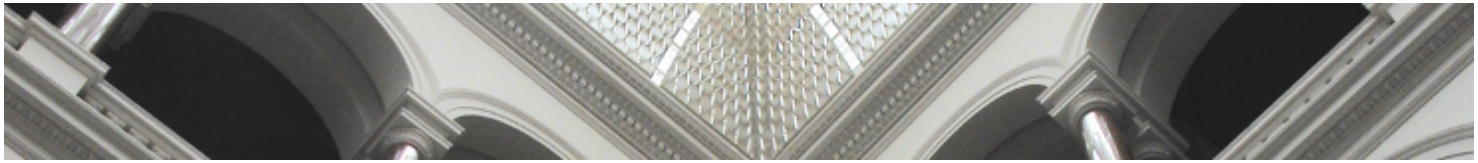
Technische Universität Berlin

Designing Recurrent Neural Networks for Explainability

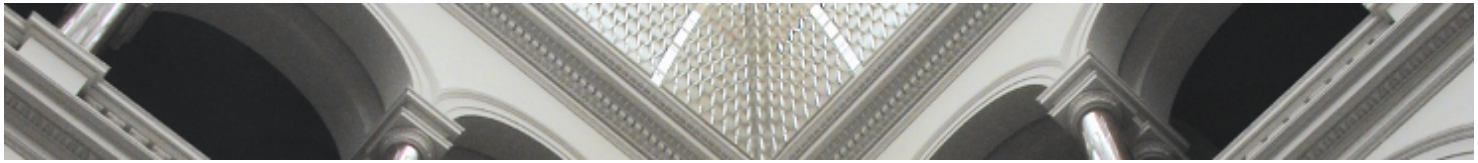
Pattarawat Chormai | Technische Universität Berlin | 14/11/2018



Technische Universität Berlin

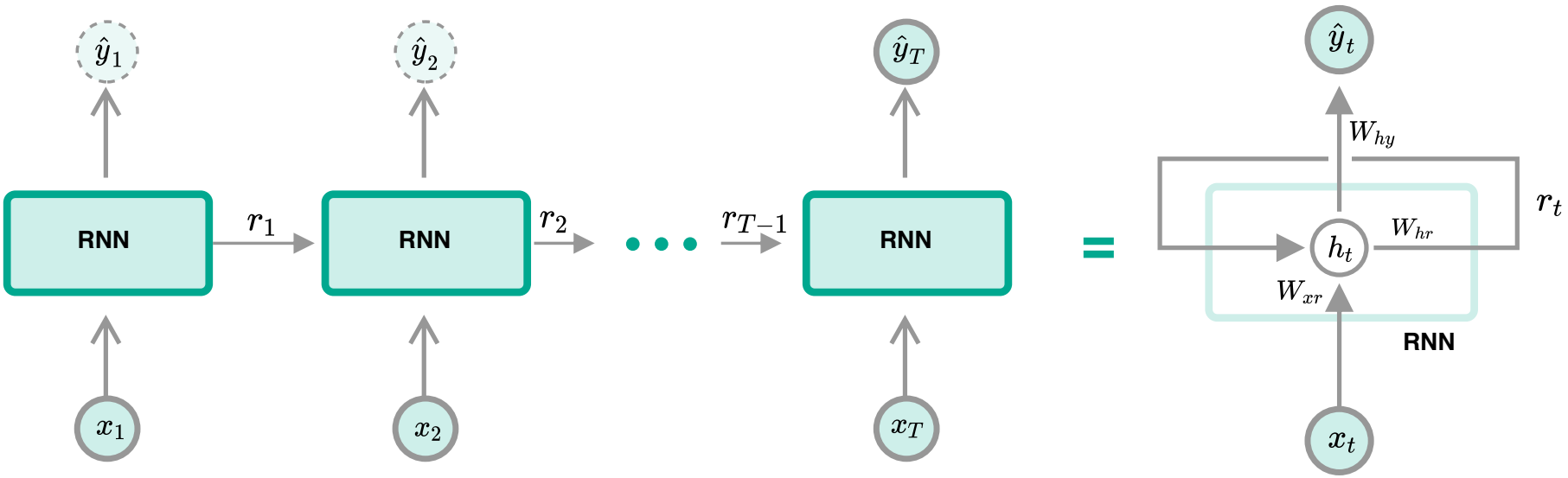
Motivation

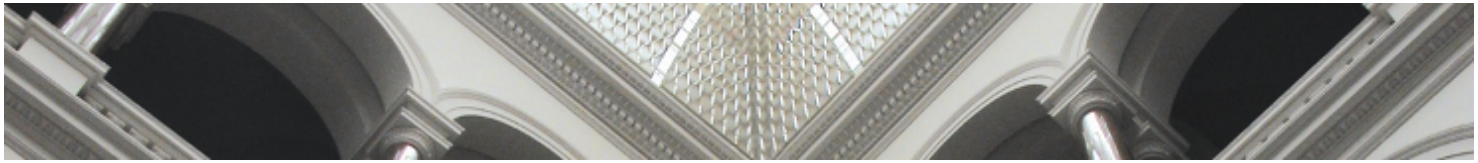
- Impact of the architecture of RNNs on explainability
- Are deeper RNNs more explainable?
- Are there ways to make RNNs more explainable?



Technische Universität Berlin

Recurrent Neural Networks (RNNs)

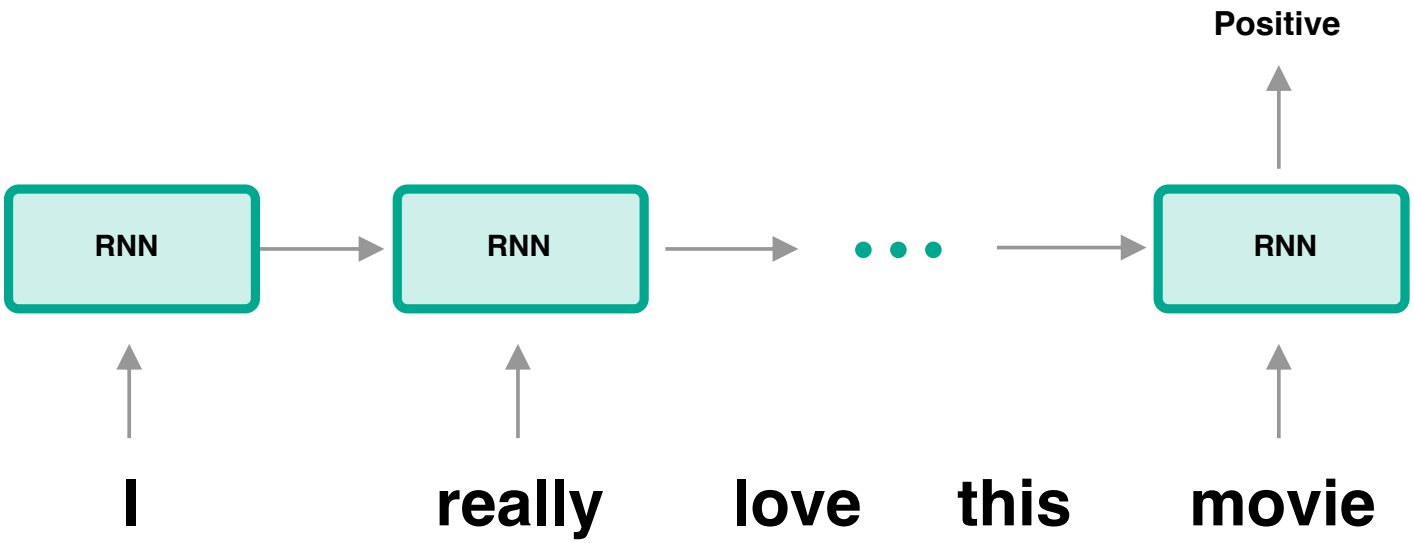


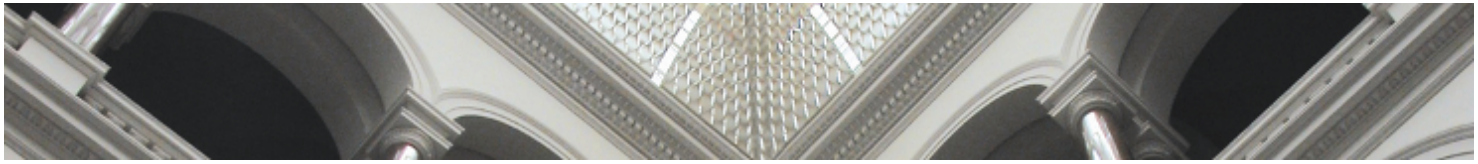


Technische Universität Berlin

Recurrent Neural Networks (RNNs)

– Applications: sentiment analysis, NLP, machine translation, and image captioning

