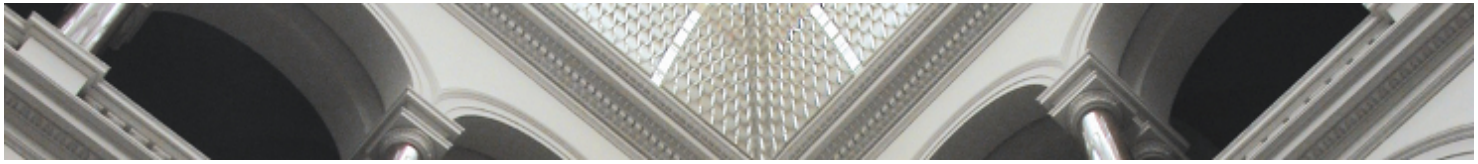
Technische Universität Berlin

Designing Recurrent Neural Networks for Explainability

Pattarawat Chormai | Technische Universität Berlin | 06/11/2018



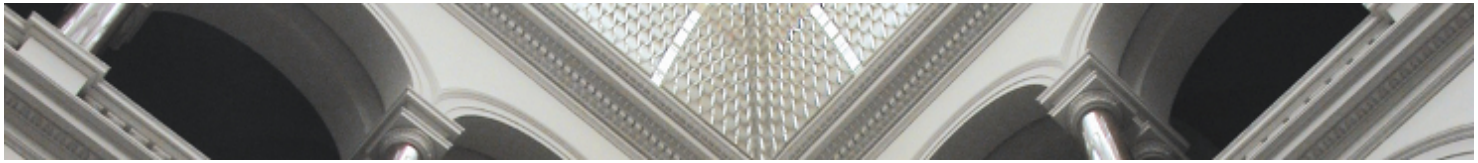
Technische Universität Berlin

Outline

Why should we care about explainability?

Explanation Methods

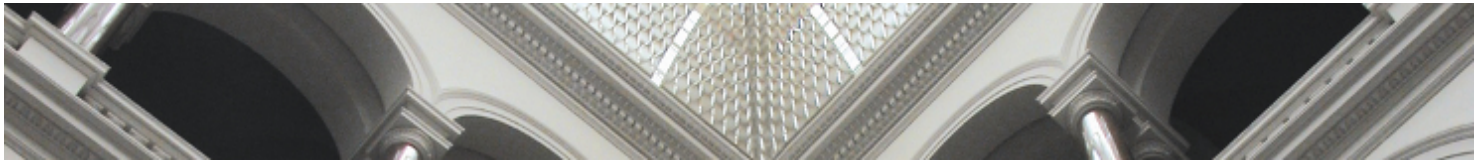
Conclusion



Technische Universität Berlin

Why should we care about explainability?

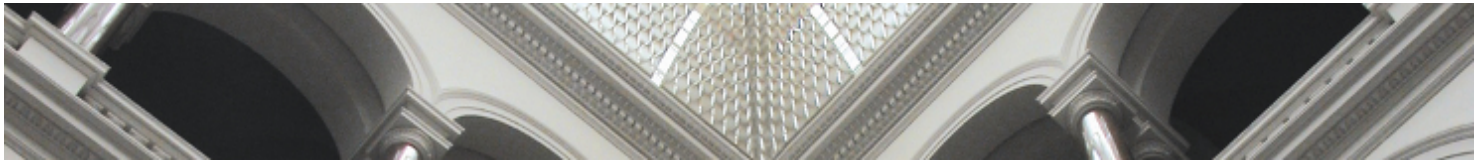
Machine Learning will soon involve in more critical applications, such as in heathcare. Thus, we need to make it accountable by being to investigate and explore how it works internally.t



Technische Universität Berlin

Explanation Methods

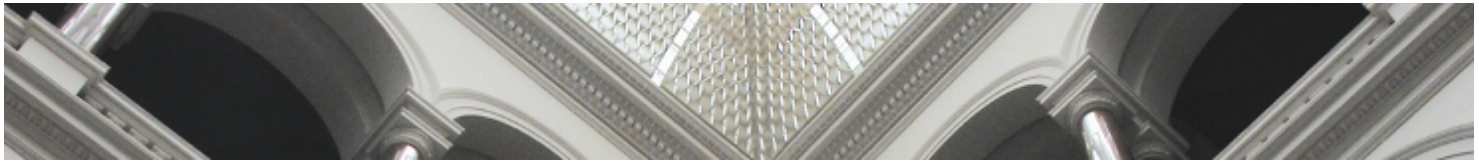
1. Sensitivity Analysis [SVZ13]



