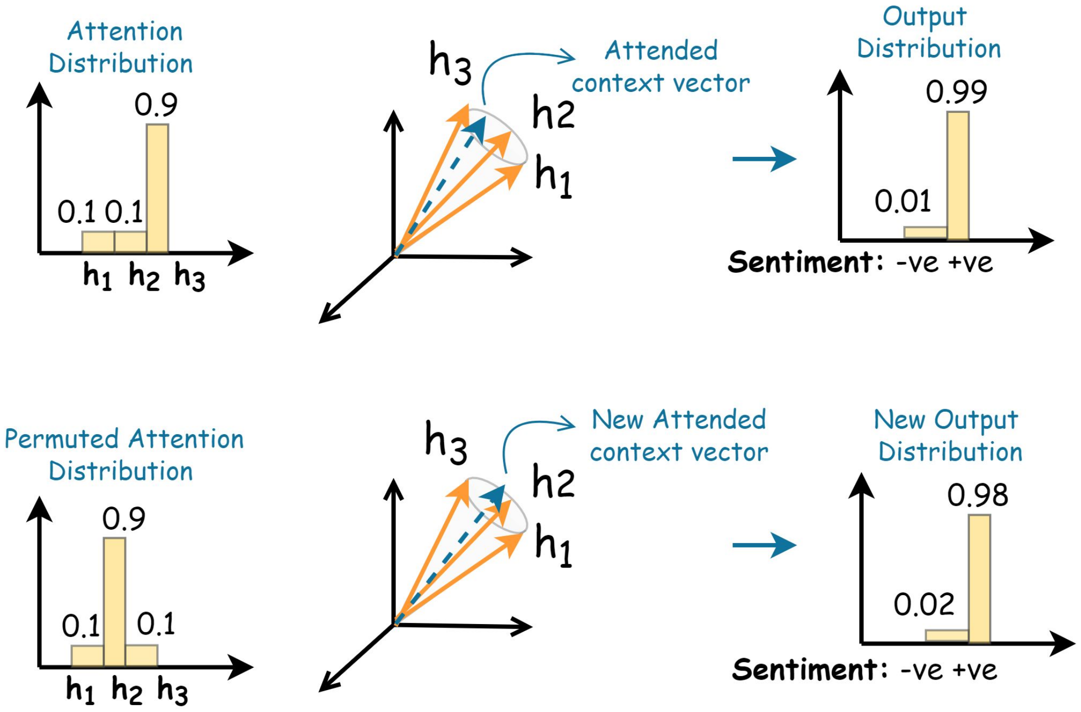






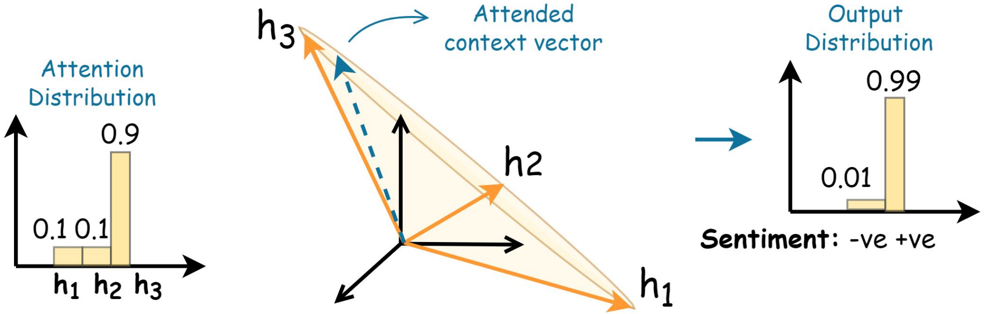
Do Attention distributions provide a faithful explanation?

Case 1: High similarity in input representations



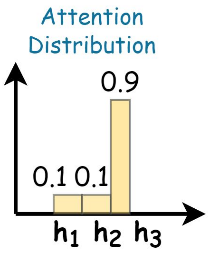
Do Attention distributions provide a faithful explanation?

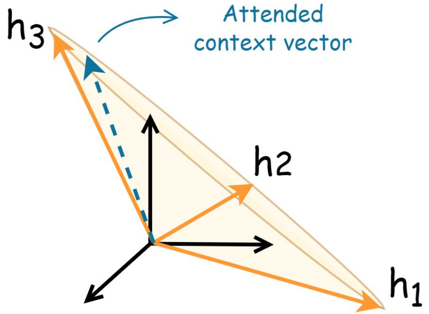
Case 2: Low similarity in input representations

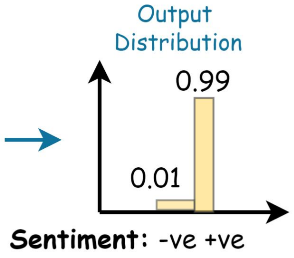


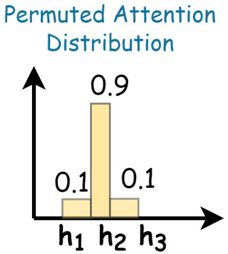
Do Attention distributions provide a faithful explanation?

Case 2: Low similarity in input representations



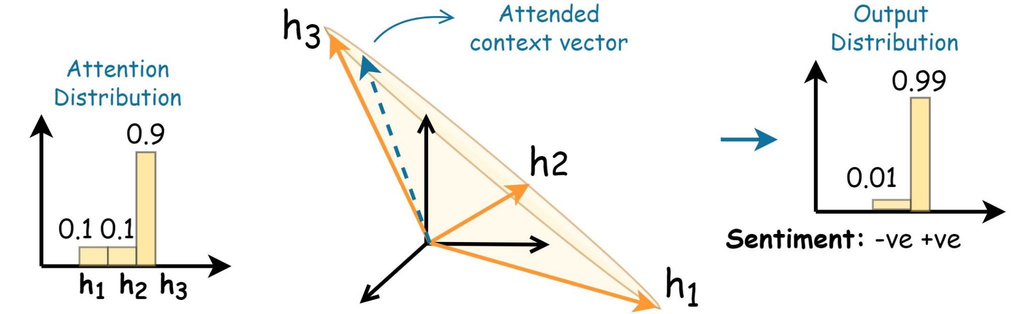


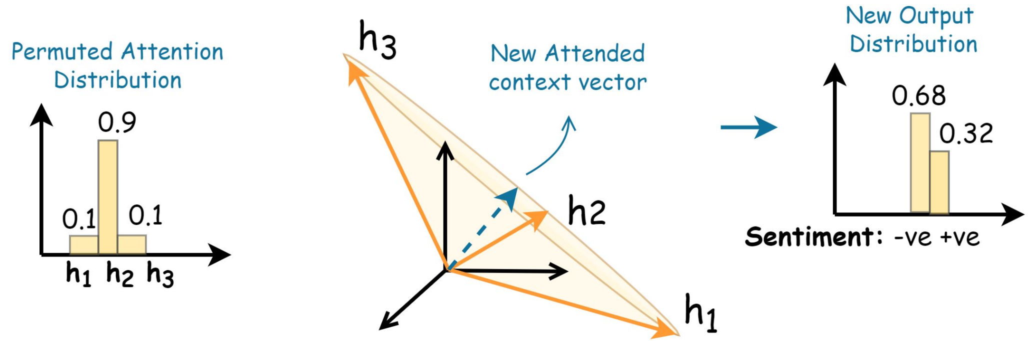




Do Attention distributions provide a faithful explanation?