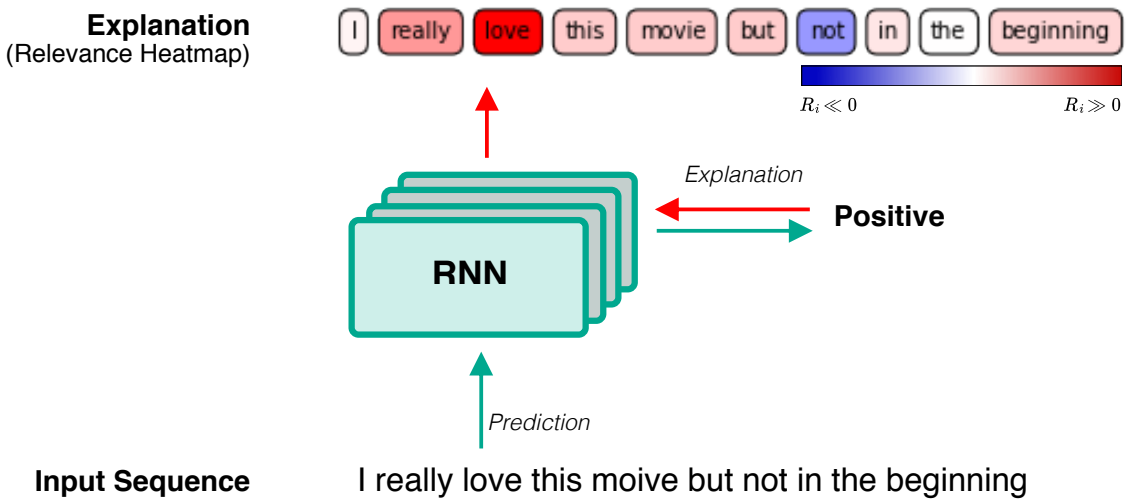


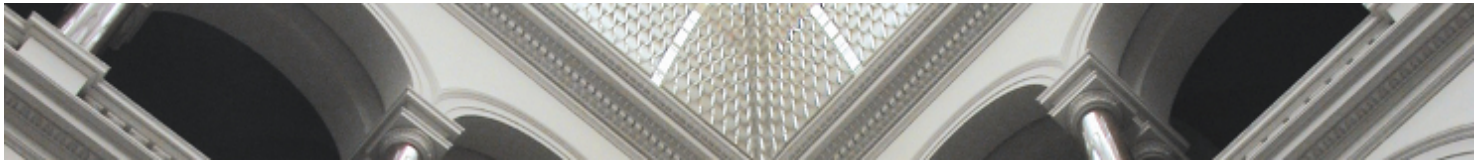


Technische Universität Berlin

Explainability I

- Ability to provide sensible explanation towards how input associates to a particular output.
- For example, why does the RNN classify this text as a positive review?





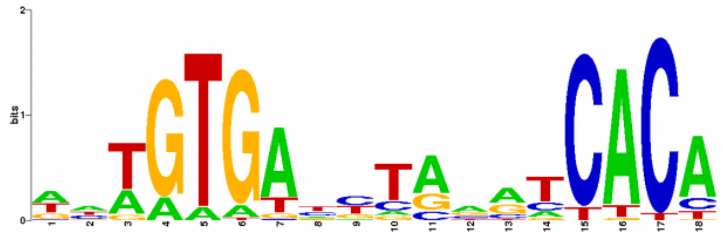
Technische Universität Berlin

Explainability II

– Why do we need explainable models?

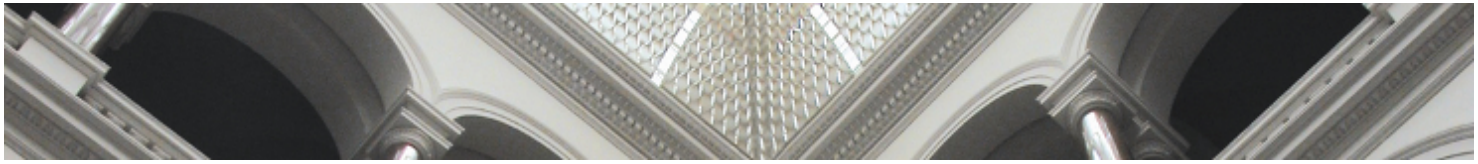
- Aid decision and establish trust in the system, especially in critical applications such as healthcare.

DNA Motif Discovery



taken from http://genomeinformatics.uni-due.de

- Developers can assure that trained models would work as expected.



Technische Universität Berlin

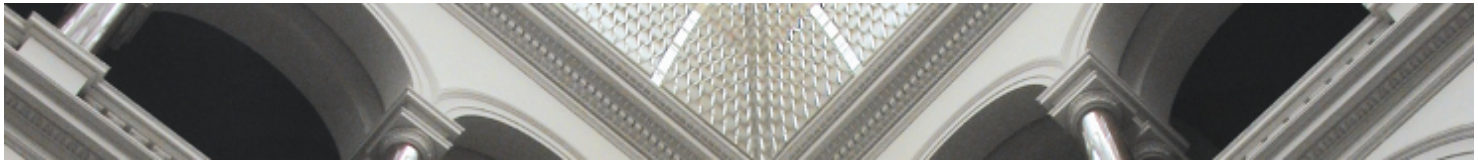
Explanation Methods I

1. Sensitivity Analysis [SVZ13]

$$R_i(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_i} \right)^2$$

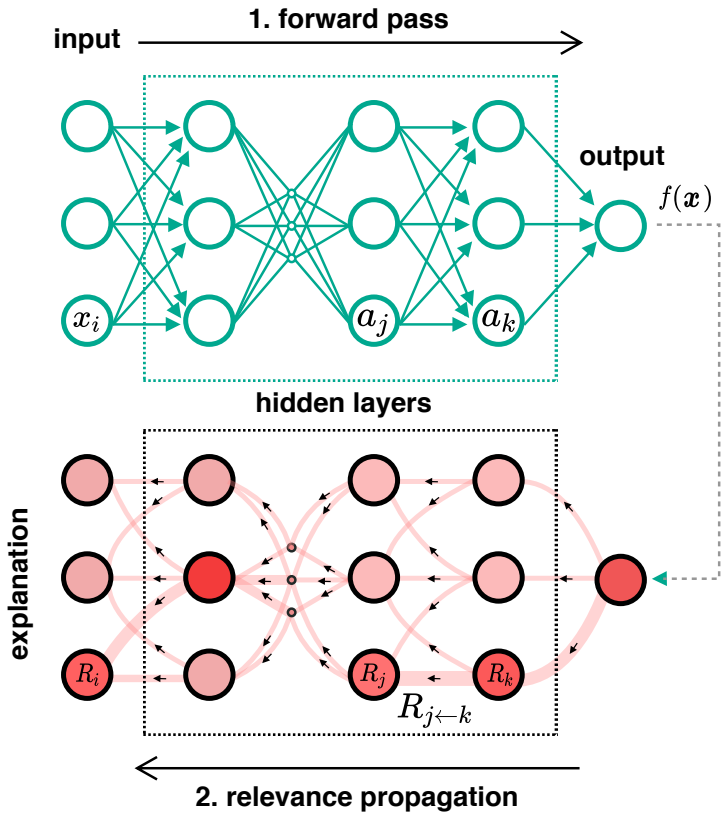
2. Guided Backprop [SDBR15]

$$\frac{\partial_* f(\boldsymbol{x})}{\partial h_j} = \mathbb{1}\left[h_j > 0 \right] \max\left(0, \frac{\partial_* f(\boldsymbol{x})}{\partial a_j} \right); \qquad R_i(\boldsymbol{x}) = \left(\frac{\partial_* f(\boldsymbol{x})}{\partial x_i} \right)^2$$

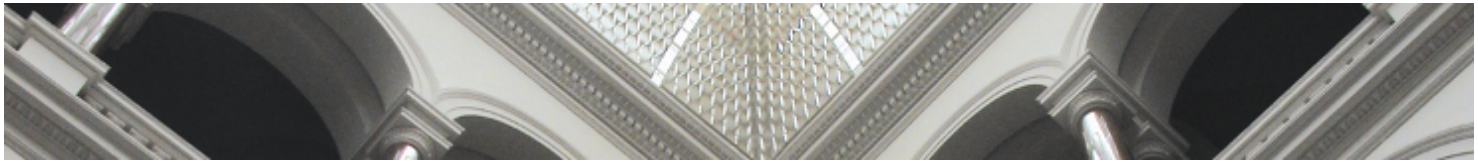


Explanation Methods II

3. Layer-wise Relevance Propagation (LRP) [BML+16]



$$R_j(\boldsymbol{x}) = \sum_k \left(\alpha \frac{a_j w_{jk}^+}{\sum_{j'} a_{j'} w_{j'k}^+} - \beta \frac{a_j w_{jk}^-}{\sum_{j'} a_{j'} w_{j'k}^-} \right) R_k(\boldsymbol{x})$$



