XAI for Transformers:

Better Explanations through Conservative Propagation

TEL AVIV UNIVERSITY

berlin

BIFOLD

Speaker: Thomas Schnake

Authors:

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, Lior Wolf.

XAI for Transformers - Introduction

XAI for Transformers - Introduction

- Transformer Models [Vaswani'17] attract increasing interest and show excellent performance in many domains.

XAI for Transformers - Introduction

- Transformer Models [Vaswani'17] attract increasing interest and show excellent performance in many domains.

- Their complexity is high (billons of parameters) and their usage without XAI can be harmful (in sensitive domains).

XAI for Transformers - Introduction

- Transformer Models [Vaswani'17] attract increasing interest and show excellent performance in many domains.

- Their complexity is high (billons of parameters) and their usage without XAI can be harmful (in sensitive domains).

- The model structure is highly non-linear with **Attention Heads** and **LayerNorm.** The interpretation is therefore very challenging.

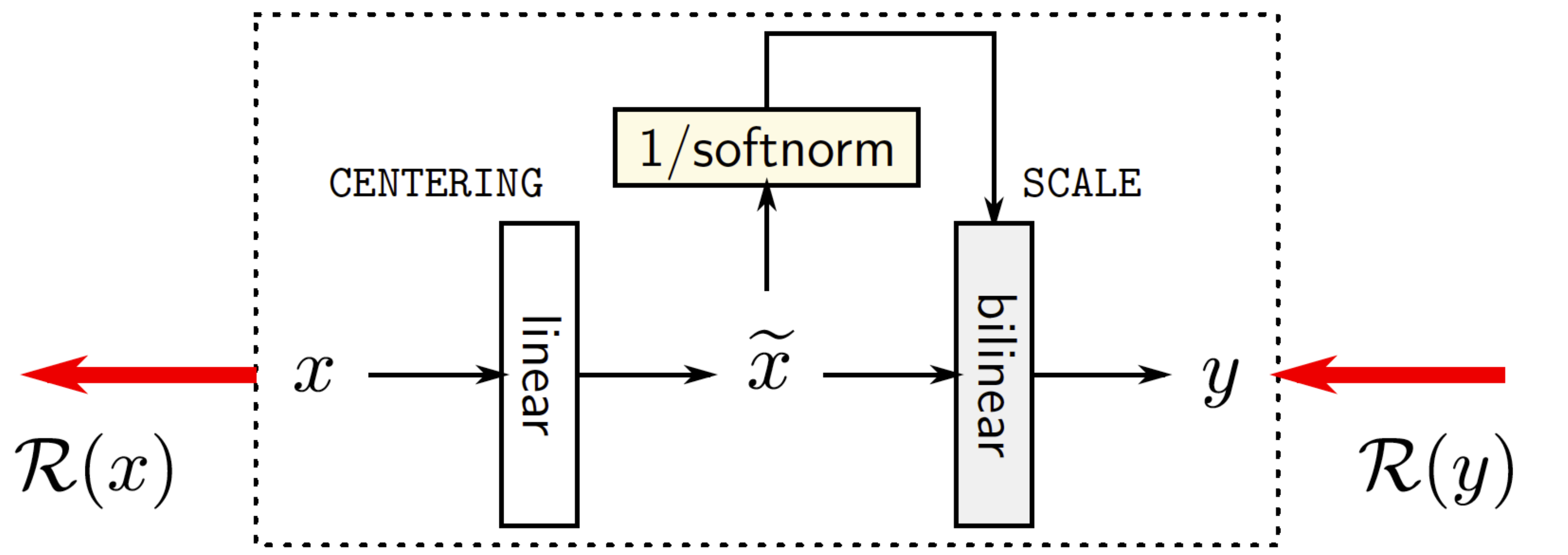
XAI for Transformers - Introduction

- Transformer Models [Vaswani'17] attract increasing interest and show excellent performance in many domains.

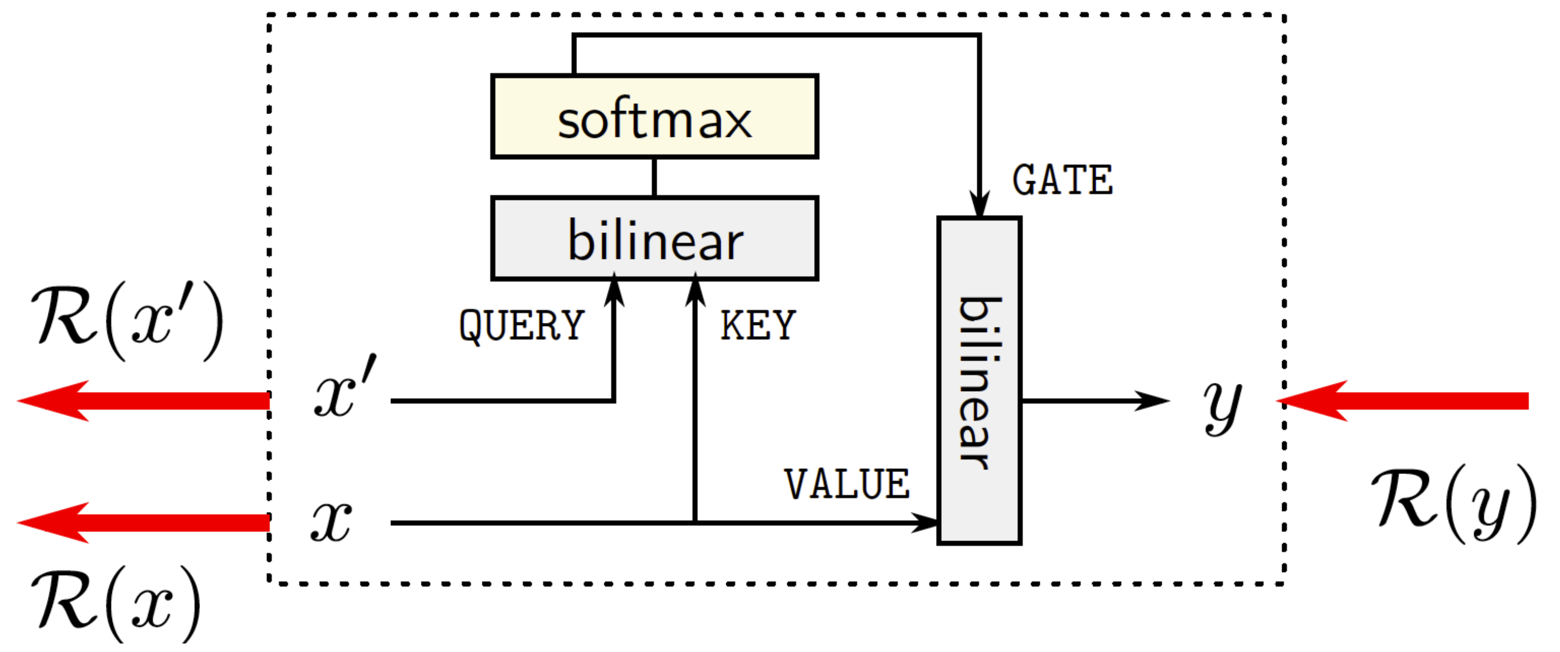
- Their complexity is high (billons of parameters) and their usage without XAI can be harmful (in sensitive domains).

- The model structure is highly non-linear with **Attention Heads** and **LayerNorm.** The interpretation is therefore very challenging.