Case 2: Low similarity in input representations



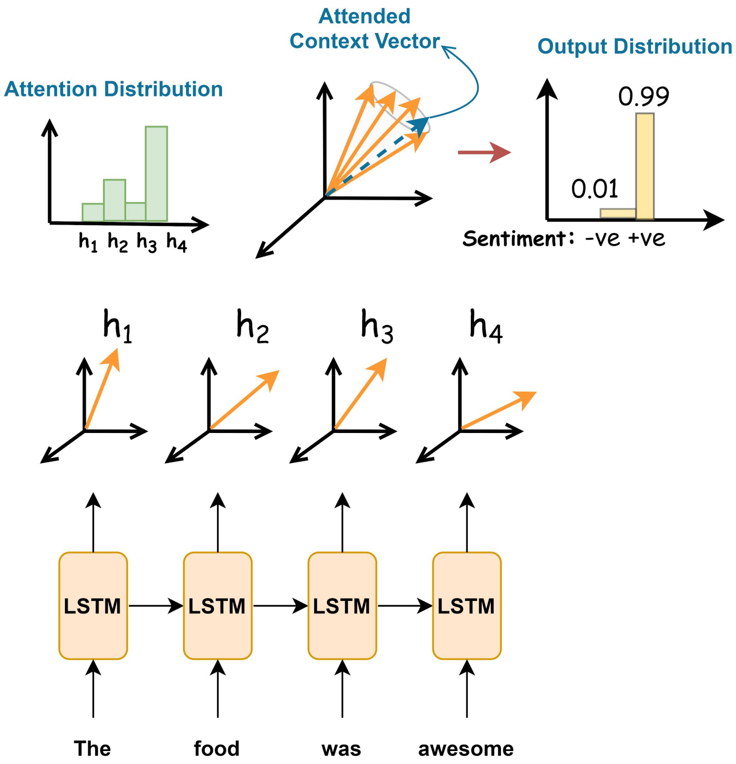


Do Attention distributions provide a faithful explanation?

Not always: When the input representations over which an attention distribution is being computed are very similar to each other, attention weights are not very meaningful

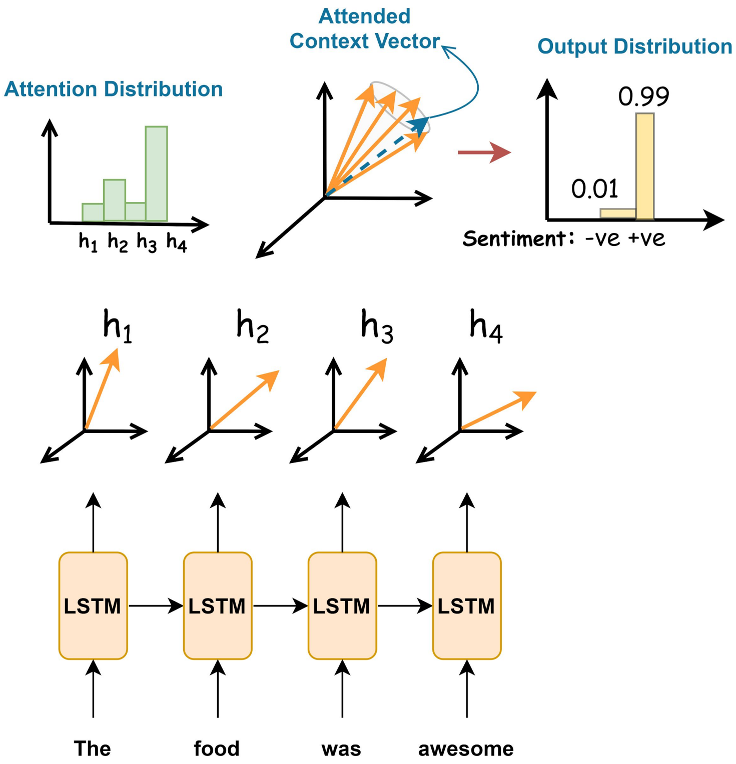
Looking closely at an LSTM based model

Are the hidden representations computed by LSTM very similar or very different?



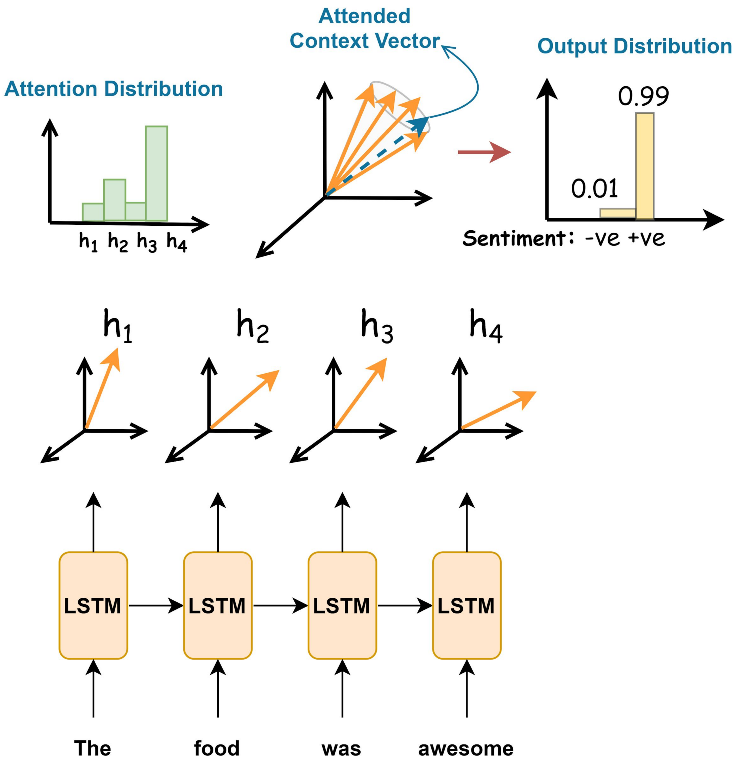
Looking closely at an LSTM based model

How do we quantify the similarity between these vectors?



Looking closely at an LSTM based model

How do we quantify the similarity between these vectors?



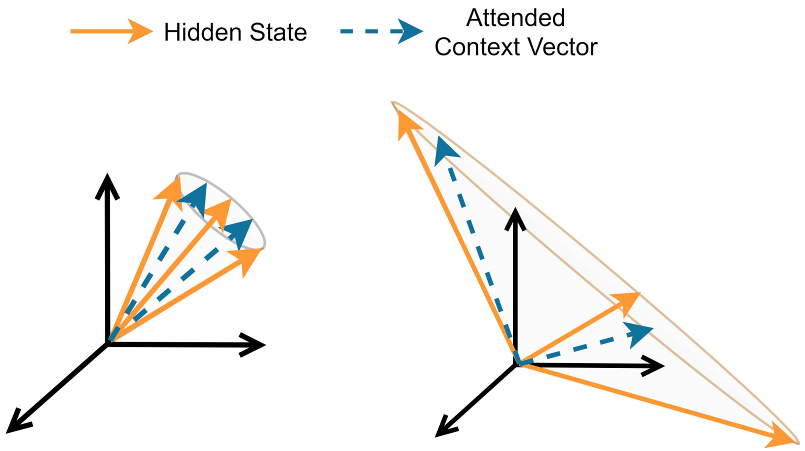
We measure the similarity between a set of vectors, $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_m\}$ using the conicity measure

$$\mathrm{ATM}(\mathbf{h}_i, \mathbf{H}) = \mathrm{cosine}(\mathbf{h}_i, \frac{1}{m}\sum_{j=1}^{m}\mathbf{h}_j)$$

$$\mathrm{conicity}(\mathbf{H}) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{ATM}(\mathbf{h}_i, \mathbf{H})$$

Looking closely at an LSTM based model

How do we quantify the similarity between these vectors?