Technische Universität Berlin

Explanation Methods

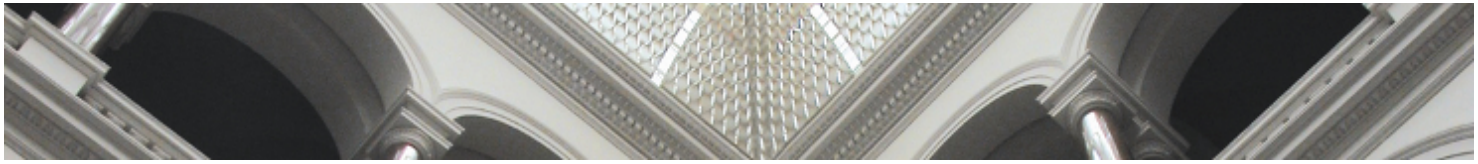
1. Sensitivity Analysis [SVZ13]
2. Guided Backprop [SDBR15]



Technische Universität Berlin

Explanation Methods

1. Sensitivity Analysis [SVZ13]
2. Guided Backprop [SDBR15]
3. LRP [BML$^+$16] and DeepTaylor Decomposition [MLB$^+$17]

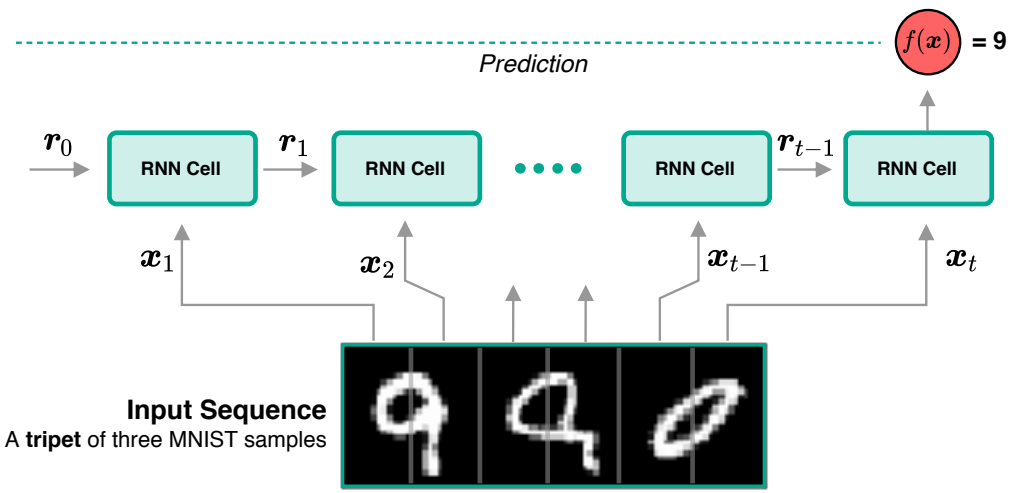


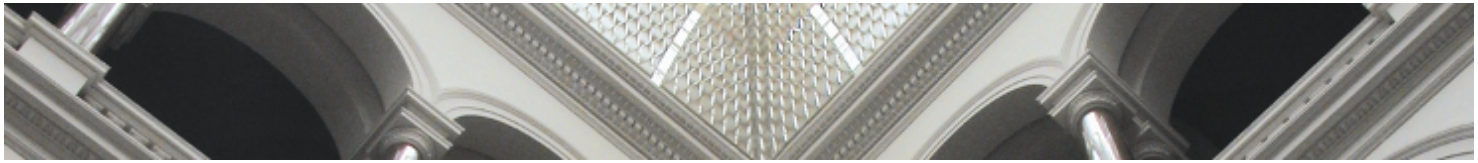
Technische Universität Berlin

Experimental Setup

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

We use Adam, Seq-length 12, archieve accuracy 98% for MNIST, 89% for FashionMNIST