LayerNorm (LN)



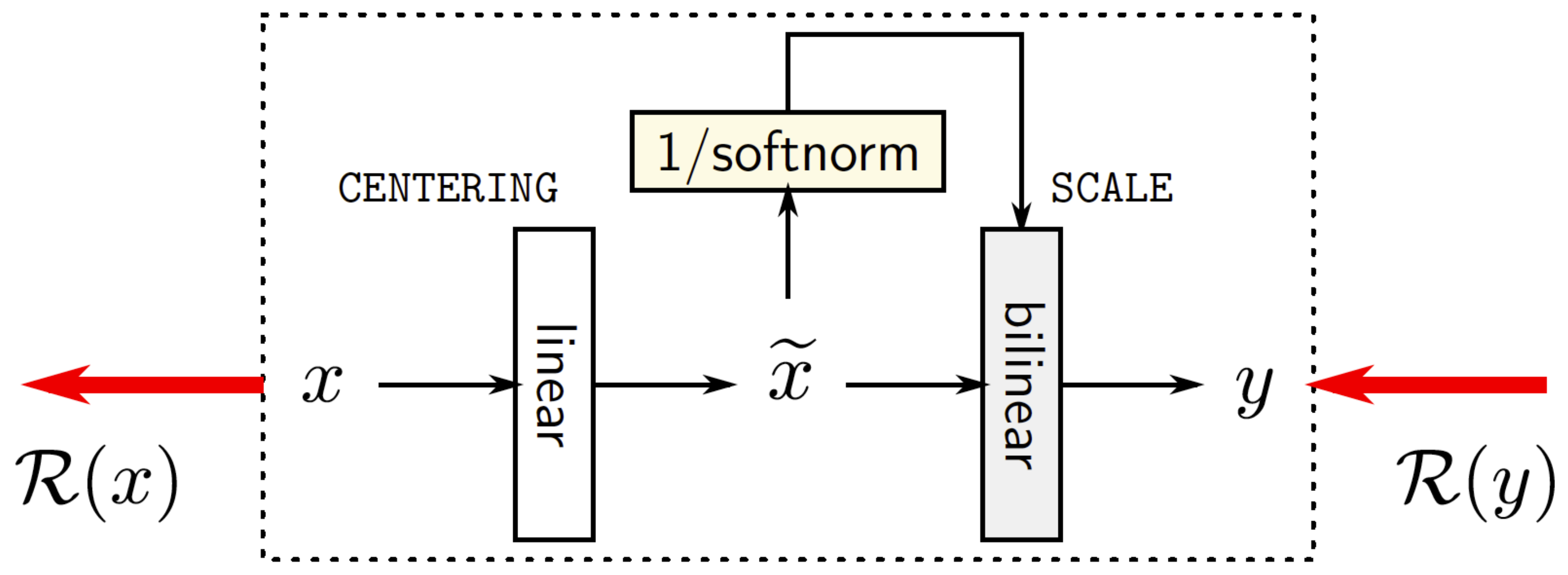
Attention Heads (AH)



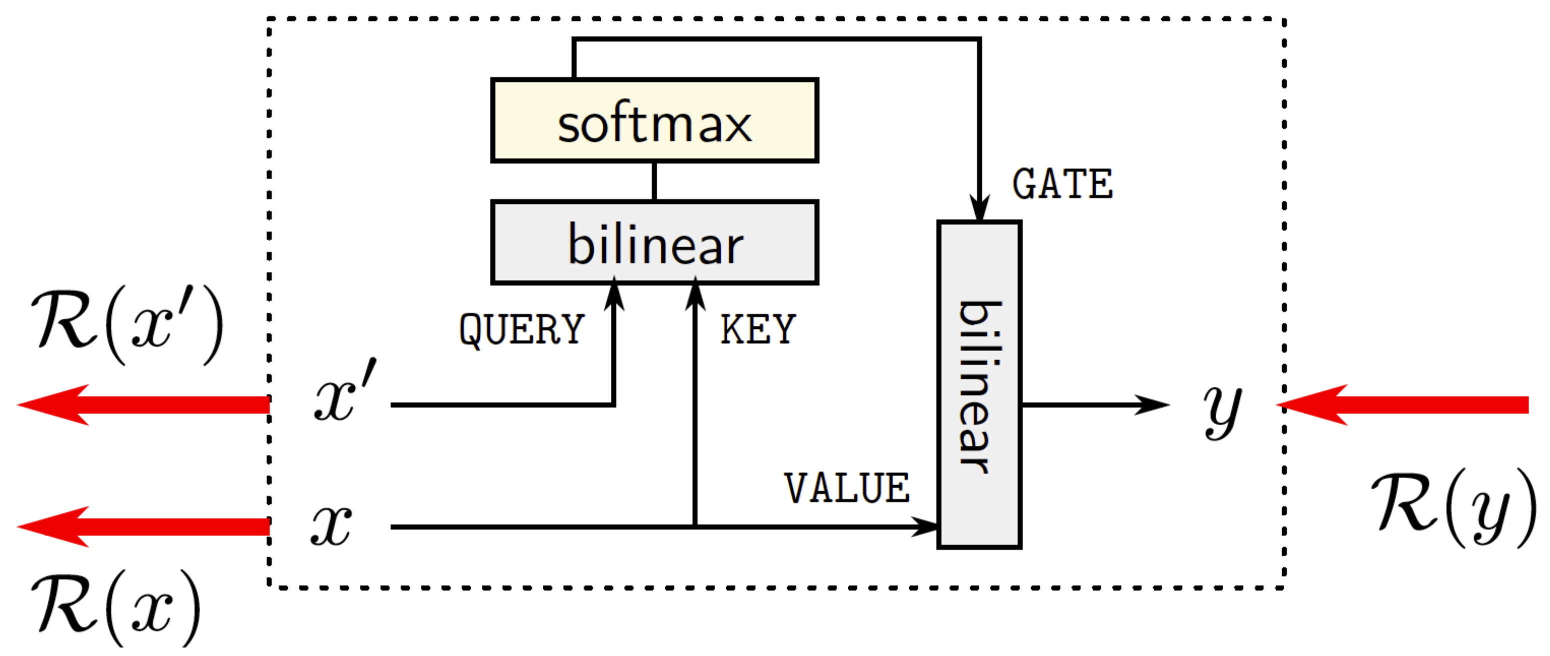
XAI for Transformers - Introduction

- Transformer Models [Vaswani'17] attract increasing interest and show excellent performance in many domains.
- Their complexity is high (billons of parameters) and their usage without XAI can be harmful (in sensitive domains).
- The model structure is highly non-linear with **Attention Heads** and **LayerNorm.** The interpretation is therefore very challenging.

LayerNorm (LN)



Attention Heads (AH)