High conicity
(Permuting attention weight will have minimum impact)

Low conicity
(Permuting attention weight will have significant impact)

We measure the similarity between a set of vectors, $\mathbf{H} = \{\mathbf{h}_1, \dots, \mathbf{h}_m\}$ using the conicity measure

$$\mathrm{ATM}(\mathbf{h}_i, \mathbf{H}) = \mathrm{cosine}(\mathbf{h}_i, \frac{1}{m}\sum_{j=1}^{m}\mathbf{h}_j)$$

$$\mathrm{conicity}(\mathbf{H}) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{ATM}(\mathbf{h}_i, \mathbf{H})$$

Conicity of LSTM Hidden states

Dataset	LSTM Accuracy	LSTM Conicity	Random Conicity
Text Classification			
SST	81.79	0.68	0.25
IMDB	89.49	0.69	0.08
YELP	95.60	0.53	0.14
Amazon	93.73	0.50	0.13
Anemia	88.54	0.46	0.02
Diabetes	92.31	0.61	0.02
20News	93.55	0.77	0.13
Tweets	87.02	0.77	0.24

Dataset	LSTM Accuracy	LSTM Conicity	Random Conicity
Natural Language Inference			
SNLI	78.23	0.56	0.27
Paraphrase Detection			
QQP	78.74	0.59	0.30
Question Answering			
Babi 1	99.10	0.56	0.19
Babi 2	40.10	0.48	0.12
Babi 3	47.70	0.43	0.07
CNN	63.07	0.45	0.04

Accuracy and conicity of the Vanilla LSTM model across different datasets. Conicity of vectors uniformly distributed with respect to direction is also reported for reference

Conicity of LSTM Hidden states

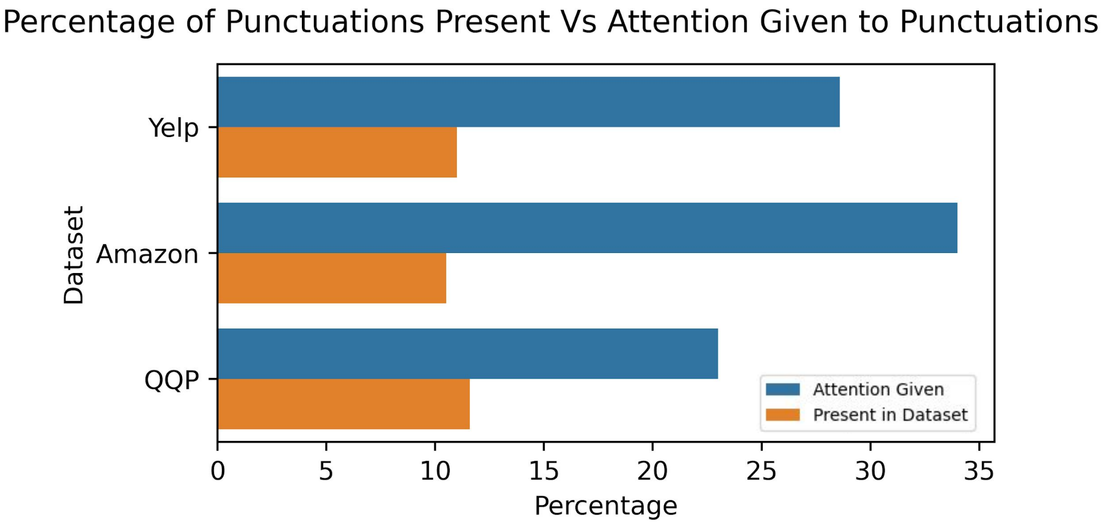
Key Insight: The LSTM representations have a high conicity, hence the learned attention distributions would not provide a faithful explanation

Dataset	LSTM Accuracy	LSTM Conicity	Random Conicity
Text Classification			
SST	81.79	0.68	0.25
IMDB	89.49	0.69	0.08
YELP	95.60	0.53	0.14
Amazon	93.73	0.50	0.13
Anemia	88.54	0.46	0.02
Diabetes	92.31	0.61	0.02
20News	93.55	0.77	0.13
Tweets	87.02	0.77	0.24

Dataset	LSTM Accuracy	LSTM Conicity	Random Conicity
Natural Language Inference			
SNLI	78.23	0.56	0.27
Paraphrase Detection			
QQP	78.74	0.59	0.30
Question Answering			
Babi 1	99.10	0.56	0.19
Babi 2	40.10	0.48	0.12
Babi 3	47.70	0.43	0.07
CNN	63.07	0.45	0.04

Accuracy and conicity of the Vanilla LSTM model across different datasets. Conicity of vectors uniformly distributed with respect to direction is also reported for reference

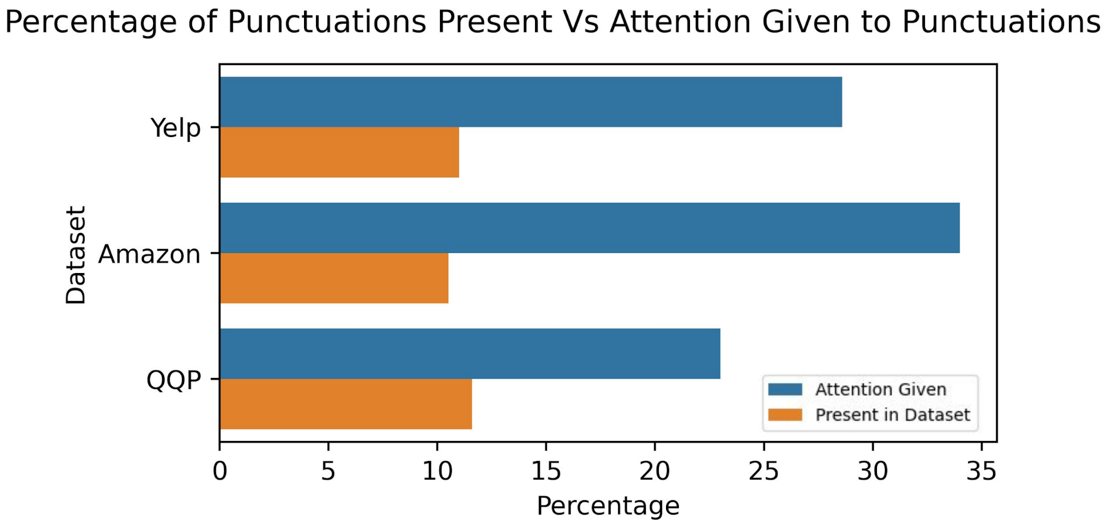
Do attention distributions provide a plausible explanation?



Percentage of total punctuation tokens present in the dataset vs percentage of total attention given to punctuation tokens by a vanilla LSTM model

Do attention distributions provide a plausible explanation?

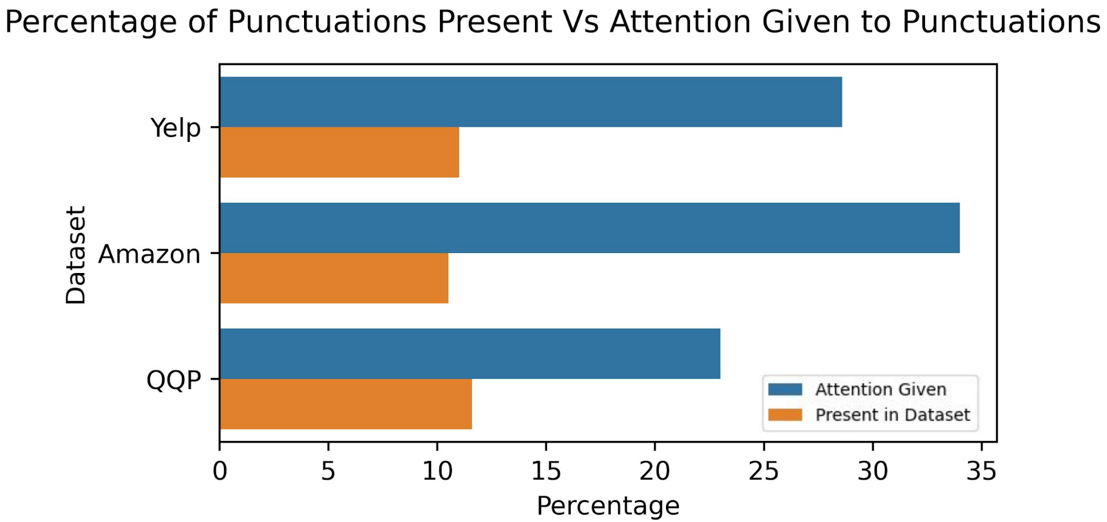
Key Insight: With significantly high attention given to punctuations, it is very doubtful whether attention distributions will provide any reasonable explanations.