Technische Universität Berlin

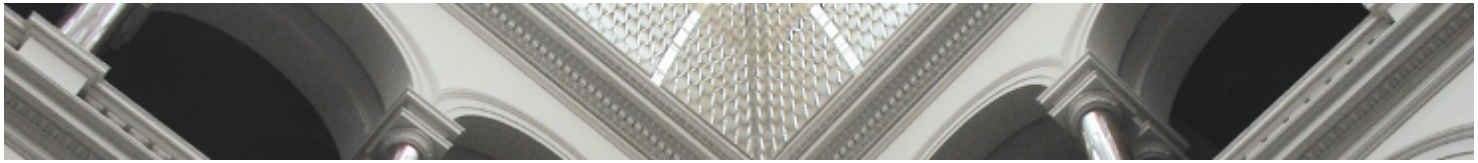
Explanation Methods III

4. Deep Taylor Decomposition [MLB$^+$17]
   - Based on LRP and derived specifically for explaining ReLU-based architectures
   - Two important propagation rules:
     – z^+ for $a_j \in \mathbb{R}^+$

$$R_j = \sum_k \frac{a_j w_{jk}^+}{\sum_j a_j w_{jk}^+} R_k$$

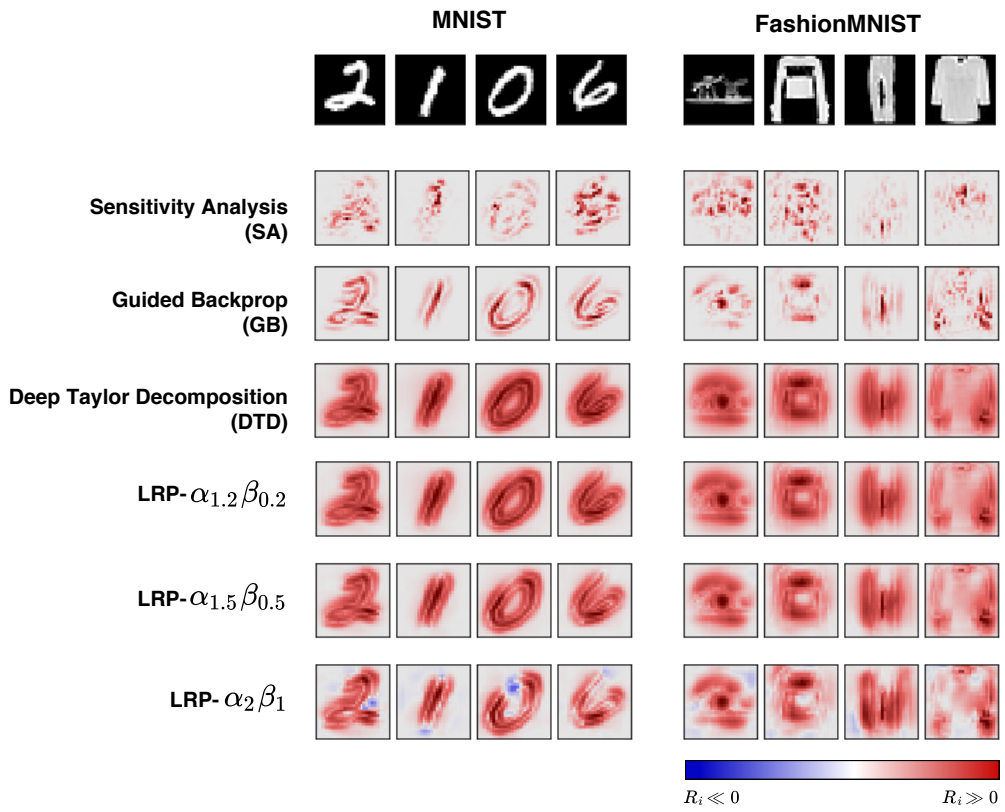
     – $z^{\mathcal{B}}$ for $a_j \in [l_j, h_j]$ where $l_j \leq 0 < h_j$

$$R_j = \sum_k \frac{a_j w_{jk} - l_j w_{jk}^+ - h_j w_{jk}^-}{\sum_{j'} a_{j'} w_{j'k} - l_{j'} w_{j'k}^+ - h_{j'} w_{j'k}^-}$$

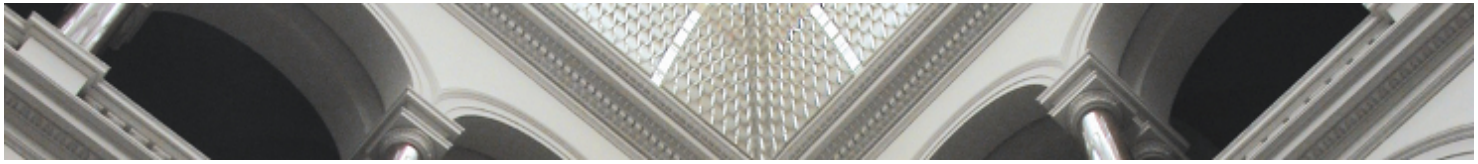


Example of relevance heatmaps

– Explaining the classification decisions of a LeNet-5 type network.



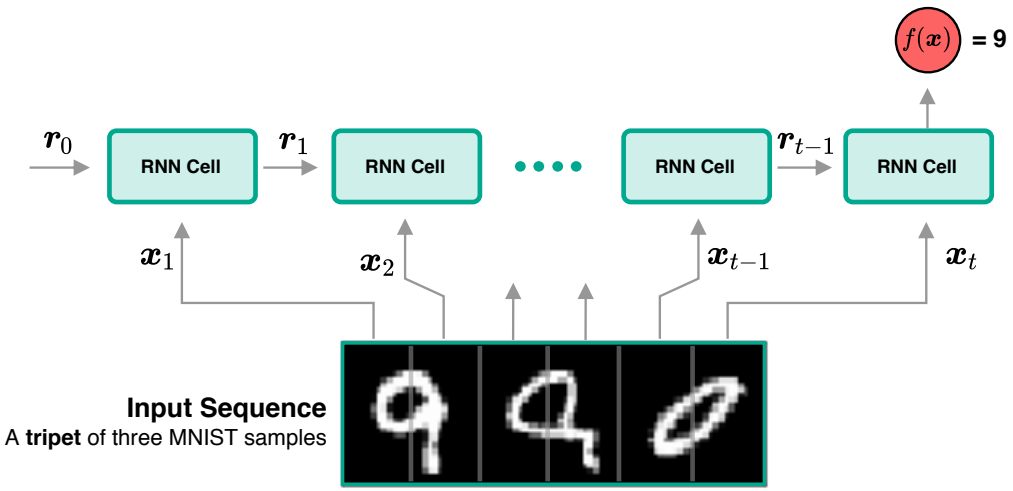
Architecture: CONV(5 × 5, 10) · AVGPOOL(2 × 2, 2, 2) · CONV(5 × 5, 25) · AVGPOOL(2 × 2, 2, 2) · CONV(4 × 4, 100)· AVGPOOL(2 × 2, 2, 2) · DENSE(100) · DENSE(10)

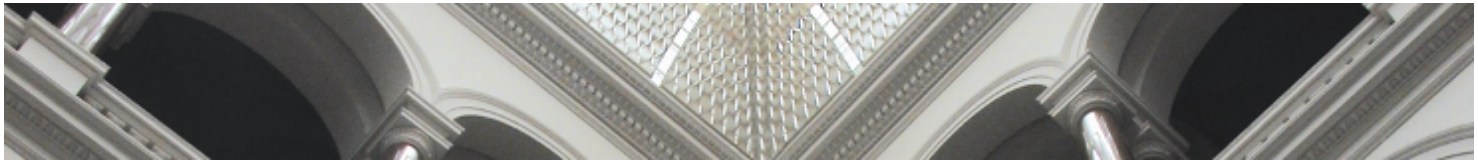


Technische Universität Berlin

Experimental Setup

- **Problem:** Majority-Sample Sequence Classification
  - ▷ Minimum accuracy 98% for MNIST and 89% for FashionMNIST
  - ▷ Sequence length: 12 ($\{\boldsymbol{x}\}_{t=1}^{12}$)