Methodology - LRP as a diagnostic tool

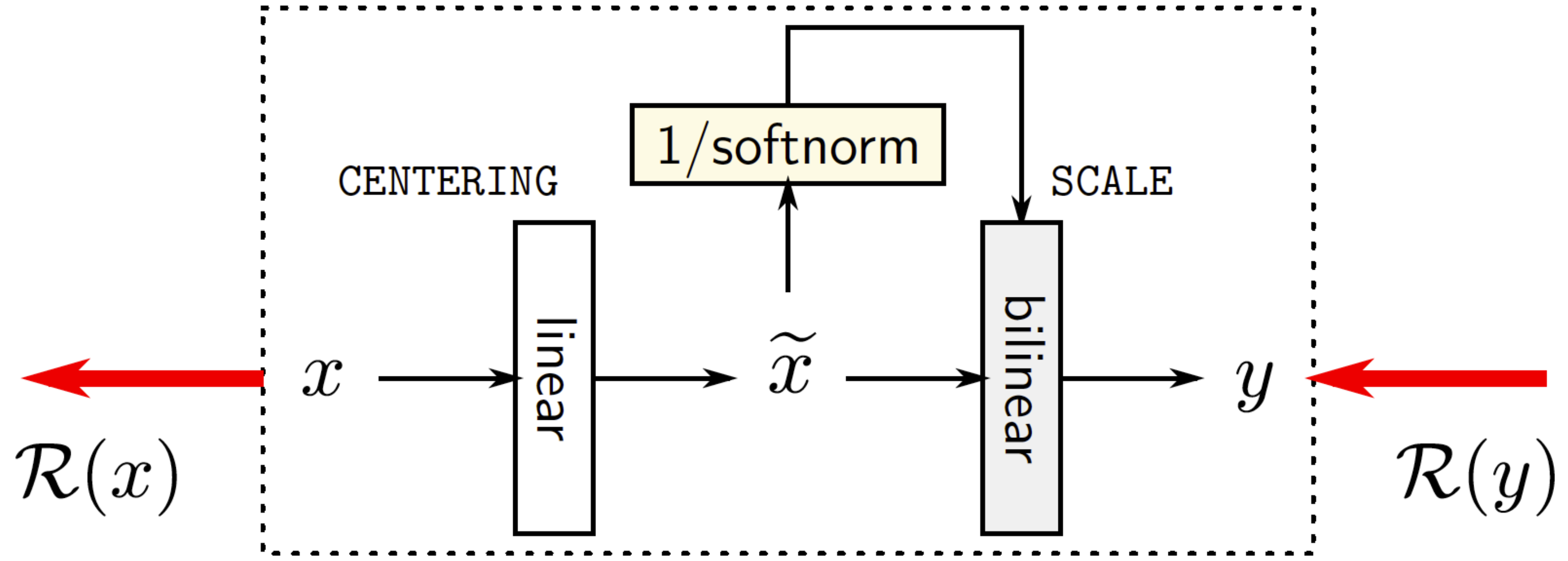
XAI for Transformers - Introduction

- Transformer Models [Vaswani'17] attract increasing interest and show excellent performance in many domains.

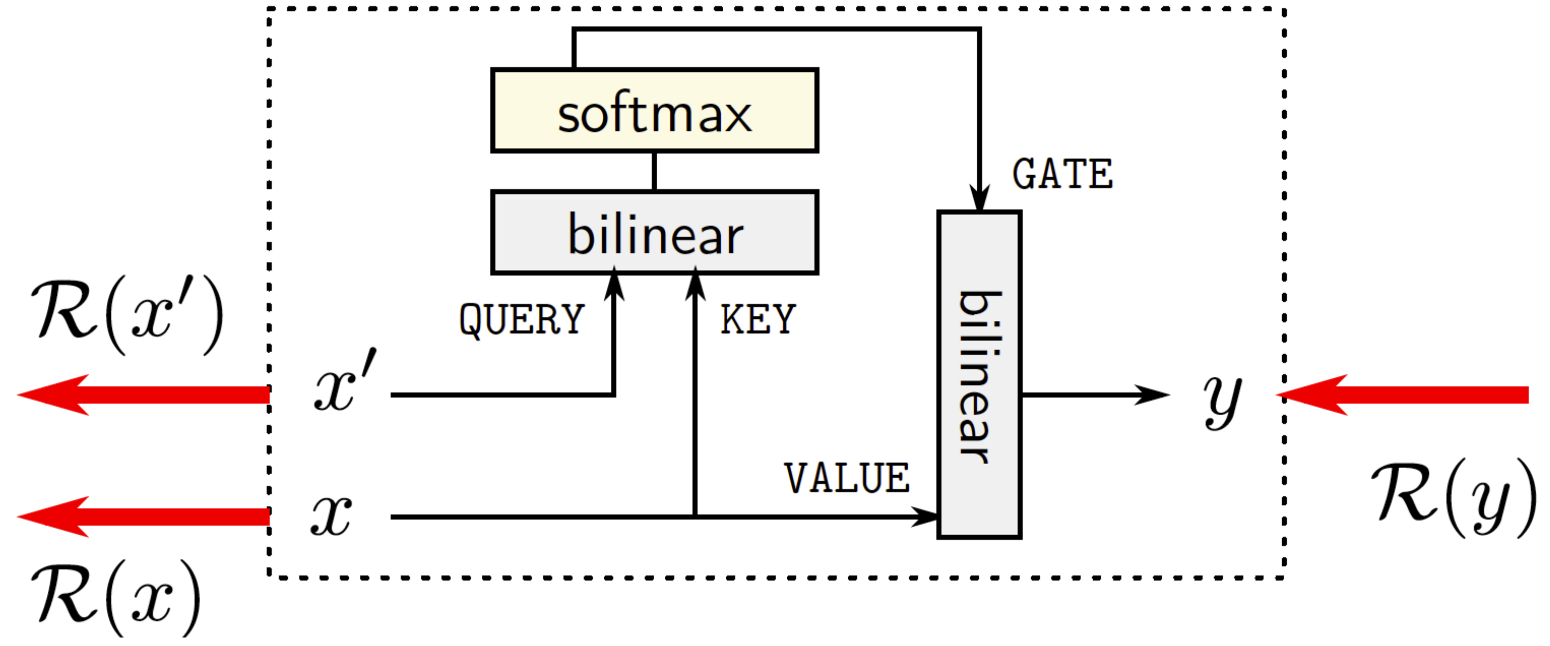
- Their complexity is high (billons of parameters) and their usage without XAI can be harmful (in sensitive domains).

- The model structure is highly non-linear with **Attention Heads** and **LayerNorm.** The interpretation is therefore very challenging.

LayerNorm (LN)



Attention Heads (AH)



Methodology - LRP as a diagnostic tool

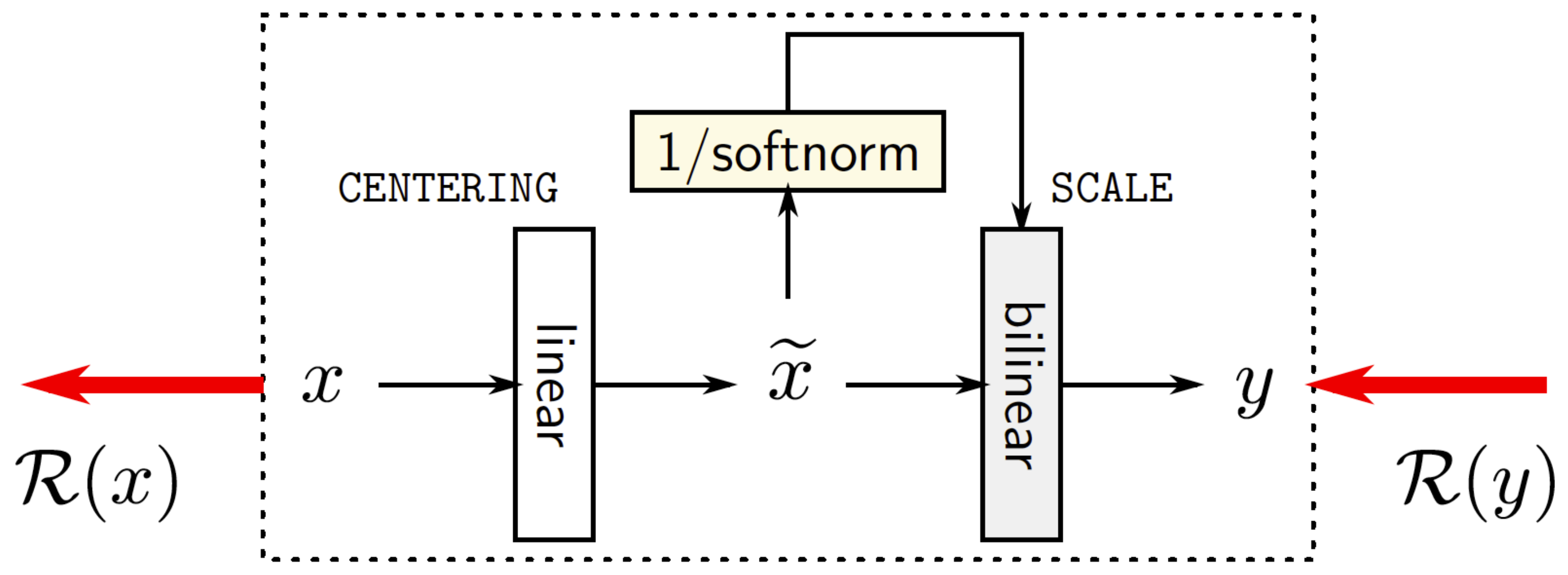
Chain rule

$$\frac{\partial f}{\partial x_i} = \sum_j \frac{\partial y_j}{\partial x_i} \frac{\partial f}{\partial y_j}$$