Percentage of total punctuation tokens present in the dataset vs percentage of total attention given to punctuation tokens by a vanilla LSTM model

Do attention distributions provide a plausible explanation?

Key Insight: With significantly high attention given to punctuations, it is very doubtful whether attention distributions will provide any reasonable explanations.



Possible reason: Hidden states might capture a summary of the entire context instead of being specific to their corresponding words as suggested by the high conicity.

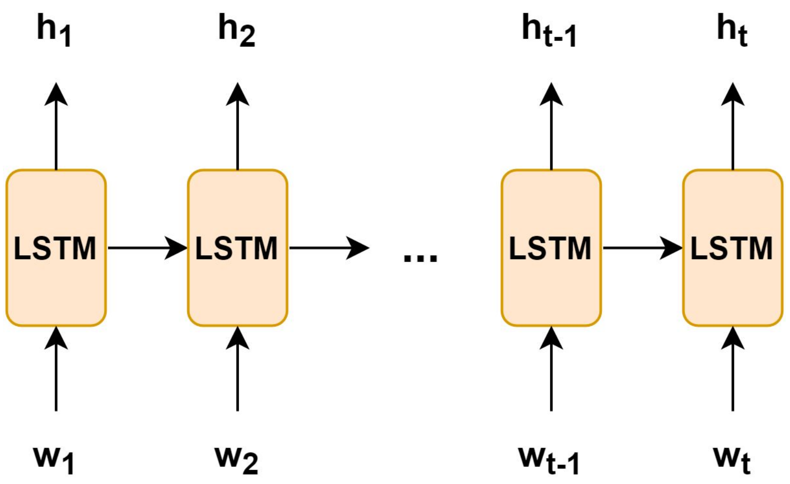
Percentage of total punctuation tokens present in the dataset vs percentage of total attention given to punctuation tokens by a vanilla LSTM model

Our Main Goal

- **Goal:** Design a model where the attention distributions provide faithful and plausible explanations
- We have observed that high conicity in hidden states can affect the transparency and explainability of attention models
- We propose two methods to promote diversity in the hidden states

Proposed Model

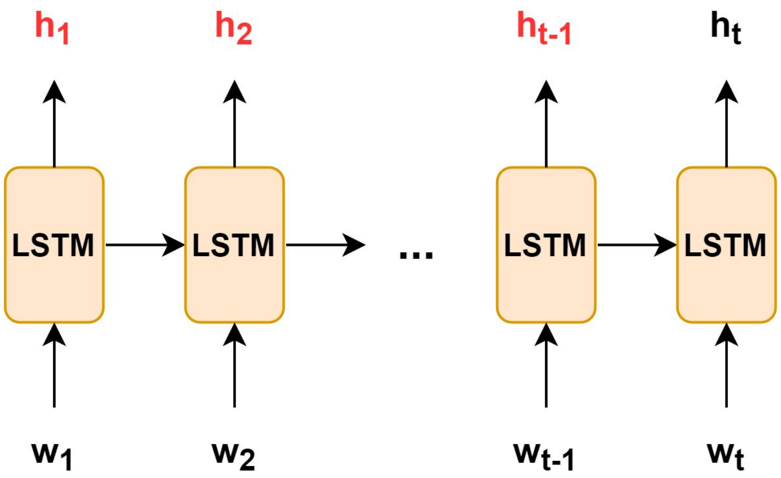
Method 1: Orthogonalization



Proposed Model

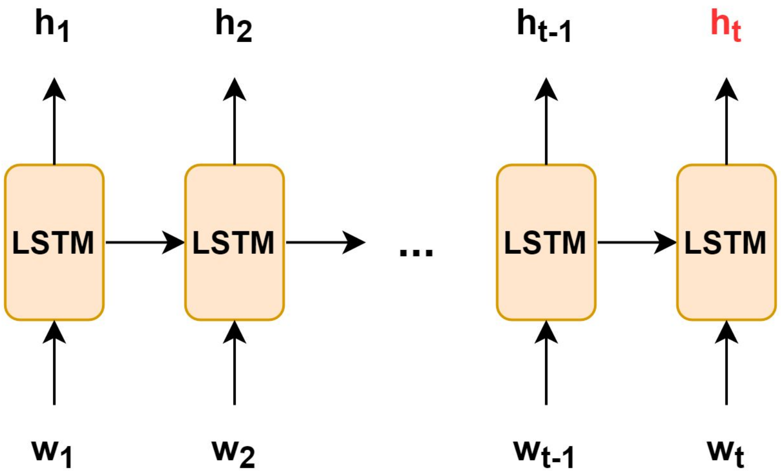
Method 1: Orthogonalization

$$\overline{\mathbf{h}}_t = \sum_{i=1}^{t-1} \mathbf{h}_i$$



Proposed Model

Method 1: Orthogonalization



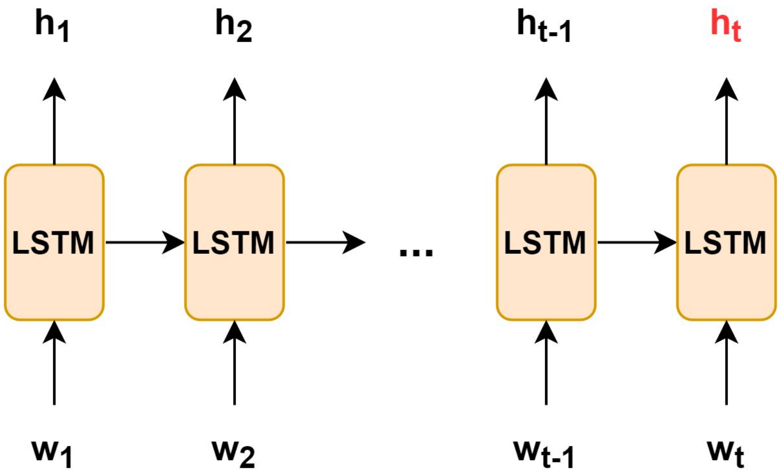
$$\overline{\mathbf{h}}_t = \sum_{i=1}^{t-1} \mathbf{h}_i$$

$$\hat{\mathbf{h}}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

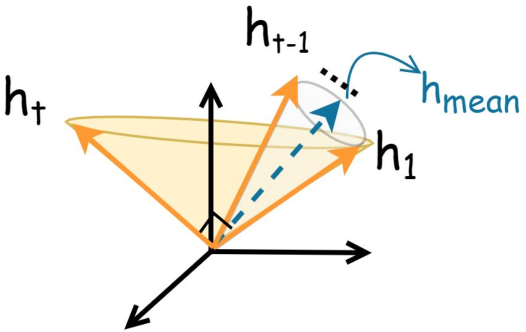
$$\mathbf{h_t} = \hat{\mathbf{h}}_t - \frac{\hat{\mathbf{h}}_t^T \overline{\mathbf{h}}_t}{\overline{\mathbf{h}}_t^T \overline{\mathbf{h}}_t} \overline{\mathbf{h}}_t$$

Proposed Model

Method 1: Orthogonalization



Orthogonal LSTM



Hidden state h_t is orthogonal to the mean of h_1, h_2, …, h_{t-1}

Proposed Model

Method 2: Diversity Driven Training

- The previous method imposes a hard orthogonality constraint between the hidden states and the previous states' mean.
- We also propose a more flexible approach where the model is jointly trained to minimize the cross entropy loss and the conicity of hidden states.

$$\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_m\}$$

$$L(\theta) = \text{Cross-Entropy}(y, \hat{y}|\theta) + \lambda \ \text{conicity}(\mathbf{H}|\theta)$$

- We call an vanilla LSTM model trained with this diversity objective as the Diversity LSTM