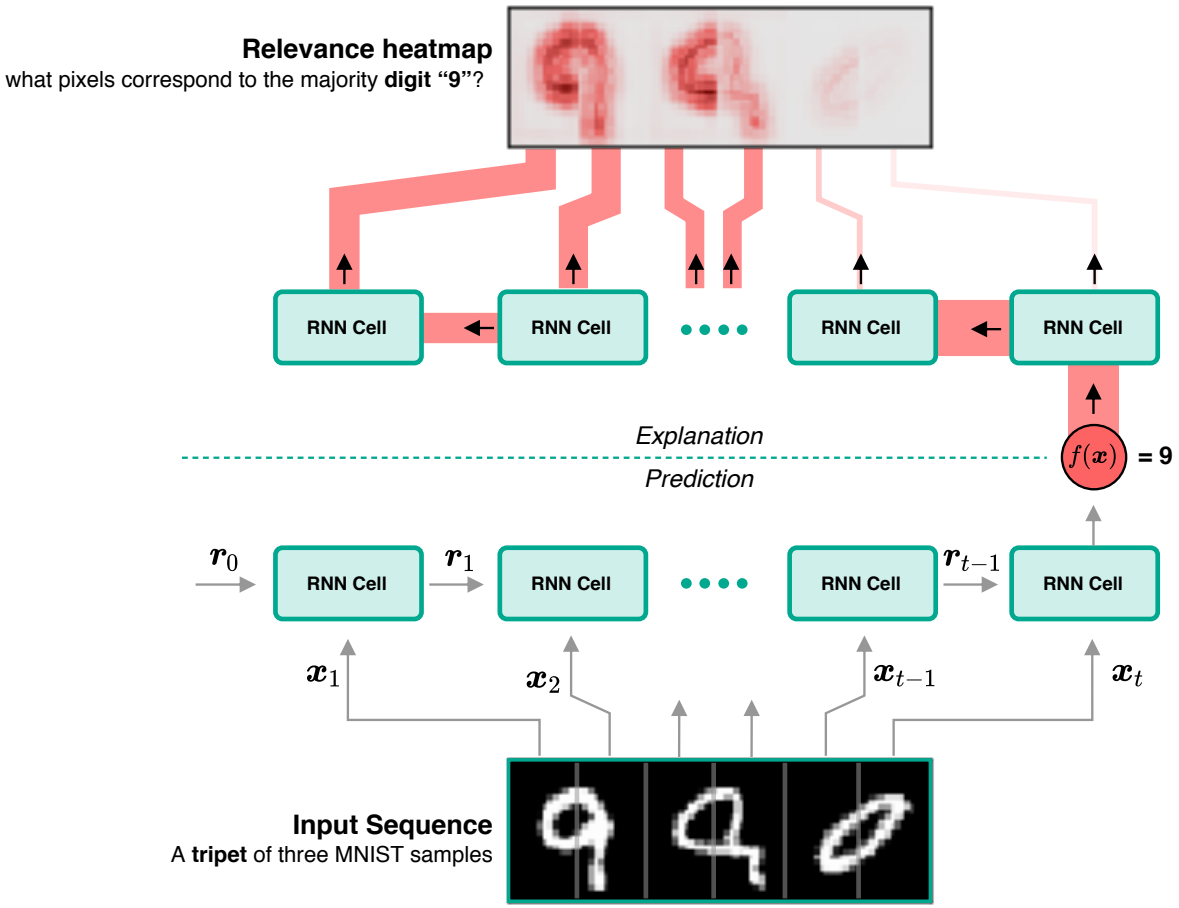


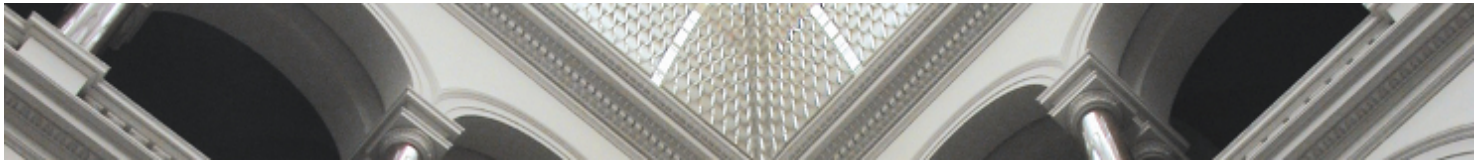


Technische Universität Berlin

Experimental Setup

– **Problem:** Majority-Sample Sequence Classification

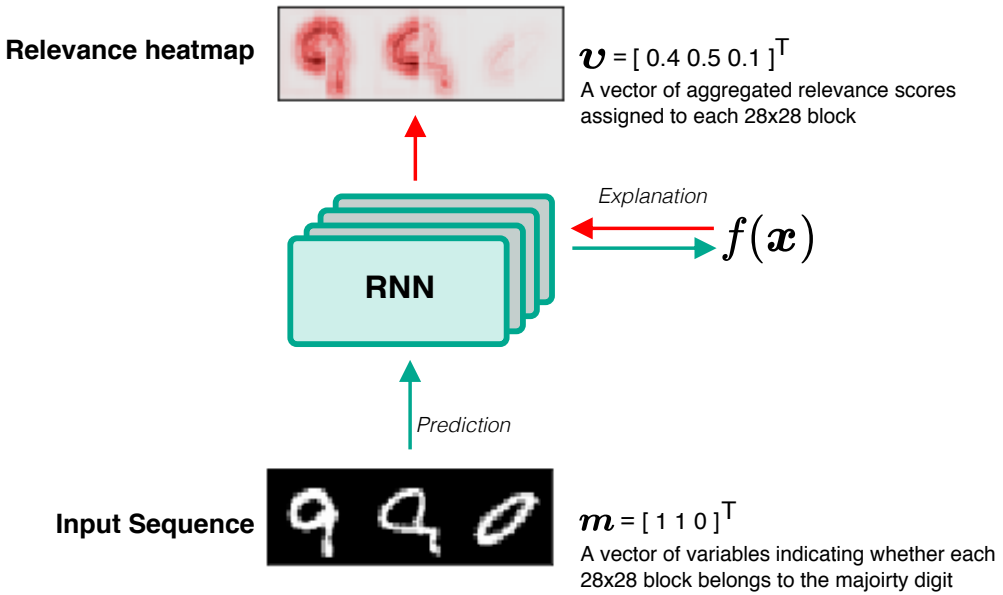




Technische Universität Berlin

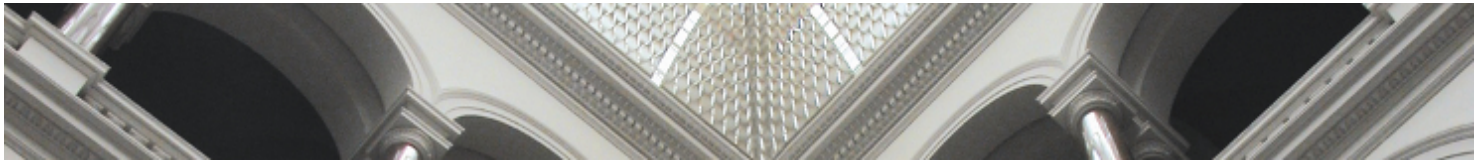
Experimental Setup

- **Problem:** Majority-Sample Sequence Classification
- **Evaluation:** Cosine similarity (averaged over test samples)



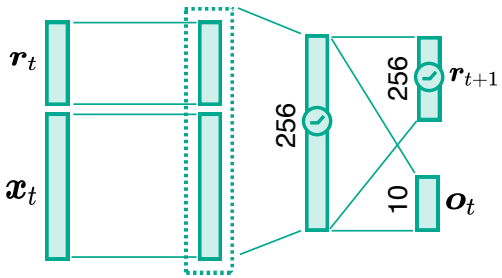
Heatmap Quality

$$\cos(\boldsymbol{m}, \boldsymbol{v}) = \frac{\boldsymbol{m}^T \boldsymbol{v}}{\|\boldsymbol{m}\| \|\boldsymbol{v}\|}$$

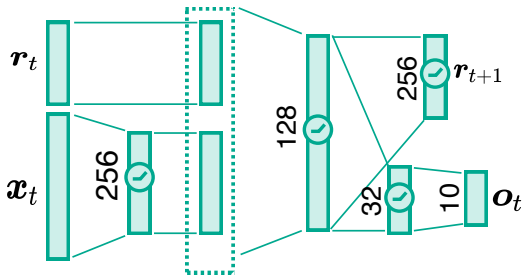


Technische Universität Berlin

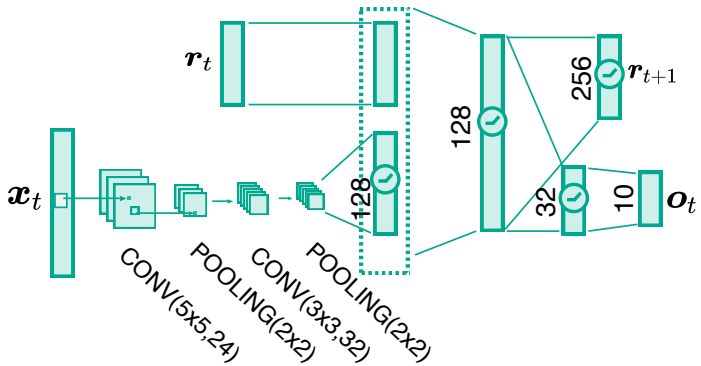
Experiment 1: Standard RNN Architectures