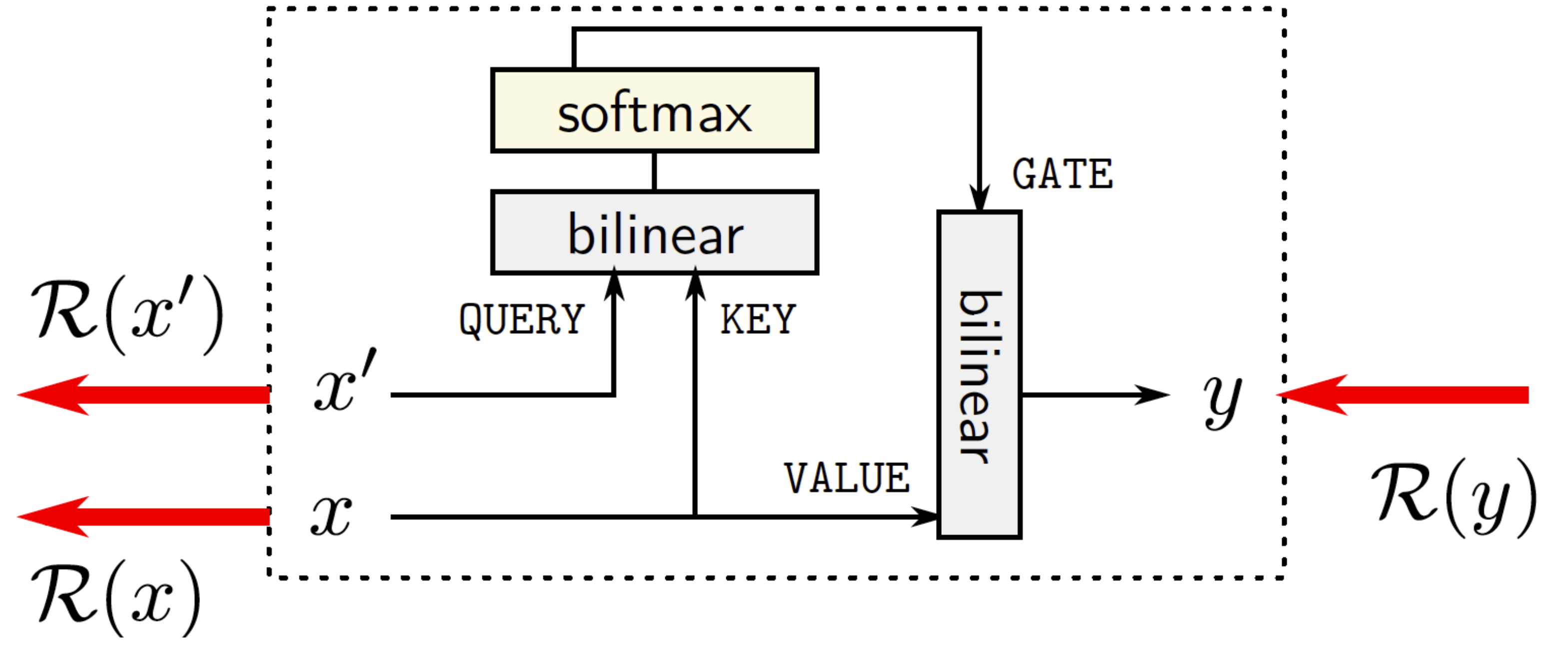
XAI for Transformers - Introduction

- Transformer Models [Vaswani'17] attract increasing interest and show excellent performance in many domains.
- Their complexity is high (billons of parameters) and their usage without XAI can be harmful (in sensitive domains).
- The model structure is highly non-linear with **Attention Heads** and **LayerNorm.** The interpretation is therefore very challenging.

LayerNorm (LN)



Attention Heads (AH)



Methodology - LRP as a diagnostic tool

Chain rule

$$\frac{\partial f}{\partial x_i} = \sum_j \frac{\partial y_j}{\partial x_i} \frac{\partial f}{\partial y_j}$$

LRP view on GI

$$\mathcal{R}(x_i) = \sum_j \frac{\partial y_j}{\partial x_i} \frac{x_i}{y_j} \mathcal{R}(y_j)$$

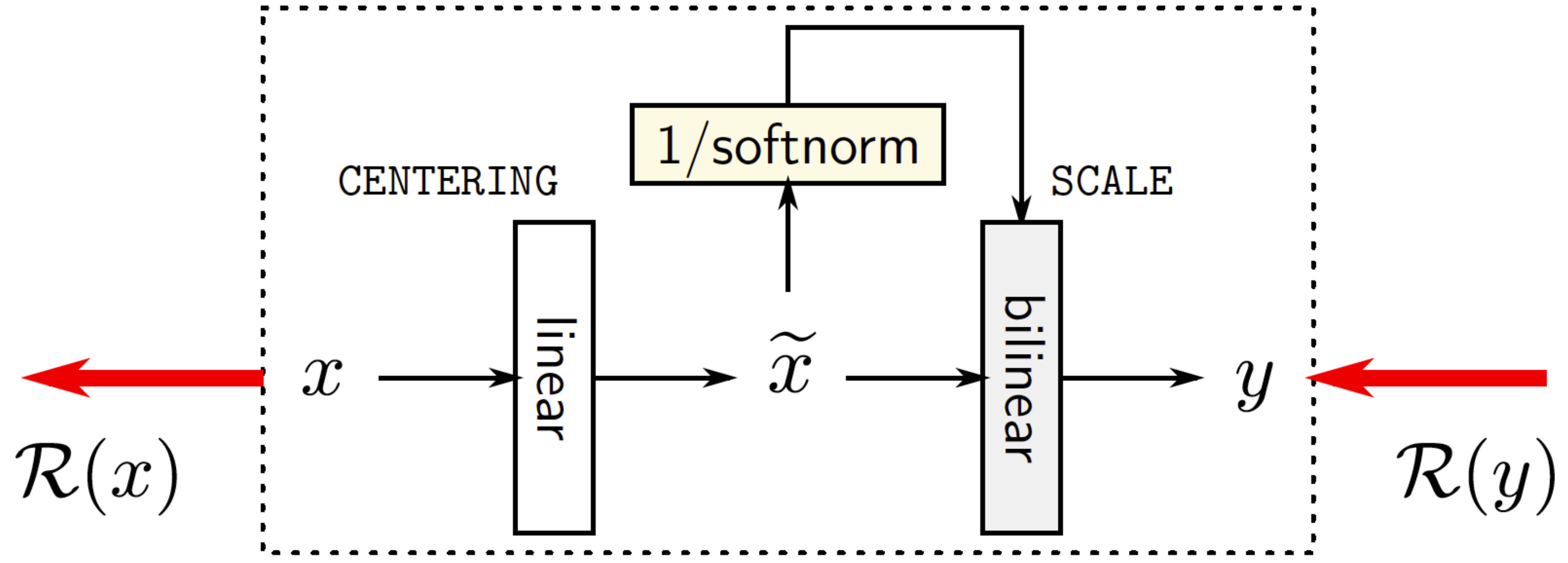
XAI for Transformers - Introduction

- Transformer Models [Vaswani'17] attract increasing interest and show excellent performance in many domains.