Empirical Evaluations: Accuracy & Conicity

Dataset	LSTM		Diversity LSTM		Orthogonal LSTM		Random	MLP
	Accuracy	Conicity	Accuracy	Conicity	Accuracy	Conicity	Conicity	Accuracy
Binary Classification								
SST	**81.79**	0.68	79.95	0.20	80.05	0.28	0.25	80.05
IMDB	**89.49**	0.69	88.54	0.08	88.71	0.18	0.08	88.29
Yelp	95.60	0.53	95.40	0.06	**96.00**	0.18	0.14	92.85
Amazon	**93.73**	0.50	92.90	0.05	93.04	0.16	0.13	87.88
Anemia	88.54	0.46	90.09	0.09	**90.17**	0.12	0.02	88.27
Diabetes	**92.31**	0.61	91.99	0.08	87.05	0.12	0.02	85.39
20News	**93.55**	0.77	91.03	0.15	92.15	0.23	0.13	87.68
Tweets	87.02	0.77	**87.04**	0.24	83.20	0.27	0.24	80.60
Natural Language Inference								
SNLI	**78.23**	0.56	76.96	0.12	76.46	0.27	0.27	75.35
Paraphrase Detection								
QQP	**78.74**	0.59	78.40	0.04	78.61	0.33	0.30	77.78
Question Answering								
bAbI 1	99.10	0.56	**100.00**	0.07	99.90	0.22	0.19	42.00
bAbI 2	40.10	0.48	40.20	0.05	**56.10**	0.21	0.12	33.20
bAbI 3	47.70	0.43	50.90	0.10	**51.20**	0.12	0.07	31.60
CNN	**63.07**	0.45	58.19	0.06	54.30	0.07	0.04	37.40

Accuracy and conicity of Vanilla, Diversity and Orthogonal LSTM across different datasets. Accuracy of a Multilayered Perceptron (MLP) model and conicity of vectors uniformly distributed with respect to direction is also reported for reference

Empirical Evaluations: Accuracy & Conicity

Dataset	LSTM		Diversity LSTM		Orthogonal LSTM		Random	MLP
	Accuracy	Conicity	Accuracy	Conicity	Accuracy	Conicity	Conicity	Accuracy
Binary Classification								
SST	**81.79**	0.68	79.95	0.20	80.05	0.28	0.25	80.05
IMDB	**89.49**	0.69	88.54	0.08	88.71	0.18	0.08	88.29
Yelp	95.60	0.53	95.40	0.06	**96.00**	0.18	0.14	92.85
Amazon	**93.73**	0.50	92.90	0.05	93.04	0.16	0.13	87.88
Anemia	88.54	0.46	90.09	0.09	**90.17**	0.12	0.02	88.27
Diabetes	**92.31**	0.61	91.99	0.08	87.05	0.12	0.02	85.39
20News	**93.55**	0.77	91.03	0.15	92.15	0.23	0.13	87.68
Tweets	87.02	0.77	**87.04**	0.24	83.20	0.27	0.24	80.60
Natural Language Inference								
SNLI	**78.23**	0.56	76.96	0.12	76.46	0.27	0.27	75.35
Paraphrase Detection								
QQP	**78.74**	0.59	78.40	0.04	78.61	0.33	0.30	77.78
Question Answering								
bAbI 1	99.10	0.56	**100.00**	0.07	99.90	0.22	0.19	42.00
bAbI 2	40.10	0.48	40.20	0.05	**56.10**	0.21	0.12	33.20
bAbI 3	47.70	0.43	50.90	0.10	**51.20**	0.12	0.07	31.60
CNN	**63.07**	0.45	58.19	0.06	54.30	0.07	0.04	37.40

Conicity of our proposed models are much lower with comparable predictive performance

Accuracy and conicity of Vanilla, Diversity and Orthogonal LSTM across different datasets. Accuracy of a Multilayered Perceptron (MLP) model and conicity of vectors uniformly distributed with respect to direction is also reported for reference

Qualitative Examples

Question 1: What is the best way to improve my spoken English soon ?
Question 2: How can I improve my English speaking ability ?
Is paraphrase (Actual & Predicted): Yes

Attention Distribution:

Vanilla LSTM	How can I improve my English speaking ability ?
Diversity LSTM	How can I improve my English speaking ability ?

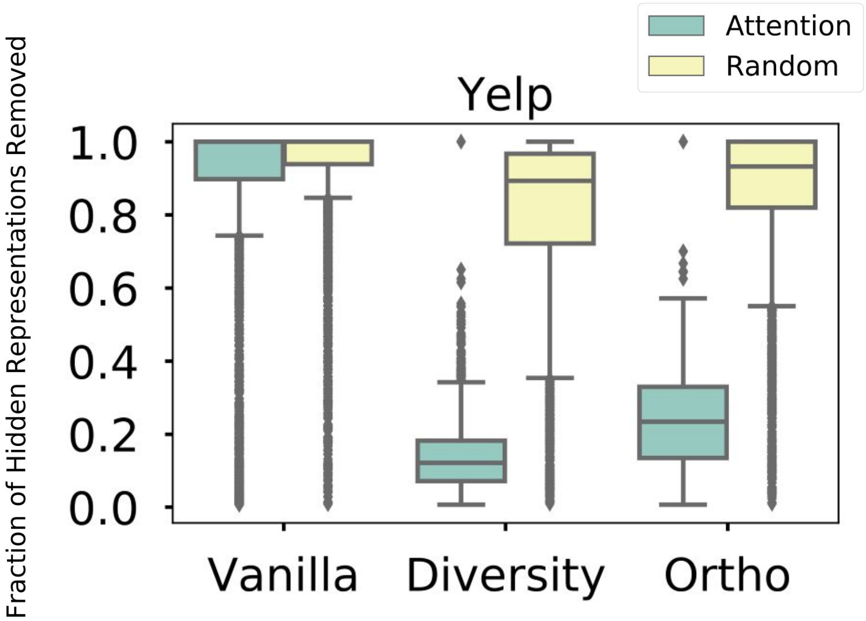
Passage: Sandra went to the garden . Daniel went to the garden.
Question: Where is Sandra?
Answer (Actual & Predicted): garden

Attention Distribution:

Vanilla LSTM	Sandra went to the garden . Daniel went to the garden
Diversity LSTM	Sandra went to the garden . Daniel went to the garden

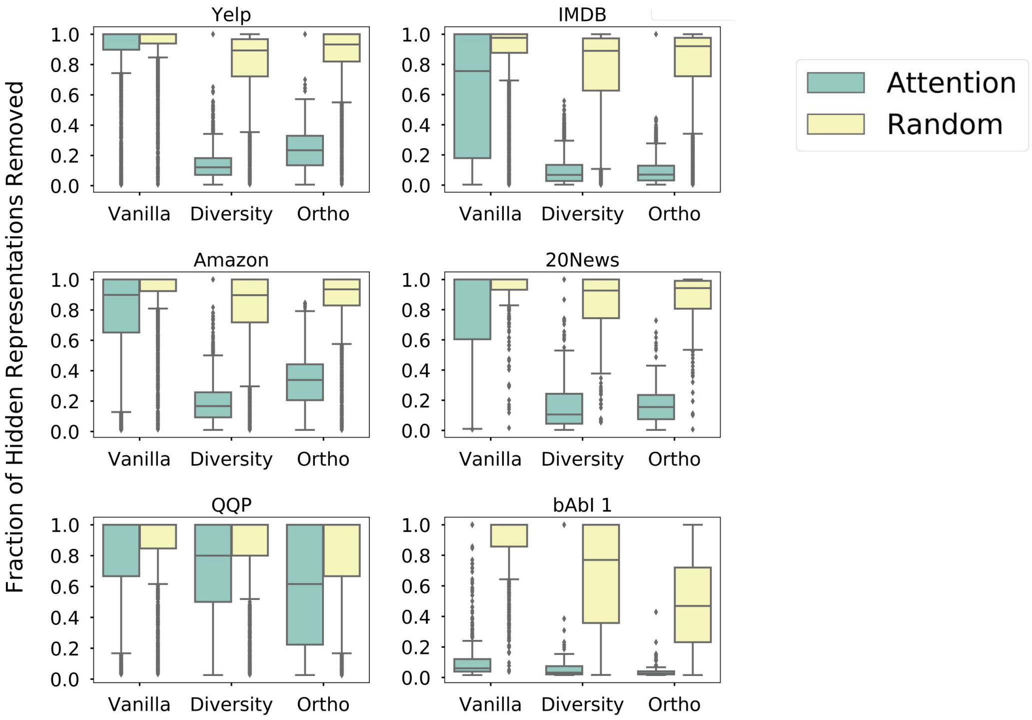
Samples of attention distribution from Vanilla and Diversity LSTM models on the Quora Question Paraphrase (QQP) and bAbi 1 datasets