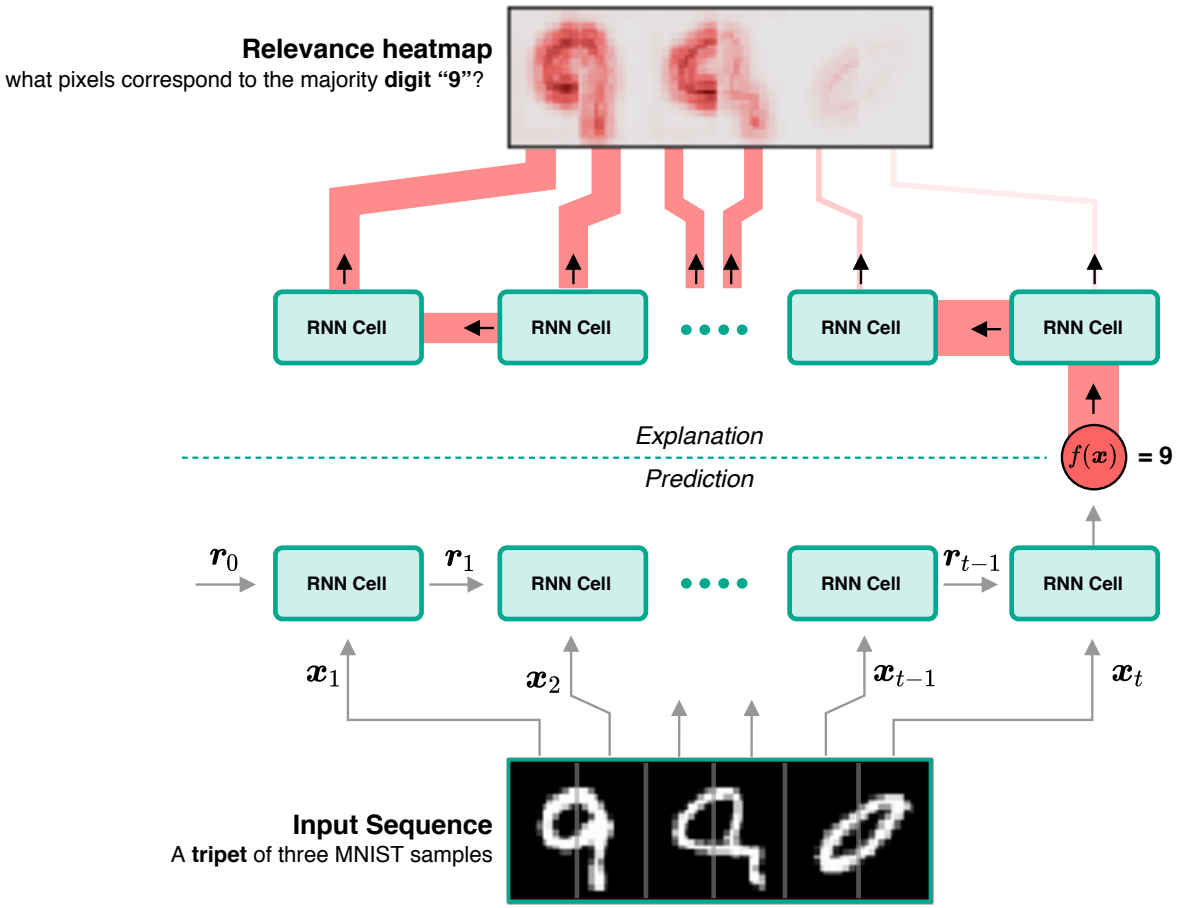


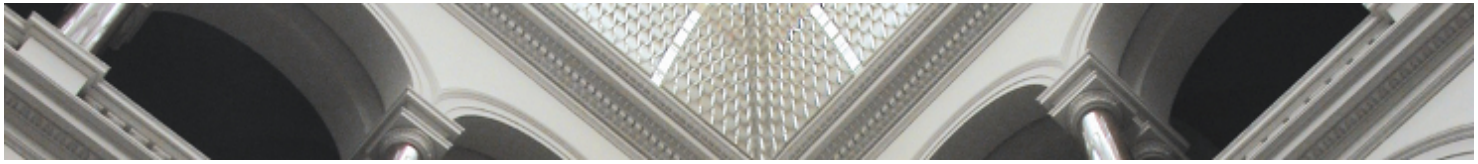


Technische Universität Berlin

Experimental Setup

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

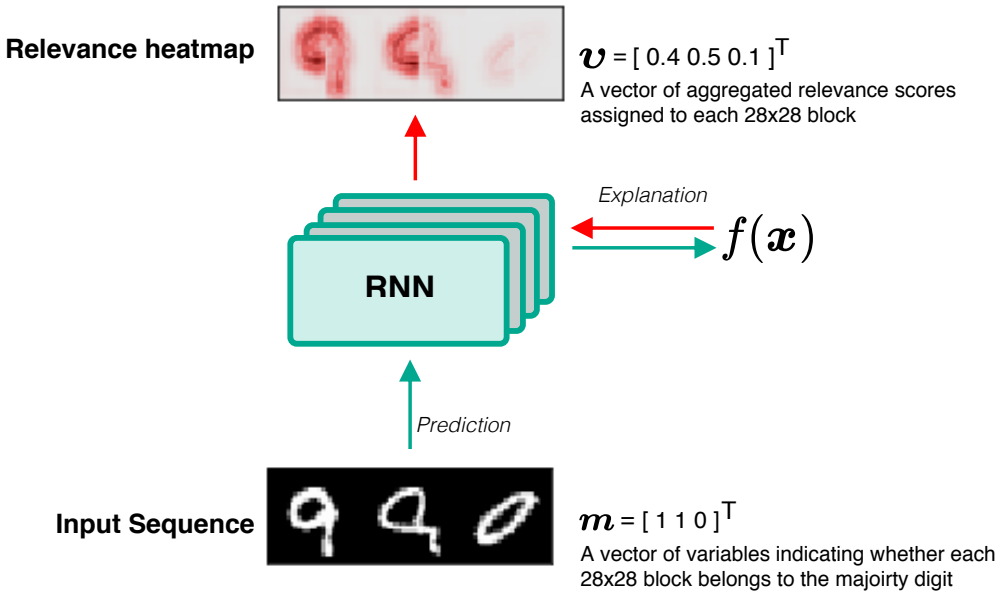




Technische Universität Berlin

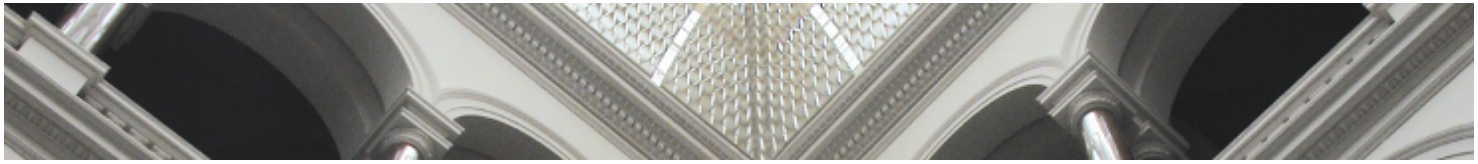
Experimental Setup