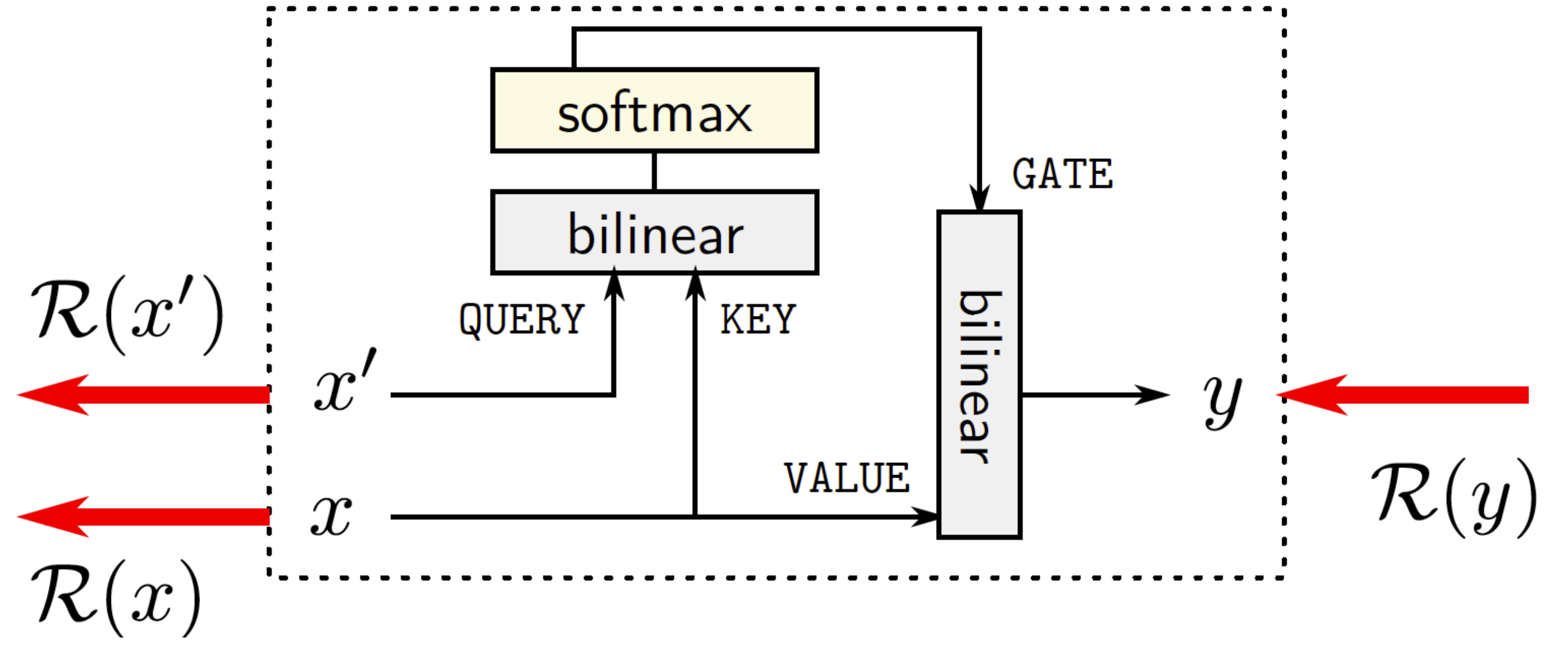
- Their complexity is high (billons of parameters) and their usage without XAI can be harmful (in sensitive domains).

- The model structure is highly non-linear with **Attention Heads** and **LayerNorm.** The interpretation is therefore very challenging.

LayerNorm (LN)



Attention Heads (AH)



Methodology - LRP as a diagnostic tool

Chain rule

$$\frac{\partial f}{\partial x_i} = \sum_j \frac{\partial y_j}{\partial x_i} \frac{\partial f}{\partial y_j}$$

LRP view on GI

$$\mathcal{R}(x_i) = \sum_j \frac{\partial y_j}{\partial x_i} \frac{x_i}{y_j} \mathcal{R}(y_j)$$

*For **conservation**, test whether*

$$\sum_i \mathcal{R}(x_i) = \sum_j \mathcal{R}(y_j)$$

Conservation Test for Attention-Heads and LayerNorm

Conservation Test for Attention-Heads and LayerNorm

Conservation Test - AH

$$\sum_i \mathcal{R}(x_i) + \sum_j \mathcal{R}(x'_j) = \sum_i \mathcal{R}(y_i) + \delta(x, x', y)$$

Alternative back-propagation

$$\mathcal{R}(x_i) = \sum_j \frac{x_i p_{ij}}{\sum_{i'} x_{i'} p_{i'j}} \mathcal{R}(y_j)$$

Conservation Test for Attention-Heads and LayerNorm

Conservation Test - AH

$$\sum_i \mathcal{R}(x_i) + \sum_j \mathcal{R}(x'_j) = \sum_i \mathcal{R}(y_i) + \underline{\delta(x, x', y)}$$

Alternative back-propagation

$$\mathcal{R}(x_i) = \sum_j \frac{x_i p_{ij}}{\sum_{i'} x_{i'} p_{i'j}} \mathcal{R}(y_j)$$

Conservation Test for Attention-Heads and LayerNorm

Conservation Test - AH

$$\sum_i \mathcal{R}(x_i) + \sum_j \mathcal{R}(x'_j) = \sum_i \mathcal{R}(y_i) + \delta(x, x', y)$$

Alternative back-propagation

$$\mathcal{R}(x_i) = \sum_j \frac{x_i p_{ij}}{\sum_{i'} x_{i'} p_{i'j}} \mathcal{R}(y_j)$$

Implementation Trick - AH

Forward pass

Before: $y_j = \sum_i x_i p(x_i, x'_j)$

After: $y_j = \sum_i x_i [p(x_i, x'_j)].\mathrm{detach}()$

Conservation Test for Attention-Heads and LayerNorm

Conservation Test - AH

$$\sum_i \mathcal{R}(x_i) + \sum_j \mathcal{R}(x'_j) = \sum_i \mathcal{R}(y_i) + \underline{\delta(x, x', y)}$$

Alternative back-propagation

$$\mathcal{R}(x_i) = \sum_j \frac{x_i p_{ij}}{\sum_{i'} x_{i'} p_{i'j}} \mathcal{R}(y_j)$$

Implementation Trick - AH

Forward pass

Before: $y_j = \sum_i x_i p(x_i, x'_j)$

After: $y_j = \sum_i x_i [p(x_i, x'_j)].\mathrm{detach}()$

Implementation Trick - LN

Forward pass

Before: $y_i = \frac{x_i - \mathbb{E}[x]}{\sqrt{\epsilon + \mathrm{Var}[x]}}$

After: $y_i = \frac{x_i - \mathbb{E}[x]}{[\sqrt{\epsilon + \mathrm{Var}[x]}].\mathrm{detach}()}$

Conservation Test for Attention-Heads and LayerNorm

Conservation Test - AH

$$\sum_i \mathcal{R}(x_i) + \sum_j \mathcal{R}(x'_j) = \sum_i \mathcal{R}(y_i) + \underline{\delta(x, x', y)}$$

Alternative back-propagation

$$\mathcal{R}(x_i) = \sum_j \frac{x_i p_{ij}}{\sum_{i'} x_{i'} p_{i'j}} \mathcal{R}(y_j)$$