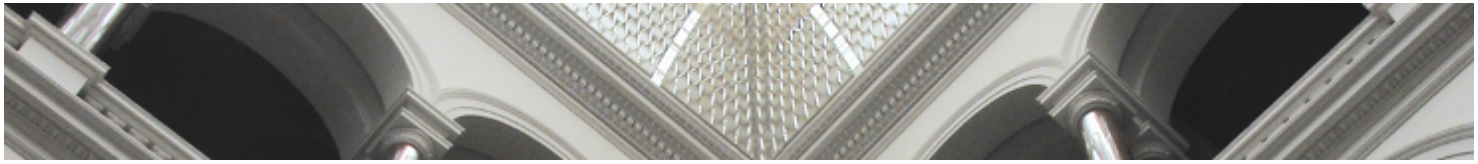
Shallow



Deep

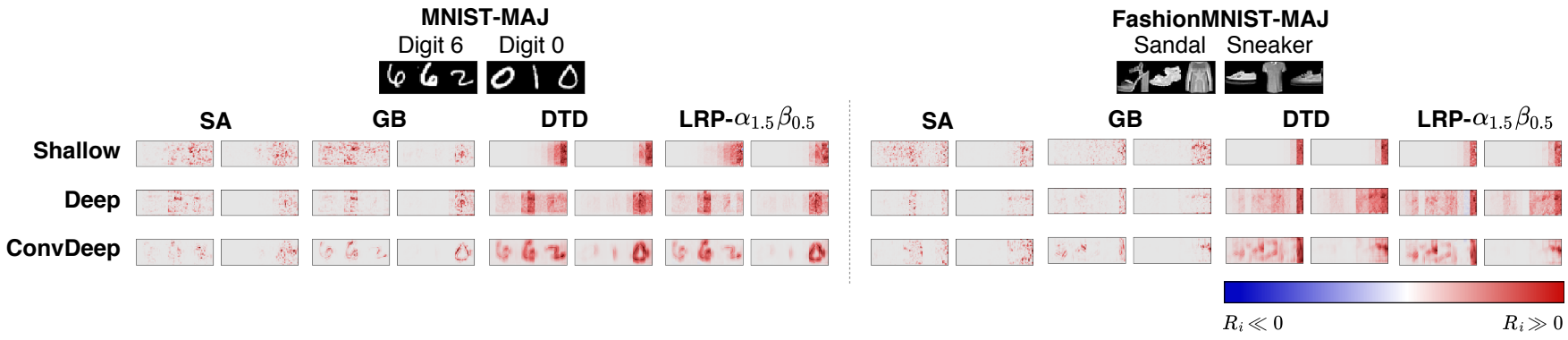


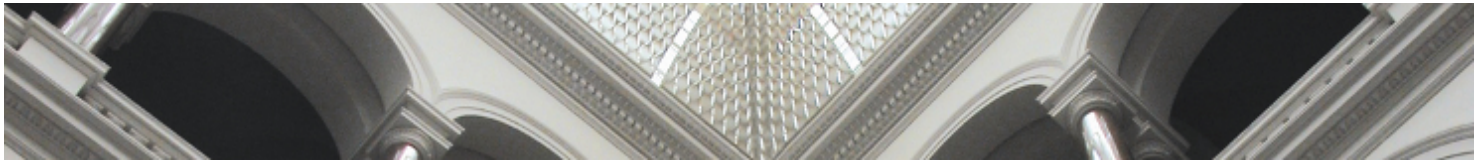
ConvDeep



Experiment 1: Standard RNN Architectures

Sample Relevance Heatmaps

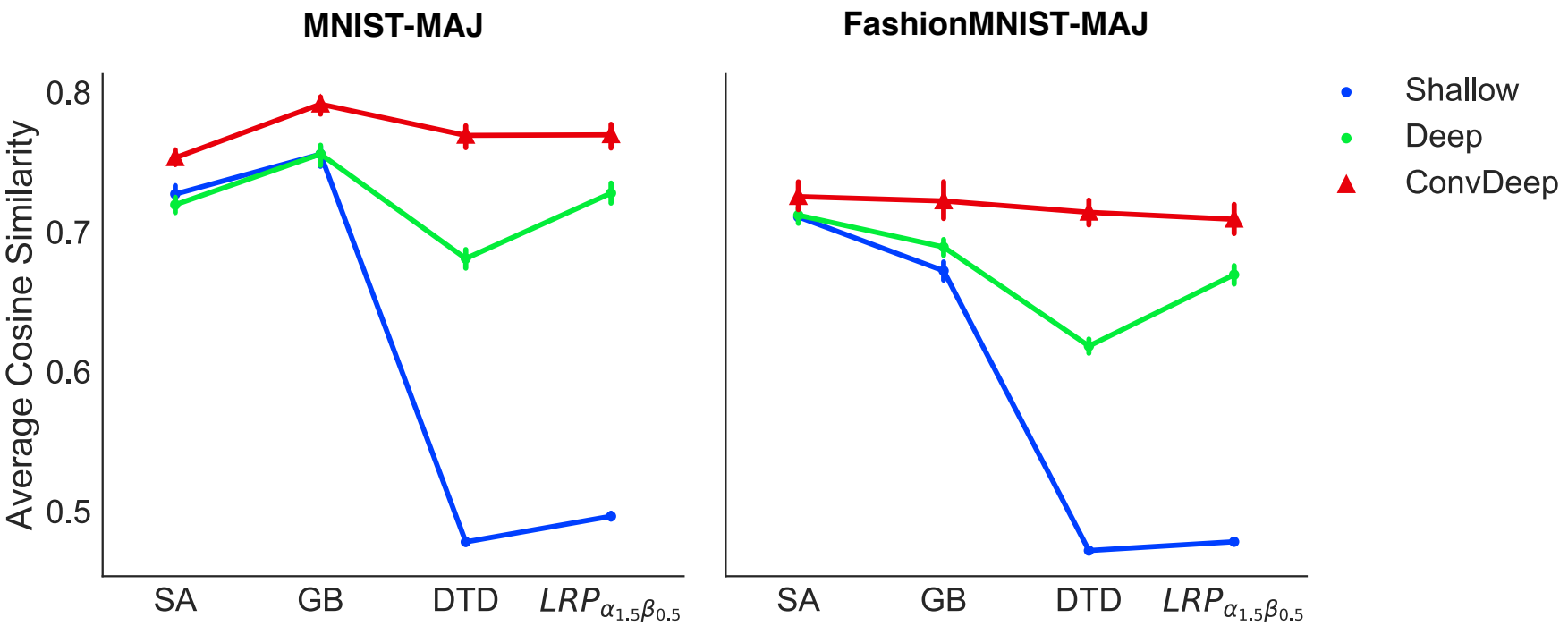


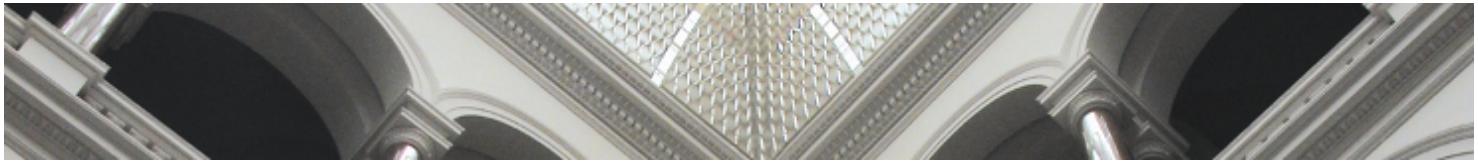


Technische Universität Berlin

Experiment 1: Standard RNN Architectures

Cosine Similarity Evaluation

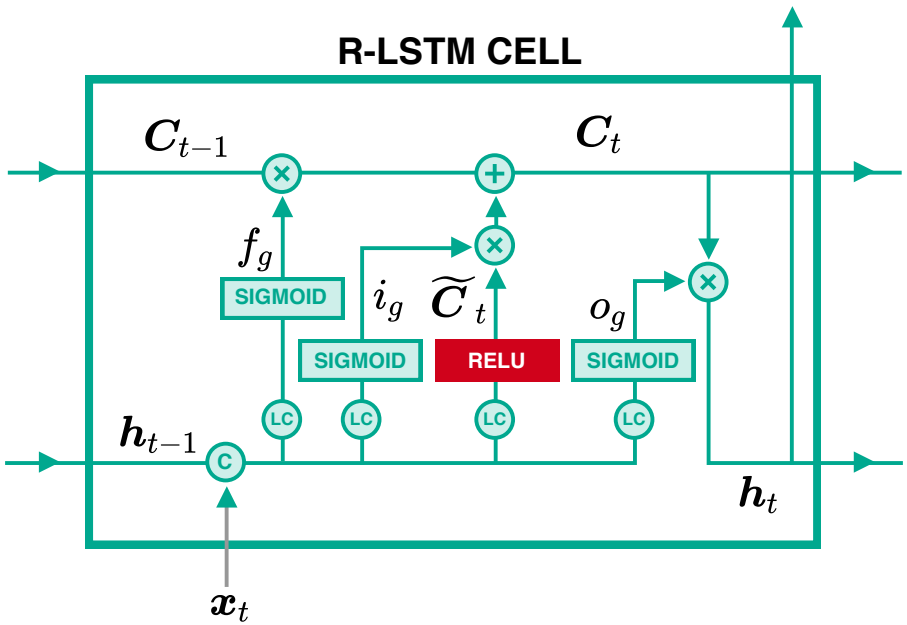




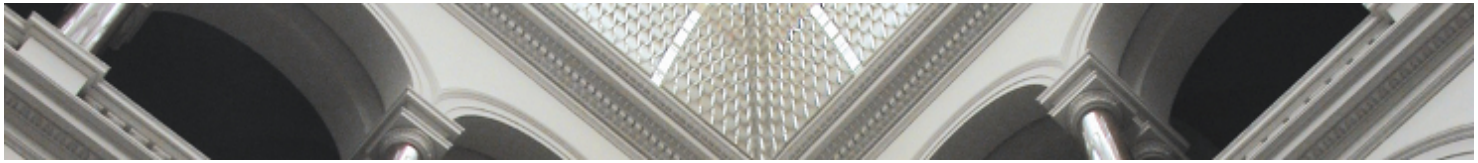
Technische Universität Berlin

Experiment 2: More Explainable Models

1. LSTM [HS97] with tanh being replaced by ReLU (R-LSTM)
   — Gating units are ignored during explaining [AMMS17].