Importance of Hidden Representations



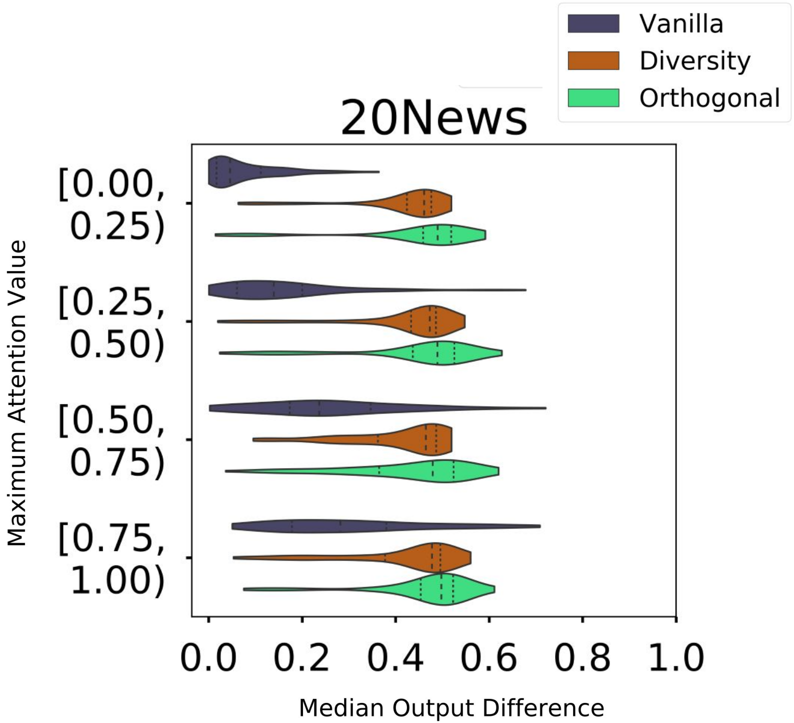
Box plots of the fraction of hidden representations erased for a decision flip when following the ranking provided by attention weights and a random ranking on the Yelp dataset. Models are mentioned at the of figure. Blue and Yellow indicate the attention and random ranking

Importance of Hidden Representations



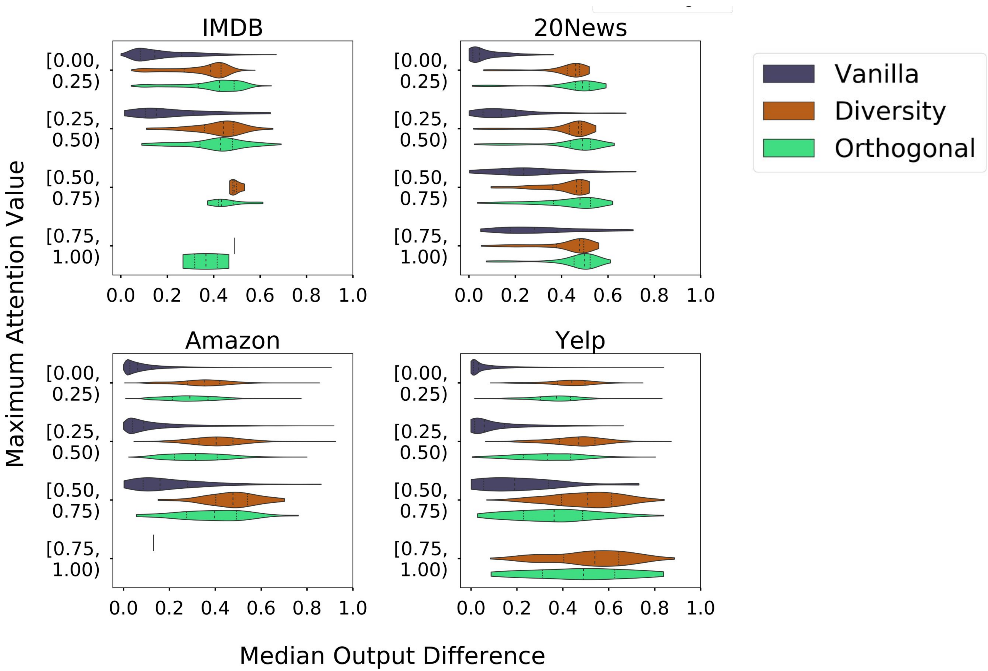
Box plots of fraction of hidden representations removed for a decision flip. Dataset and models are mentioned at the top and bottom of figures. Blue and Yellow indicate the attention and random ranking

Permuting Attention



Comparison of Median output difference on randomly permuting the attention weights in the vanilla, Diversity and Orthogonal LSTM models for the 20News dataset

Permuting Attention



Comparison of Median output difference on randomly permuting the attention weights in the vanilla, Diversity and Orthogonal LSTM models. The Dataset names are mentioned at the top of each figure. Colors indicate the different models as shown legend

Comparison with Rationales

- We analyze how much attention is given to words in the sentence that are important for the prediction
- Specifically, we find the minimum subset of words in the input sentence with which the model can accurately make predictions, which are also known as rationales.
- An extractive rationale generator is trained using the REINFORCE algorithm to maximize the foll

$$R = \log p_{model}(y|\mathbf{Z}) - \alpha||\mathbf{Z}||$$

where y is the ground truth class, Z is the extracted rationale, |Z| represents the length of the rationale.

Comparison with Rationales

Dataset	Vanilla LSTM		Diversity LSTM	
	Rationale Attention	Rationale Length	Rationale Attention	Rationale Length
SST	0.348	0.240	0.624	0.175
IMDB	0.472	0.217	0.761	0.169
Yelp	0.438	0.173	0.574	0.160
Amazon	0.346	0.162	0.396	0.240
Anemia	0.611	0.192	0.739	0.237
Diabetes	0.742	0.458	0.825	0.354
20News	0.627	0.215	0.884	0.173
Tweets	0.284	0.225	0.764	0.306

Mean Attention given to the generated rationales with their mean lengths (in fraction)