Implementation Trick - AH

Forward pass

Before: $y_j = \sum_i x_i p(x_i, x'_j)$

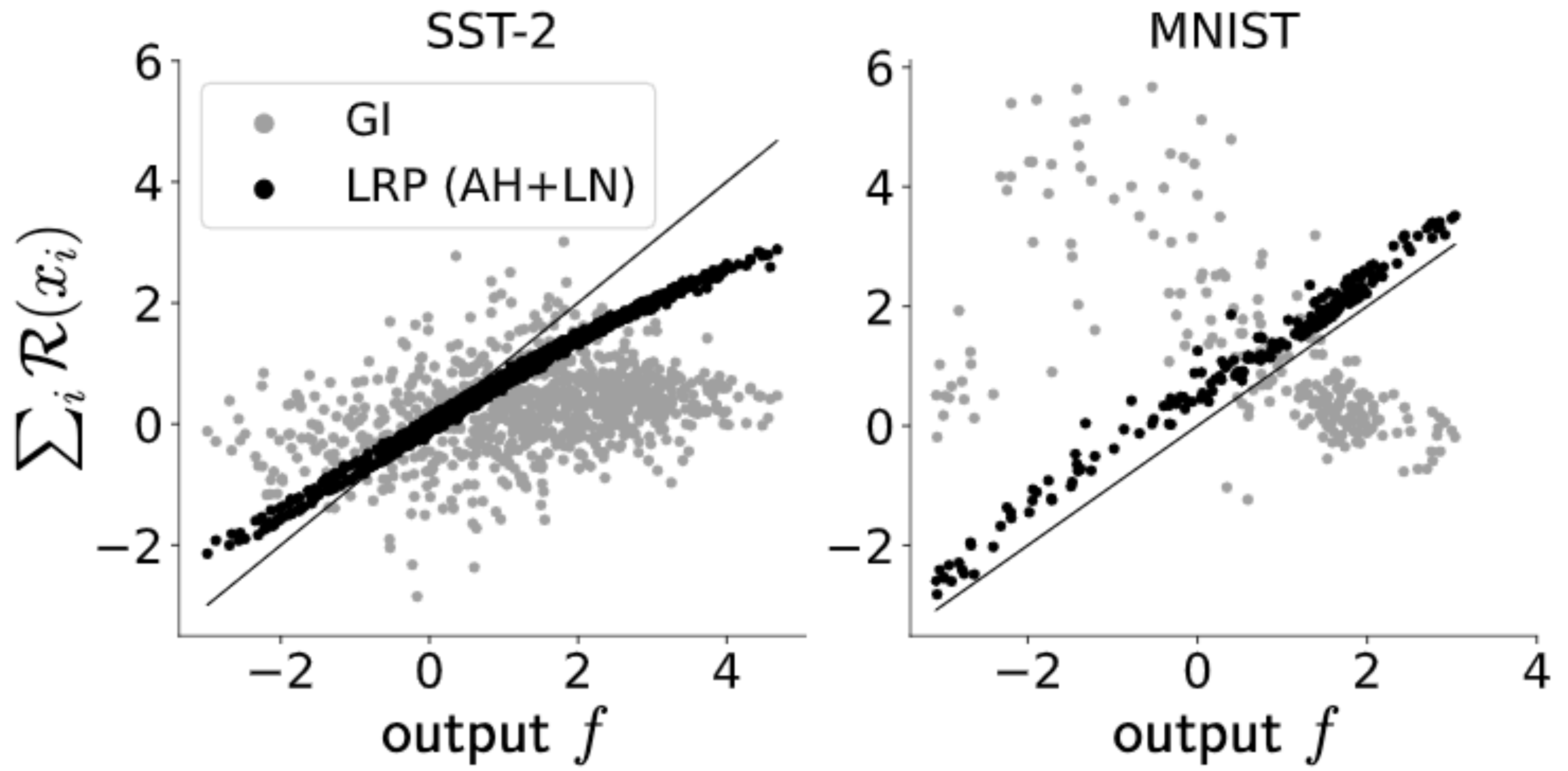
After: $y_j = \sum_i x_i [p(x_i, x'_j)].\text{detach}()$

Implementation Trick - LN

Forward pass

Before: $y_i = \dfrac{x_i - \mathbb{E}[x]}{\sqrt{\epsilon + \text{Var}[x]}}$

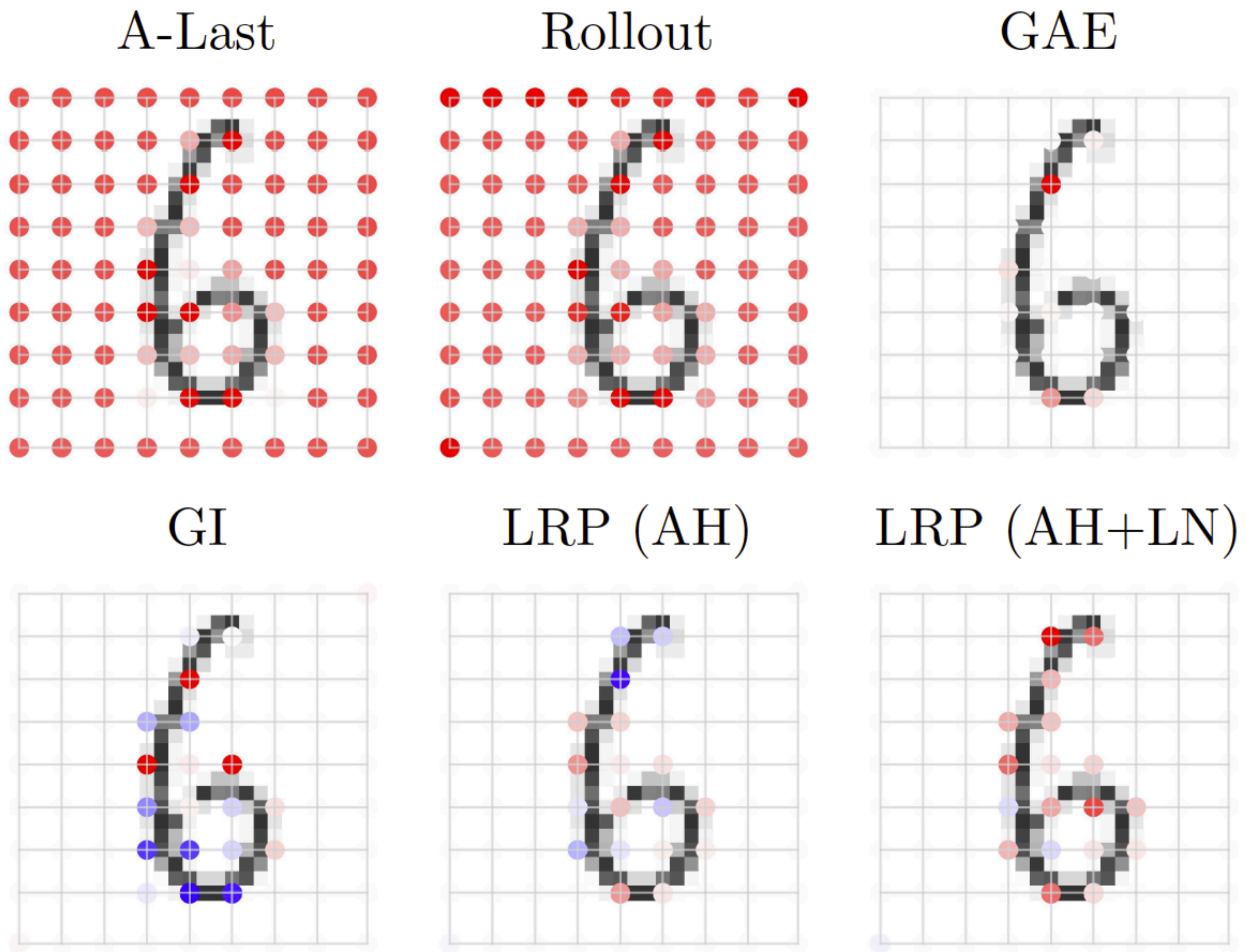
After: $y_i = \dfrac{x_i - \mathbb{E}[x]}{[\sqrt{\epsilon + \text{Var}[x]}].\text{detach}()}$