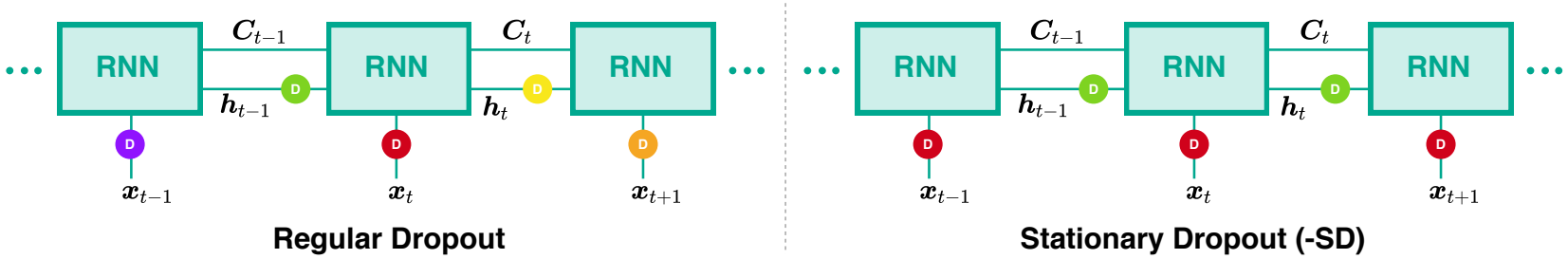
C Vector concatenation LC Linear combination ⊗ Element-wise multiplication ⊕ Element-wise addition

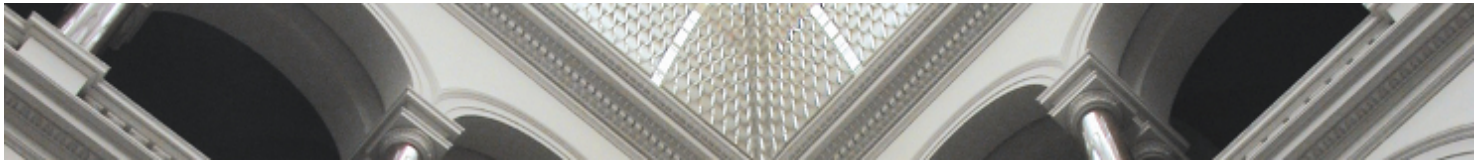


Technische Universität Berlin

Experiment 2: More Explainable Models

1. LSTM [HS97] with tanh being replaced by ReLU (R-LSTM)
2. Stationary Dropout [GG16] (-SD)

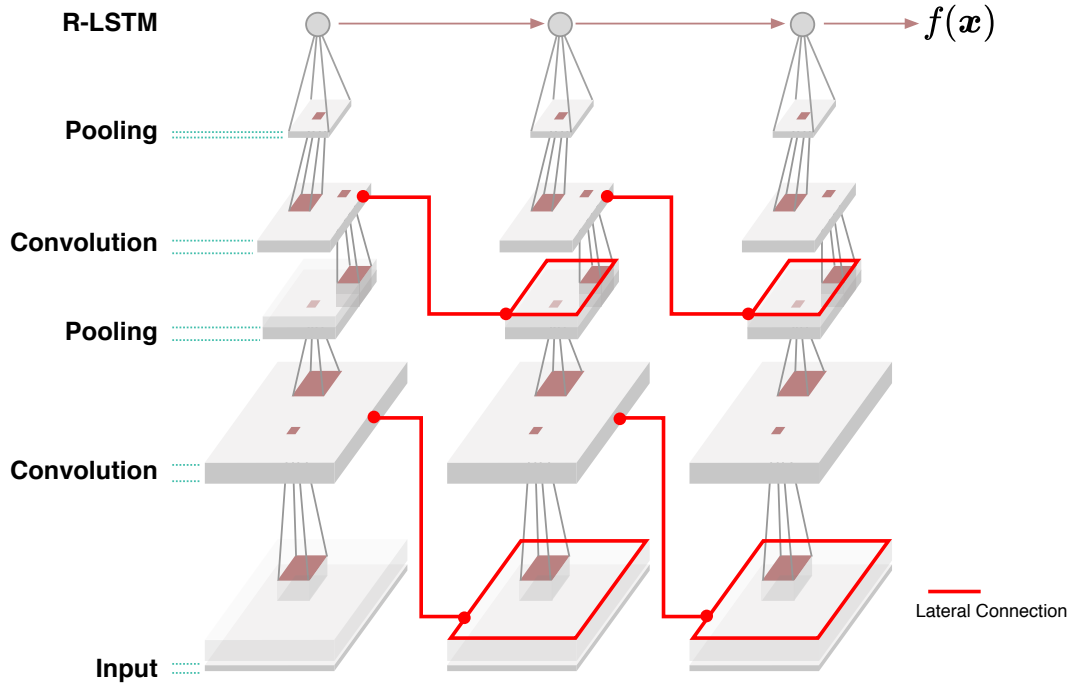


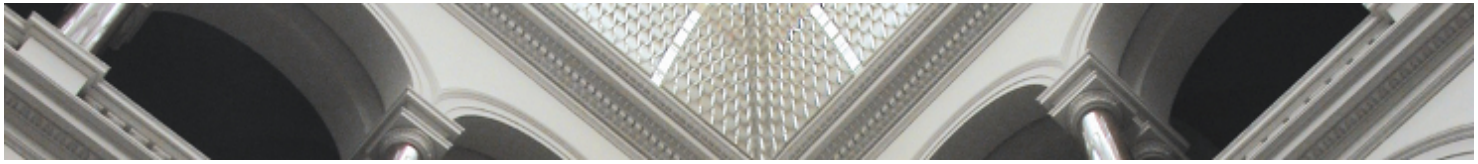


Technische Universität Berlin

Experiment 2: More Explainable Models

1. LSTM [HS97] with tanh being replaced by ReLU (R-LSTM)
2. Stationary Dropout [GG16] (-SD)
3. Lateral Connections for convolutional layers (Conv^{+})