Comparison with attribution methods

Dataset	Pearson Correlation ↑				JS Divergence ↓			
	Gradients (Mean ± Std.)		Integrated Gradients (Mean ± Std.)		Gradients (Mean ± Std.)		Integrated Gradients (Mean ± Std.)	
	Vanilla	Diversity	Vanilla	Diversity	Vanilla	Diversity	Vanilla	Diversity
Text Classification								
SST	0.71 ± 0.21	0.83 ± 0.19	0.62 ± 0.24	0.79 ± 0.22	0.10 ± 0.04	0.08 ± 0.05	0.12 ± 0.05	0.09 ± 0.05
IMDB	0.80 ± 0.07	0.89 ± 0.04	0.68 ± 0.09	0.78 ± 0.07	0.09 ± 0.02	0.09 ± 0.01	0.13 ± 0.02	0.13 ± 0.02
Yelp	0.55 ± 0.16	0.79 ± 0.12	0.40 ± 0.19	0.79 ± 0.14	0.15 ± 0.04	0.13 ± 0.04	0.19 ± 0.05	0.19 ± 0.05
Amazon	0.43 ± 0.19	0.77 ± 0.14	0.43 ± 0.19	0.77 ± 0.14	0.17 ± 0.04	0.12 ± 0.04	0.21 ± 0.06	0.12 ± 0.04
Anemia	0.63 ± 0.12	0.72 ± 0.10	0.43 ± 0.15	0.66 ± 0.11	0.20 ± 0.04	0.19 ± 0.03	0.34 ± 0.05	0.23 ± 0.04
Diabetes	0.65 ± 0.15	0.76 ± 0.13	0.55 ± 0.14	0.69 ± 0.18	0.26 ± 0.05	0.20 ± 0.04	0.36 ± 0.04	0.24 ± 0.06
20News	0.72 ± 0.28	0.96 ± 0.08	0.65 ± 0.32	0.67 ± 0.11	0.15 ± 0.07	0.06 ± 0.04	0.21 ± 0.06	0.07 ± 0.05
Tweets	0.65 ± 0.24	0.80 ± 0.21	0.56 ± 0.25	0.74 ± 0.22	0.08 ± 0.03	0.12 ± 0.07	0.08 ± 0.04	0.15 ± 0.06
Natural Language Inference								
SNLI	0.58 ± 0.33	0.51 ± 0.35	0.38 ± 0.40	0.26 ± 0.39	0.11 ± 0.07	0.10 ± 0.06	0.16 ± 0.09	0.13 ± 0.06
Paraphrase Detection								
QQP	0.19 ± 0.34	0.58 ± 0.31	-0.06 ± 0.34	0.21 ± 0.36	0.15 ± 0.08	0.10 ± 0.05	0.19 ± 0.10	0.15 ± 0.06
Question Answering								
Babi 1	0.56 ± 0.34	0.91 ± 0.10	0.33 ± 0.37	0.91 ± 0.10	0.33 ± 0.12	0.21 ± 0.08	0.43 ± 0.13	0.24 ± 0.08
Babi 2	0.16 ± 0.23	0.70 ± 0.13	0.05 ± 0.22	0.75 ± 0.10	0.53 ± 0.09	0.23 ± 0.06	0.58 ± 0.09	0.19 ± 0.05
Babi 3	0.39 ± 0.24	0.67 ± 0.19	-0.01 ± 0.08	0.47 ± 0.25	0.46 ± 0.08	0.37 ± 0.07	0.64 ± 0.05	0.41 ± 0.08
CNN	0.58 ± 0.25	0.75 ± 0.20	0.45 ± 0.28	0.66 ± 0.23	0.22 ± 0.07	0.17 ± 0.08	0.30 ± 0.10	0.21 ± 0.10

Mean and standard deviation of Pearson correlation and Jensen–Shannon divergence between Attention weights and Gradients/Integrated Gradients in Vanilla and Diversity LSTM models

Comparison with attribution methods

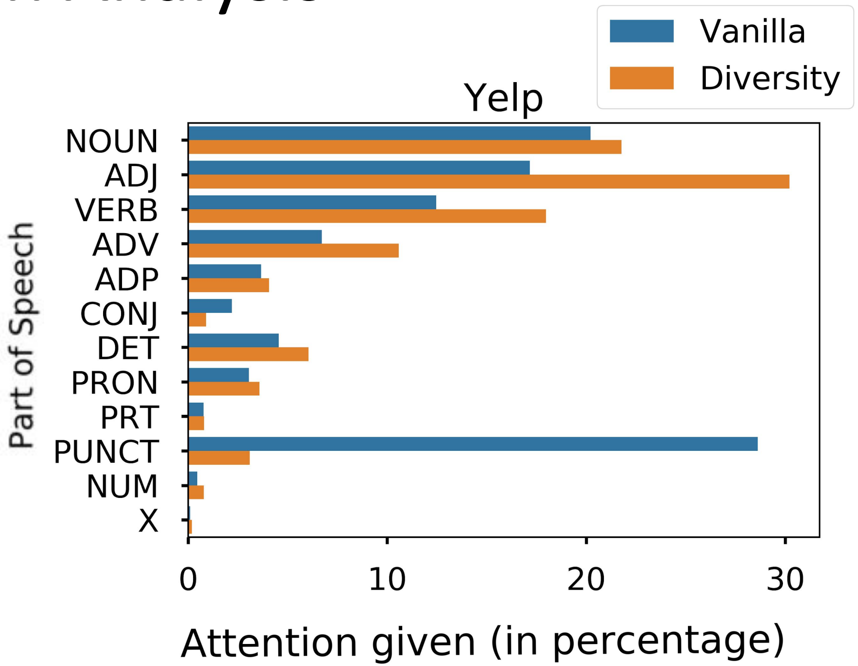
Dataset	Pearson Correlation ↑ Gradients (Mean ± Std.) Vanilla	Pearson Correlation ↑ Gradients (Mean ± Std.) Diversity	Pearson Correlation ↑ Integrated Gradients (Mean ± Std.) Vanilla	Pearson Correlation ↑ Integrated Gradients (Mean ± Std.) Diversity	JS Divergence ↓ Gradients (Mean ± Std.) Vanilla	JS Divergence ↓ Gradients (Mean ± Std.) Diversity	JS Divergence ↓ Integrated Gradients (Mean ± Std.) Vanilla	JS Divergence ↓ Integrated Gradients (Mean ± Std.) Diversity
Text Classification								
SST	0.71 ± 0.21	0.83 ± 0.19	0.62 ± 0.24	0.79 ± 0.22	0.10 ± 0.04	0.08 ± 0.05	0.12 ± 0.05	0.09 ± 0.05
IMDB	0.80 ± 0.07	0.89 ± 0.04	0.68 ± 0.09	0.78 ± 0.07	0.09 ± 0.02	0.09 ± 0.01	0.13 ± 0.02	0.13 ± 0.02
Yelp	0.55 ± 0.16	0.79 ± 0.12	0.40 ± 0.19	0.79 ± 0.14	0.15 ± 0.04	0.13 ± 0.04	0.19 ± 0.05	0.19 ± 0.05
Amazon	0.43 ± 0.19	0.77 ± 0.14	0.43 ± 0.19	0.77 ± 0.14	0.17 ± 0.04	0.12 ± 0.04	0.21 ± 0.06	0.12 ± 0.04
Anemia	0.63 ± 0.12	0.72 ± 0.10	0.43 ± 0.15	0.66 ± 0.11	0.20 ± 0.04	0.19 ± 0.03	0.34 ± 0.05	0.23 ± 0.04
Diabetes	0.65 ± 0.15	0.76 ± 0.13	0.55 ± 0.14	0.69 ± 0.18	0.26 ± 0.05	0.20 ± 0.04	0.36 ± 0.04	0.24 ± 0.06
20News	0.72 ± 0.28	0.96 ± 0.08	0.65 ± 0.32	0.67 ± 0.11	0.15 ± 0.07	0.06 ± 0.04	0.21 ± 0.06	0.07 ± 0.05
Tweets	0.65 ± 0.24	0.80 ± 0.21	0.56 ± 0.25	0.74 ± 0.22	0.08 ± 0.03	0.12 ± 0.07	0.08 ± 0.04	0.15 ± 0.06
Natural Language Inference								
SNLI	0.58 ± 0.33	0.51 ± 0.35	0.38 ± 0.40	0.26 ± 0.39	0.11 ± 0.07	0.10 ± 0.06	0.16 ± 0.09	0.13 ± 0.06
Paraphrase Detection								
QQP	0.19 ± 0.34	0.58 ± 0.31	-0.06 ± 0.34	0.21 ± 0.36	0.15 ± 0.08	0.10 ± 0.05	0.19 ± 0.10	0.15 ± 0.06
Question Answering								
Babi 1	0.56 ± 0.34	0.91 ± 0.10	0.33 ± 0.37	0.91 ± 0.10	0.33 ± 0.12	0.21 ± 0.08	0.43 ± 0.13	0.24 ± 0.08
Babi 2	0.16 ± 0.23	0.70 ± 0.13	0.05 ± 0.22	0.75 ± 0.10	0.53 ± 0.09	0.23 ± 0.06	0.58 ± 0.09	0.19 ± 0.05
Babi 3	0.39 ± 0.24	0.67 ± 0.19	-0.01 ± 0.08	0.47 ± 0.25	0.46 ± 0.08	0.37 ± 0.07	0.64 ± 0.05	0.41 ± 0.08
CNN	0.58 ± 0.25	0.75 ± 0.20	0.45 ± 0.28	0.66 ± 0.23	0.22 ± 0.07	0.17 ± 0.08	0.30 ± 0.10	0.21 ± 0.10