- **Problem:** Majority-Sample Sequence Classification
- **Evaluation:** Cosine Similarity

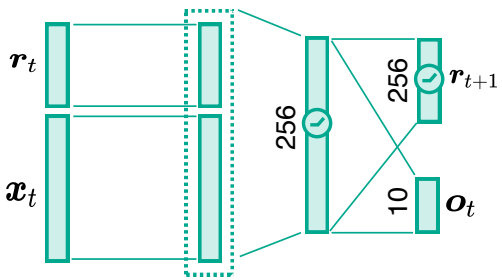


Heatmap Quality

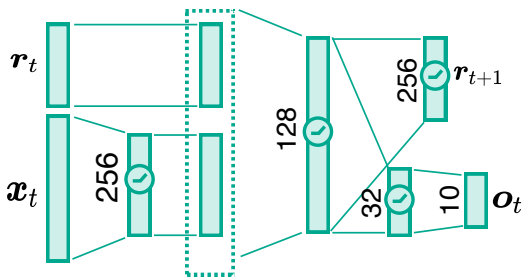
$$\cos(\boldsymbol{m}, \boldsymbol{v}) = \frac{\boldsymbol{m}^T \boldsymbol{v}}{\|\boldsymbol{m}\| \|\boldsymbol{v}\|}$$