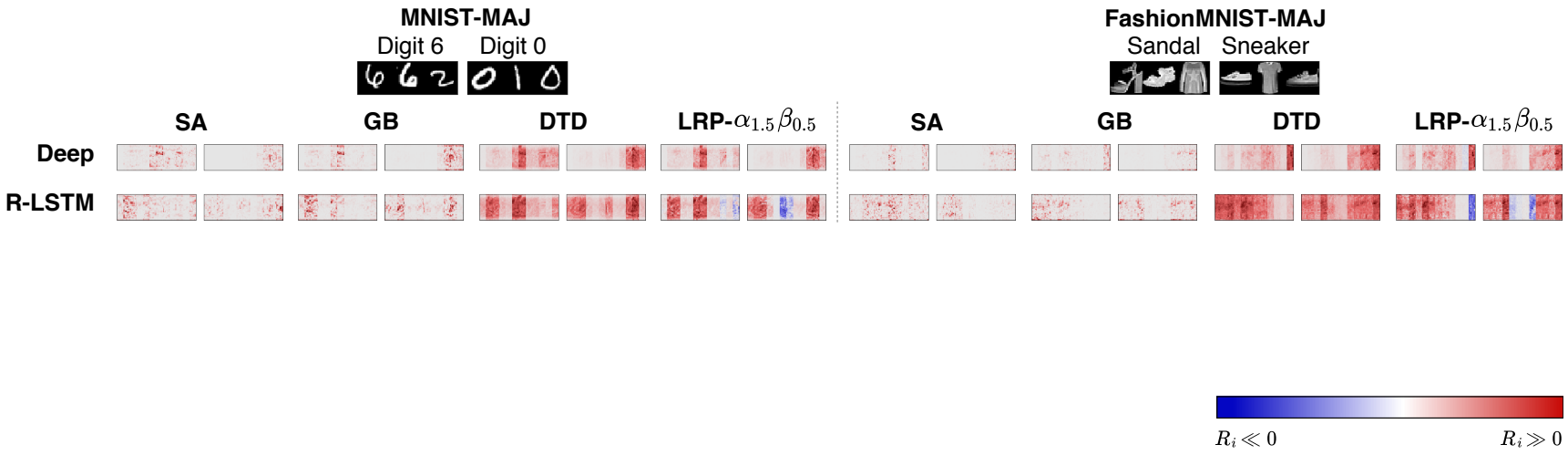


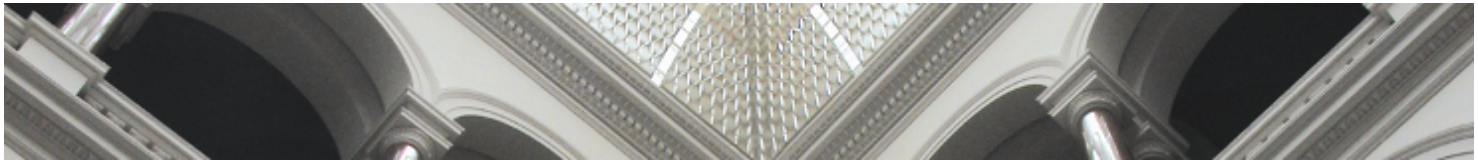


Technische Universität Berlin

Experiment 2: More Explainable Models

Sample Relevance Heatmaps

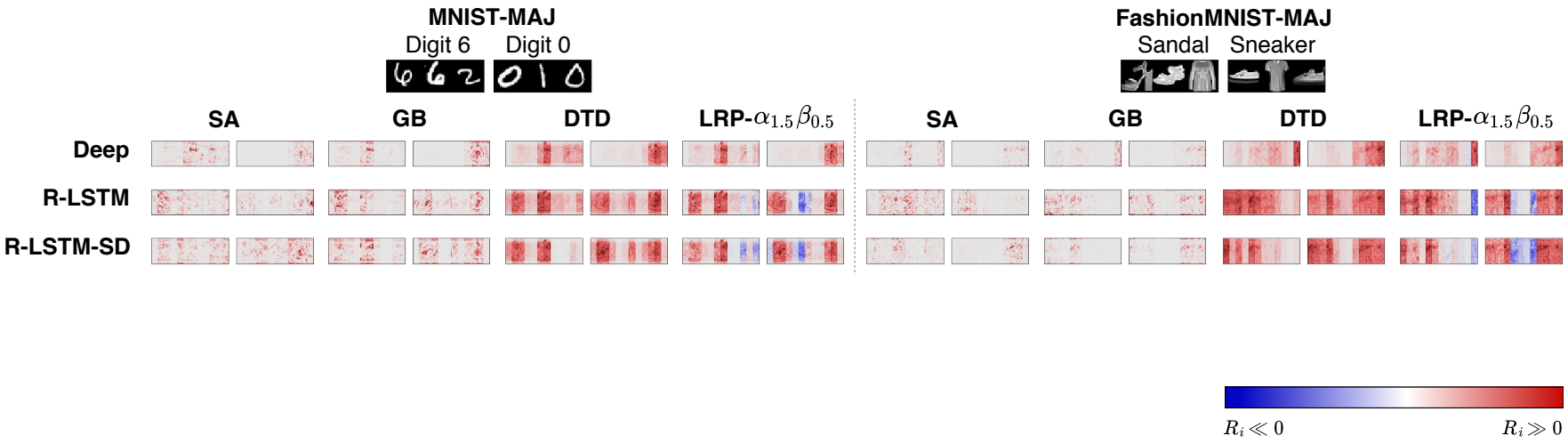


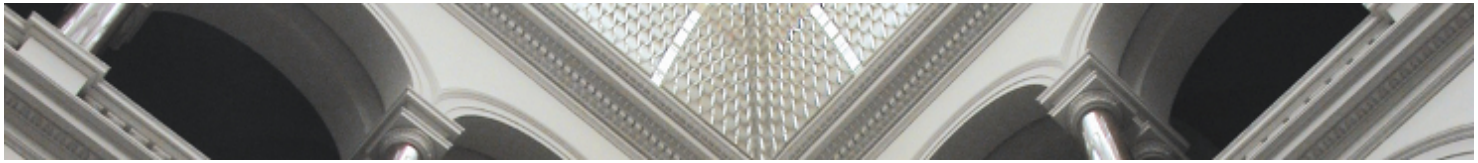


Technische Universität Berlin

Experiment 2: More Explainable Models

Sample Relevance Heatmaps

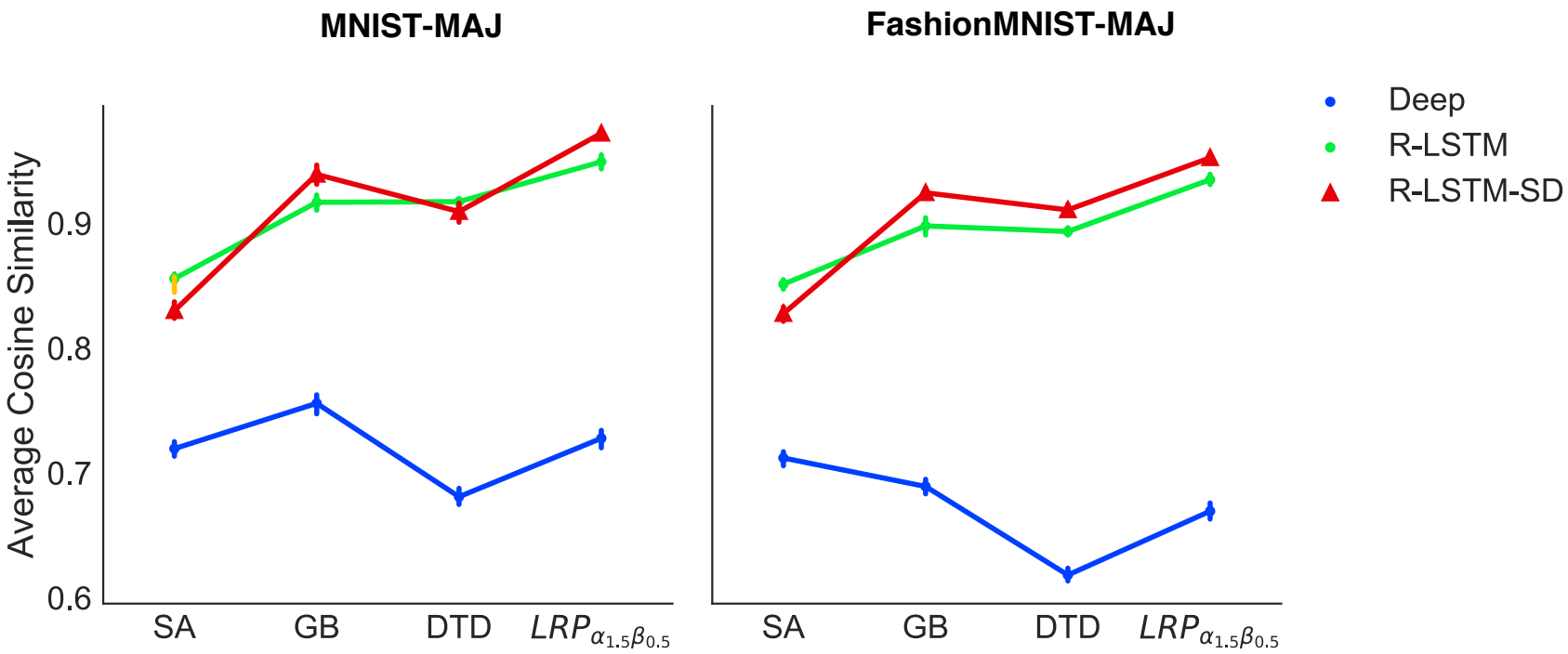


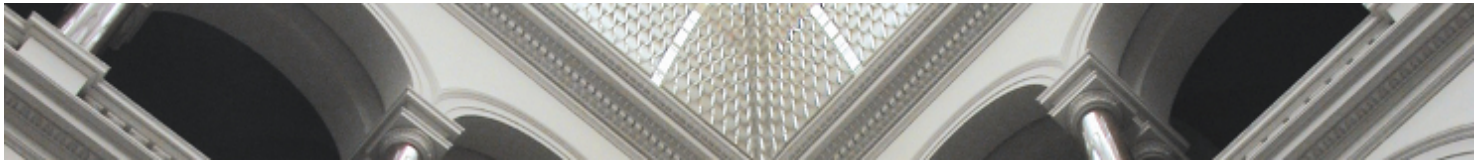


Technische Universität Berlin

Experiment 2: More Explainable Models

Cosine Similarity Evaluation

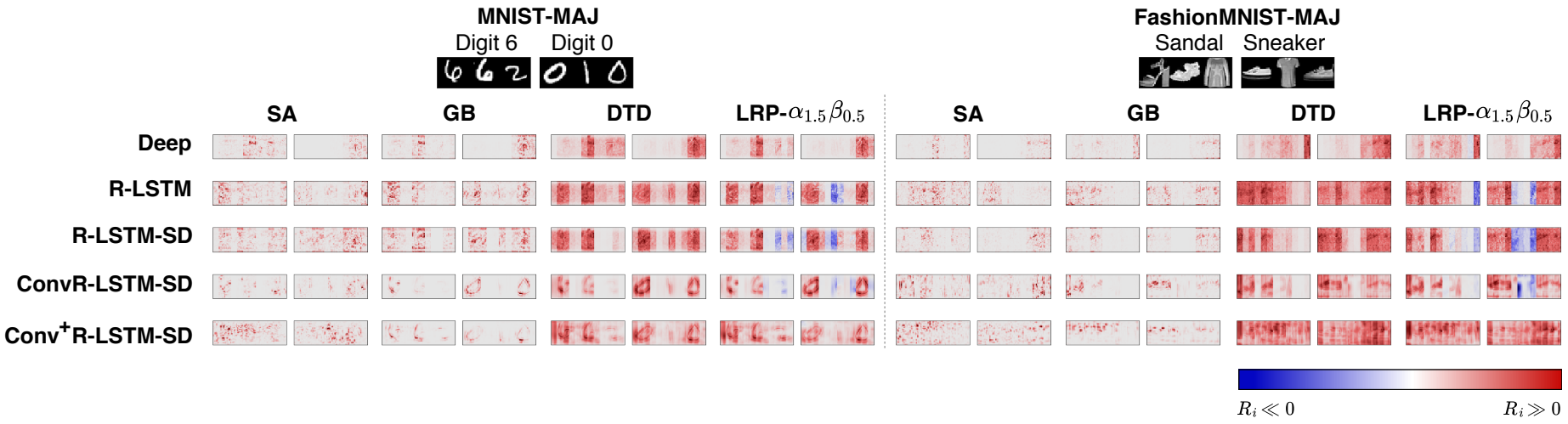


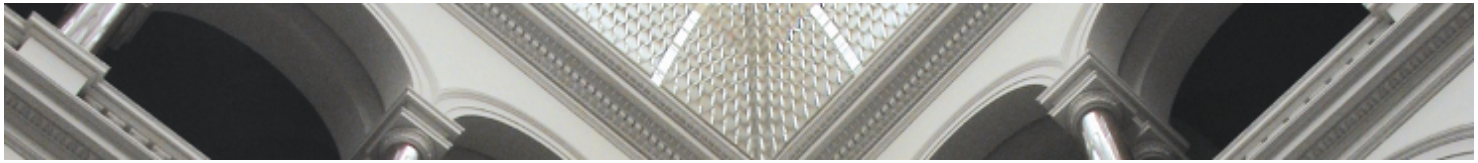


Technische Universität Berlin

Experiment 2: More Explainable Models

Sample Relevance Heatmaps

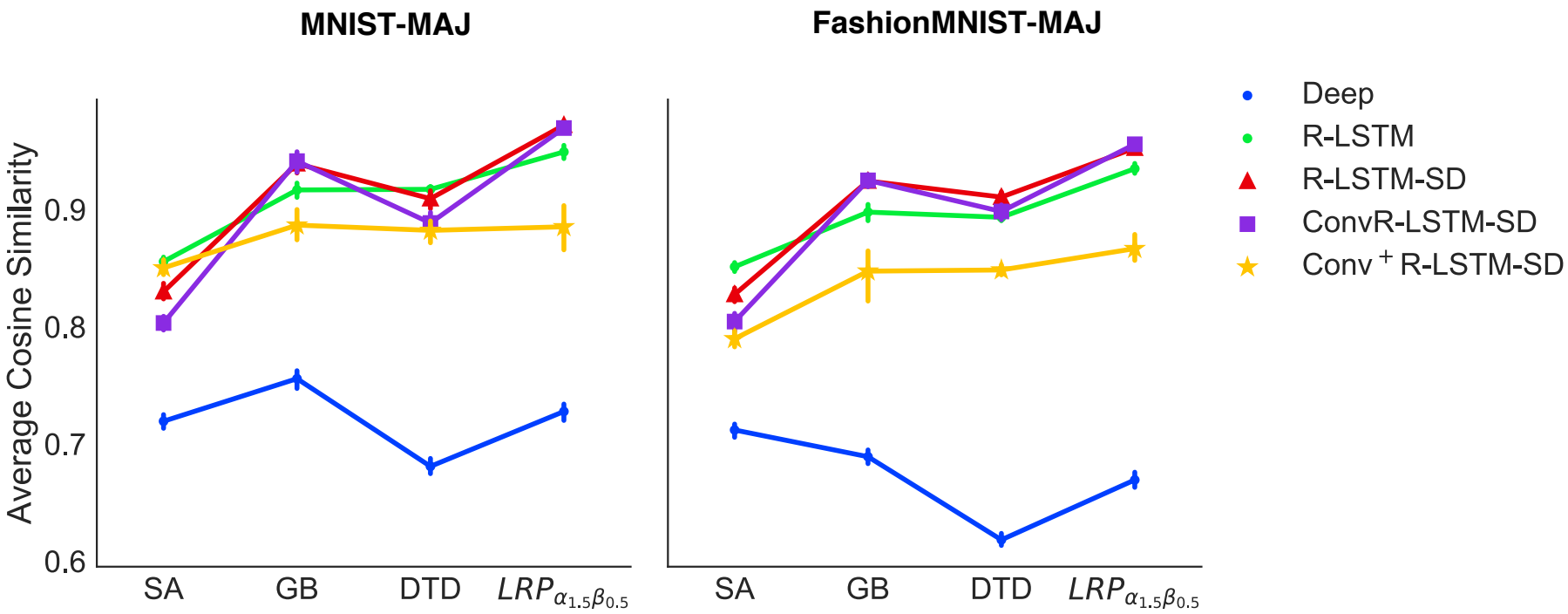


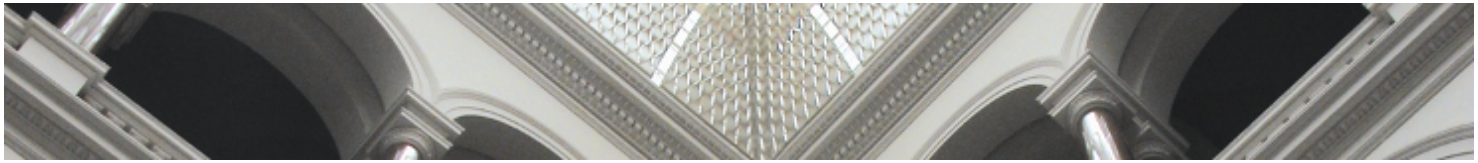


Technische Universität Berlin

Experiment 2: More Explainable Models

Cosine Similarity Evaluation

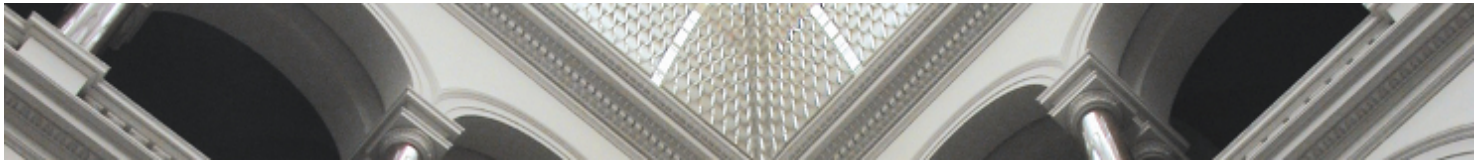




Technische Universität Berlin

Conclusion

- **Deeper and LSTM-type RNNs** are more explainable.
- **Stationary dropout** could improve model's explainability.
- Explainability of RNNs should be considered in two aspects:
  - **Coarse-gained**: relevance adequately propagated to the relevant input steps
    — related to recurrent mechanism (solution: LSTMs)
  - **Fine-gained**: soundness of each input's explanation.
    — related to the choice of input layers (i.e. convolutional layer for image data)

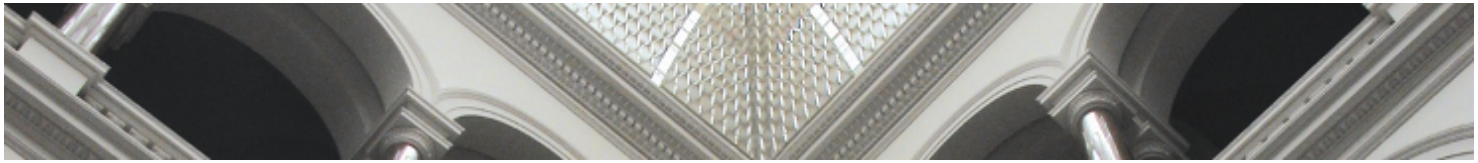


Technische Universität Berlin

References I

[AMMS17] Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek, **Explaining Recurrent Neural Network Predictions in Sentiment Analysis**, Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017 (Alexandra Balahur, Saif M. Mohammad, and Erik van der Goot, eds.), Association for Computational Linguistics, 2017, pp. 159–168.

[BML+16] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek, **Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers**, Artificial Neural Networks and Machine Learning – ICANN 2016, Lecture Notes in Computer Science, Springer, Cham, 2016, pp. 63–71.



Technische Universität Berlin

References II

[GG16] Yarin Gal and Zoubin Ghahramani, **A Theoretically Grounded Application of Dropout in Recurrent Neural Networks**, Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 1019–1027.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber, **Long Short-Term Memory**, Neural Computation **9** (1997), no. 8, 1735–1780.

[MLB+17] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller, **Explaining nonlinear classification decisions with deep Taylor decomposition**, Pattern Recognition **65** (May 1, 2017), 211–222.

[SDBR15] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, **Striving for Simplicity: The All Convolutional Net**, ICLR (Workshop Track), 2015.

[SVZ13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, **Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps**, CoRR **abs/1312.6034** (2013).