Average increase of 64.84% pearson correlation with gradients

Average decrease of 17.18% in JS divergence with gradients

Mean and standard deviation of Pearson correlation and Jensen–Shannon divergence between Attention weights and Gradients/Integrated Gradients in Vanilla and Diversity LSTM models

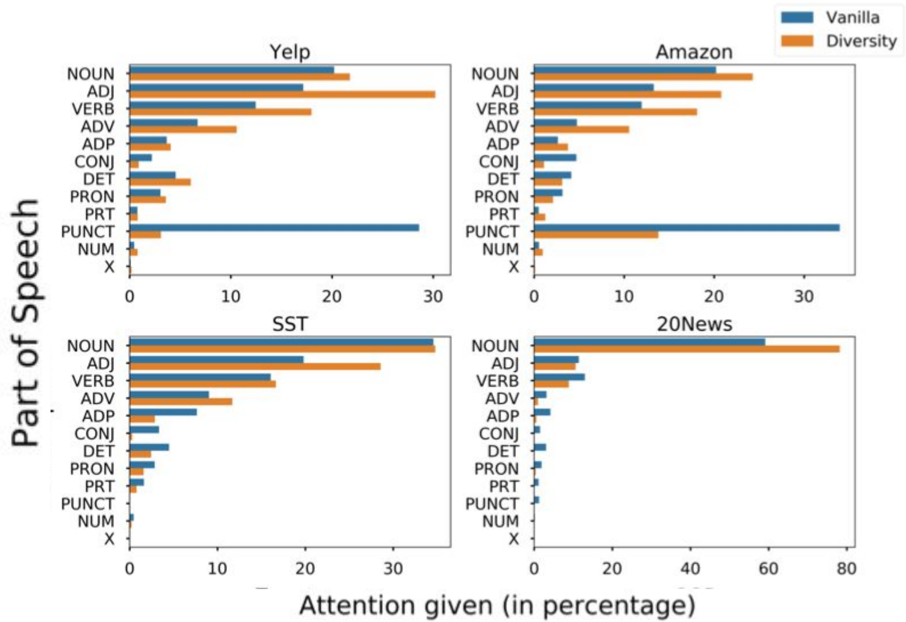
Part-of-Speech Analysis

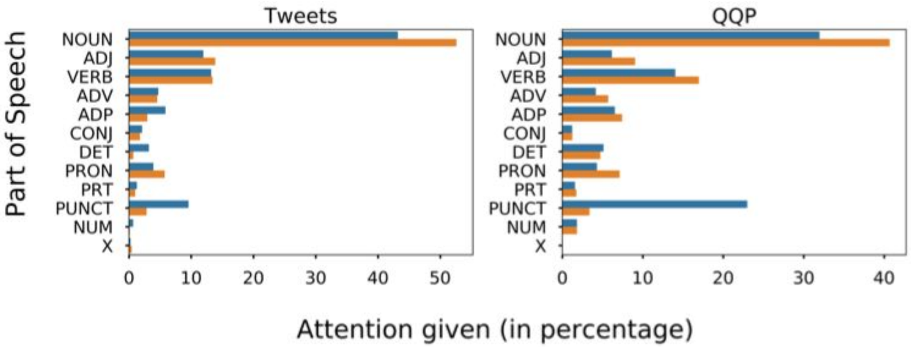


Distribution of cumulative attention given to different part-of-speech tags in the test dataset. Blue and Orange indicate the vanilla and Diversity LSTMs.

Part-of-Speech Analysis

An average increase of 49.27% attention given to adjectives across the four sentiment analysis datasets





Distribution of cumulative attention given to different part-of-speech tags in the test dataset. Blue and Orange indicate the vanilla and Diversity LSTMs.

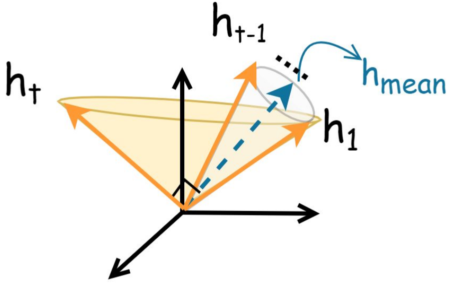
Human Evaluations

Dataset	Overall	Completness	Correctness
	Vanilla/Divers.	Vanilla/Divers.	Vanilla/Divers.
Yelp	27.7% / 72.3%	35.1% / 64.9%	10.5% / 89.5%
SNLI	37.8% / 62.2%	32.3% / 67.7%	38.9% / 61.1%
QQP	11.6% / 88.4%	11.8% / 88.2%	7.9% / 92.1%
bAbI 1	1.0% / 99.0%	4.2% / 95.8%	1.0% / 99.0%

Percentage preference given to Vanilla vs Diversity model by human annotators based on 3 criteria

Conclusion & Future Work

- In this work, we characterize why attention weights in LSTM architectures fail to provide explanations that are either faithful or plausible.
- In particular, we observe that low diversity in the hidden states induced by an LSTM tend to affect the interpretability of the resulting attention distributions.
- We then propose a orthogonalization technique and a regularization scheme aimed at improving the diversity of hidden representations.



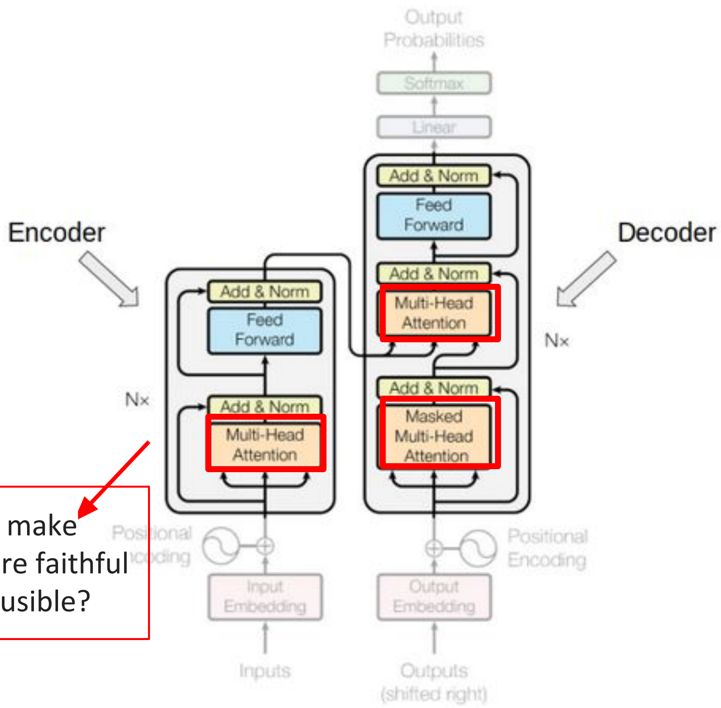
Orthogonal LSTM: Hidden state h_t is orthogonal to the mean of h_1, h_2, ..., h_{t-1}

$$L(\theta) = \text{Cross-Entropy}(y, \hat{y}|\theta) + \lambda\ \text{conicity}(\mathbf{H}|\theta)$$

Diversity Driven Training objective

Conclusion & Future Work

- Through a series of experiments, we show that our proposed methods result in more faithful and plausible attention distributions.
- As future work, we would like to extend our analysis and proposed techniques to transformer-based models and more complex downstream tasks



Can we make this more faithful and plausible?