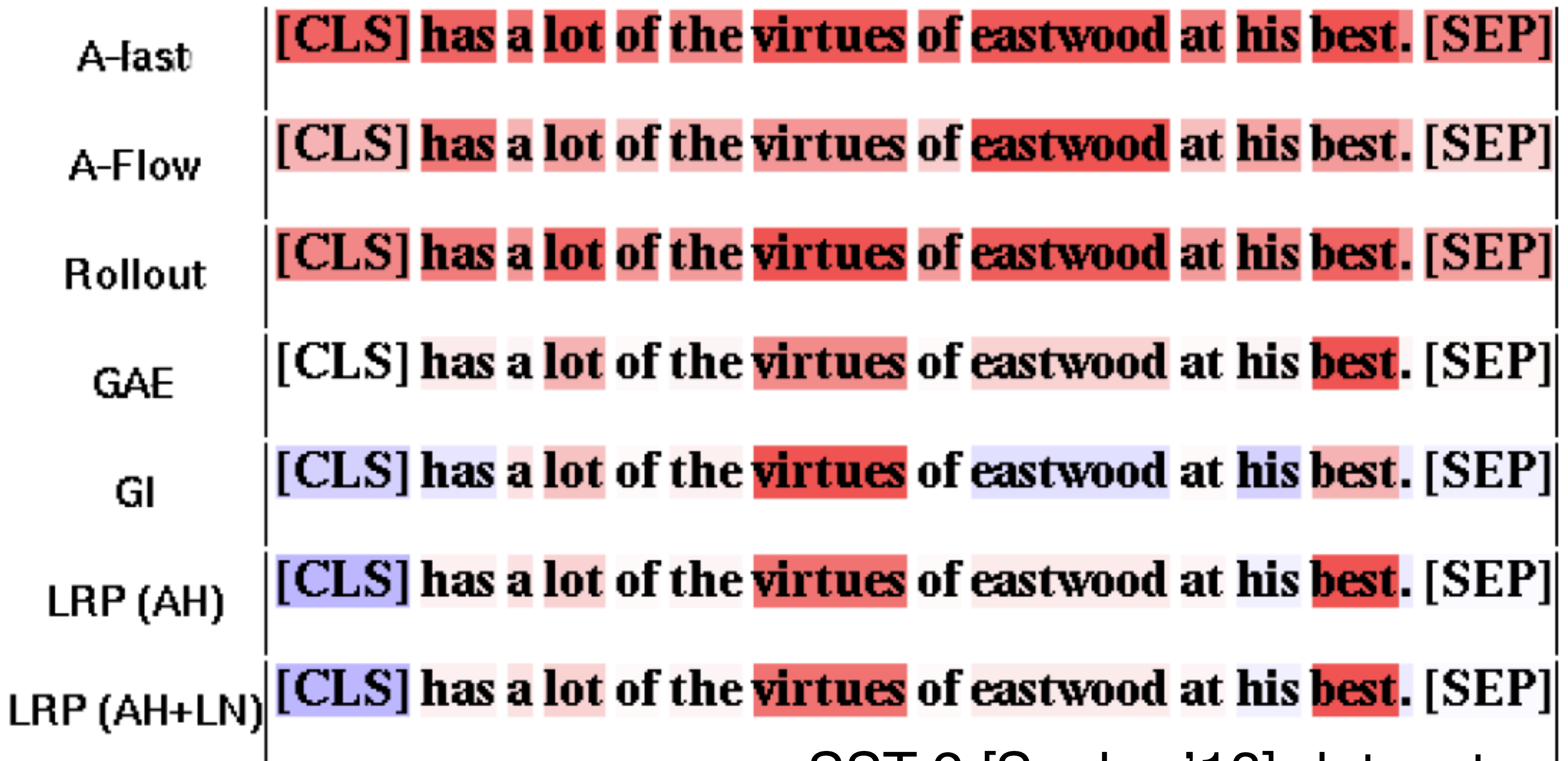
Evaluation

Evaluation

Qualitative



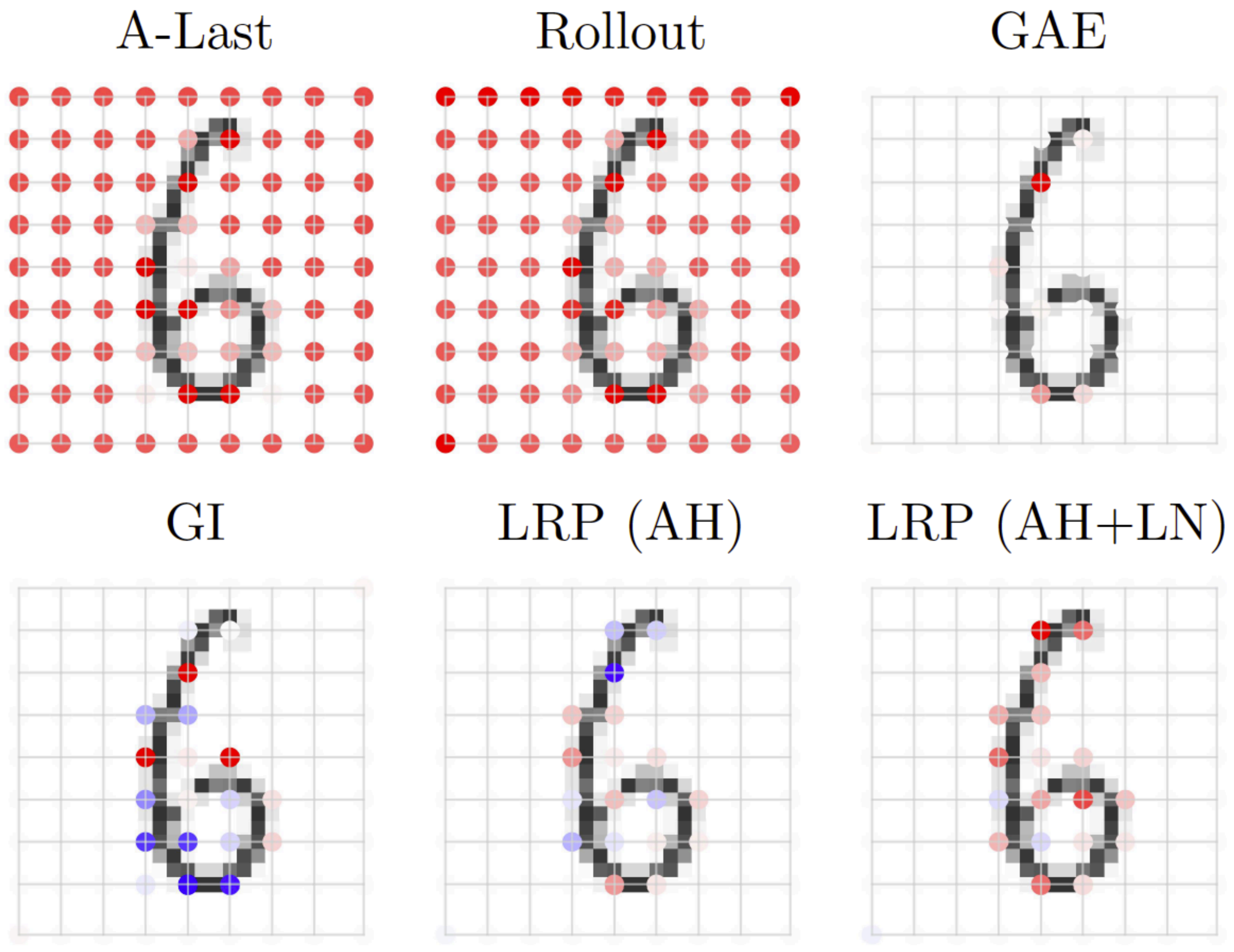
MNIST [Deng'12] dataset



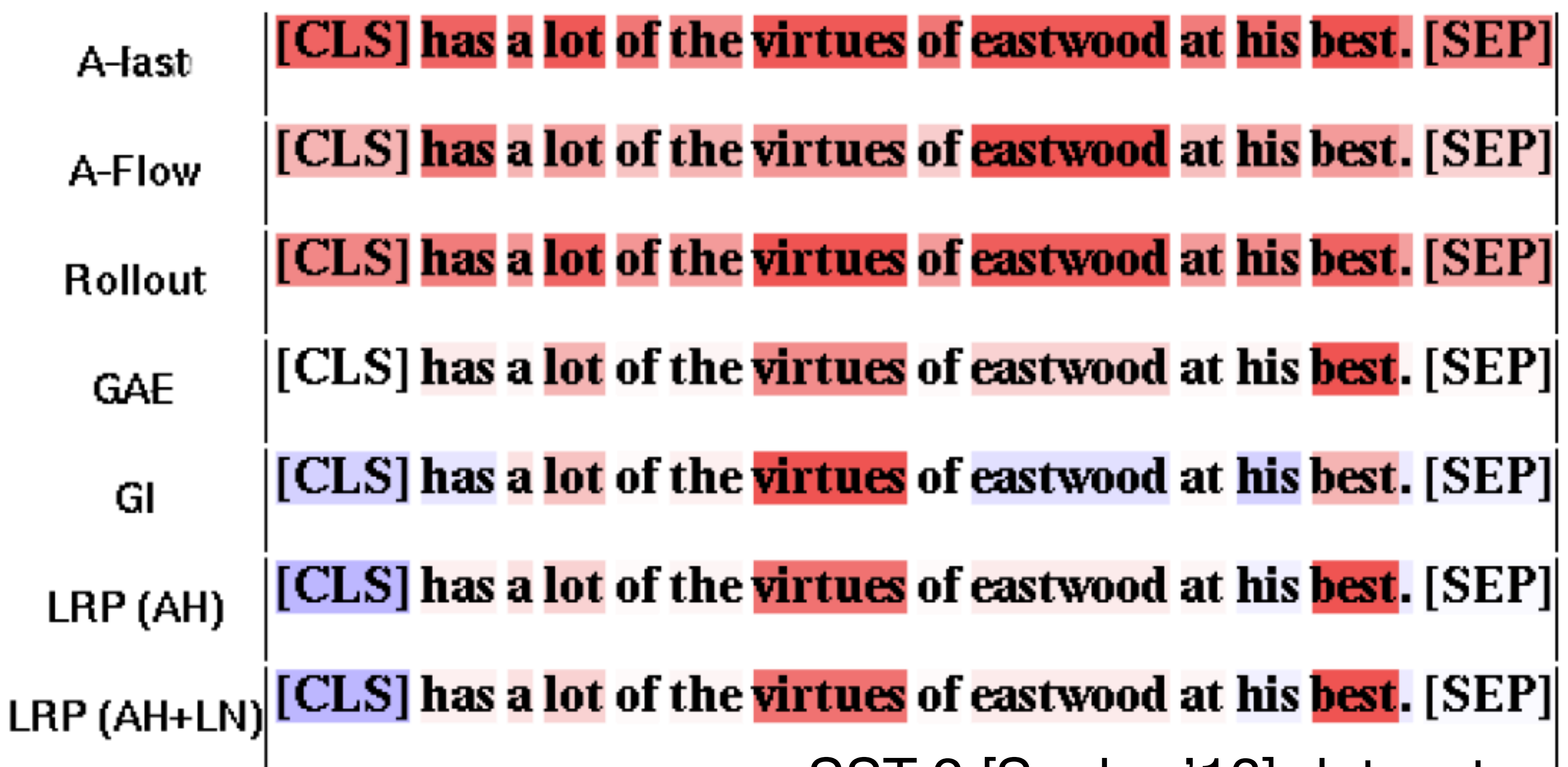
SST-2 [Socher'13] dataset

Evaluation

Qualitative



MNIST [Deng'12] dataset



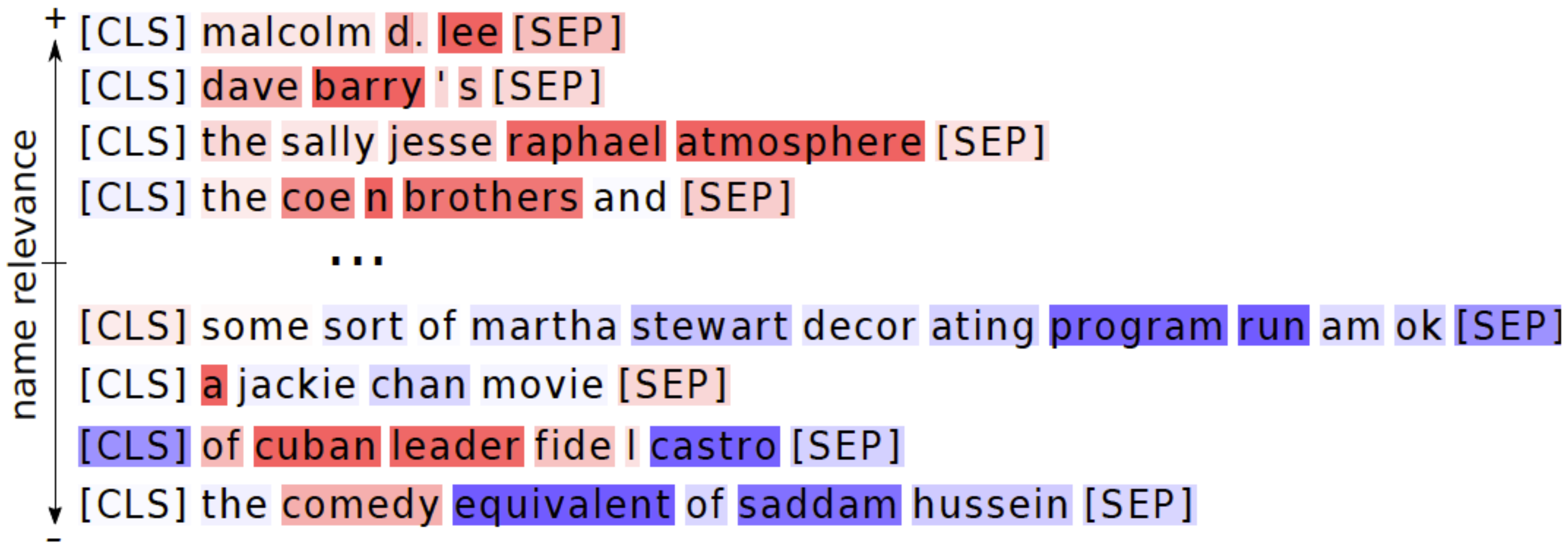
SST-2 [Socher'13] dataset

Quantitative

Area under the activation curve (AUAC)

Method	IMDB	SST-2	BACE	MNIST	T-Emotions	T-Hate	T-Sentiment	Meld-S	Semaine
Random	.673	.664	.624	.324	.516	.640	.484	.460	.432
A-Last	.708	.712	.620	.862	.542	.663	.515	.483	.451
A-Flow	-	.711	.637	-	-	-	-	-	-
Rollout	.738	.713	.653	.358	.554	.659	.520	.489	.441
GAE	.872	.821	.675	.426	.675	.762	.611	.548	.532
GI	.920	.847	.646	.942	.652	.772	.651	.591	.529
LRP(AH)	.911	.855	.645	.942	.675	.797	.668	.594	.544
LRP (LN)	.935	.907	.702	.947	.735	.829	.710	.632	.593
LRP(AH+LN)	**.939**	**.908**	**.707**	**.948**	**.750**	**.838**	**.713**	**.635**	**.606**

Bias of names on the SST-Task



References

[Vaswani’17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *In Advances in neural information processing systems*, pp. 5998–6008, 2017.

[Socher’13] R. Socher, A. Perelygin, J.Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, *in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistic, 2013, pp. 1631–1642.

[Subramanian’16] G. Subramanian, B. Ramsundar, V. Pande, and R. A. Denny. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, vol. 56, no. 10, pp. 1936–1949, 2016.

[Deng’12] Deng L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*. vol. 29, no. 6: pp. 141–142, 2012.

[Bach'15] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015.

[Chefer'21a] Chefer, H., Gur, S., and Wolf, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 397–406, 2021a.

[Chefer'21b] Chefer, H., Gur, S., and Wolf, L. Transformer interpretability beyond attention visualization. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021b.

[Montavon'18] Montavon, G., Samek, W., and Müller, K. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, 73:1–15, 2018.