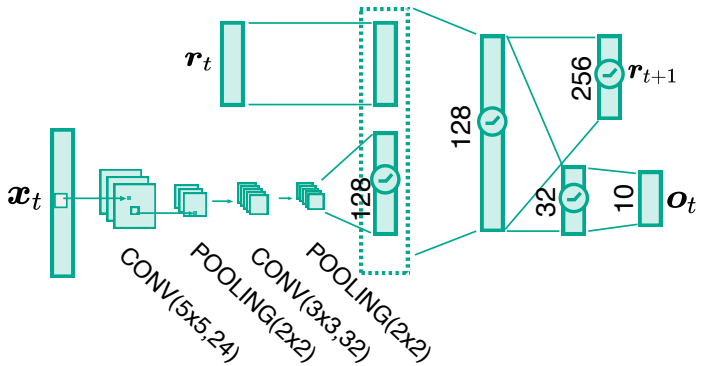
Experiment 1: Standard RNN Architectures



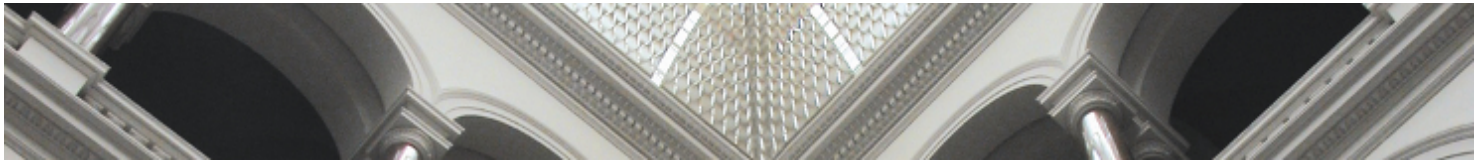
Shallow



Deep

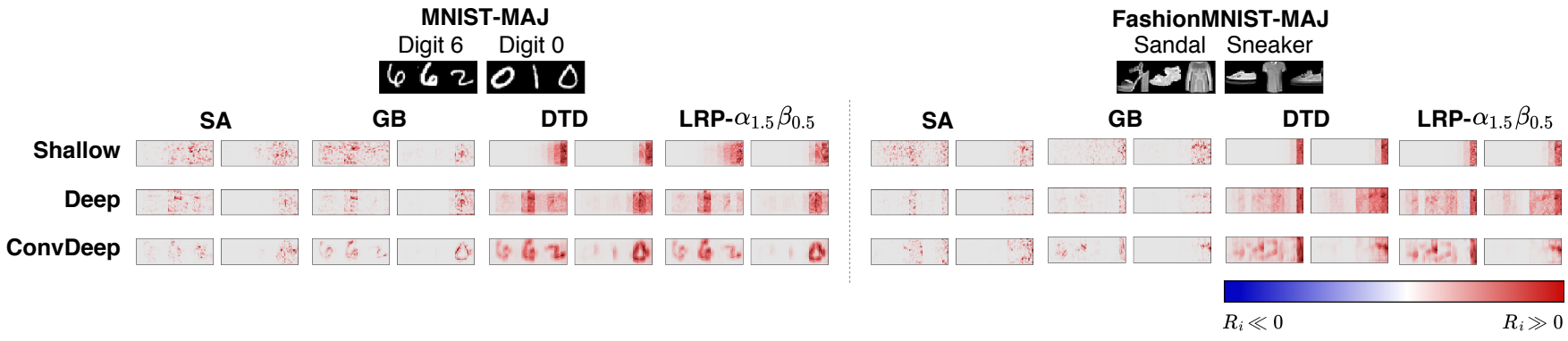


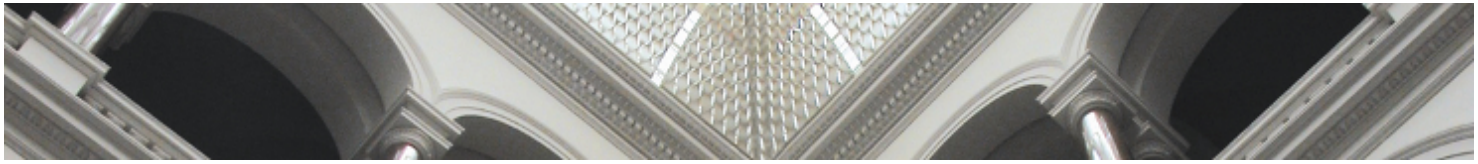
ConvDeep



Experiment 1: Standard RNN Architectures

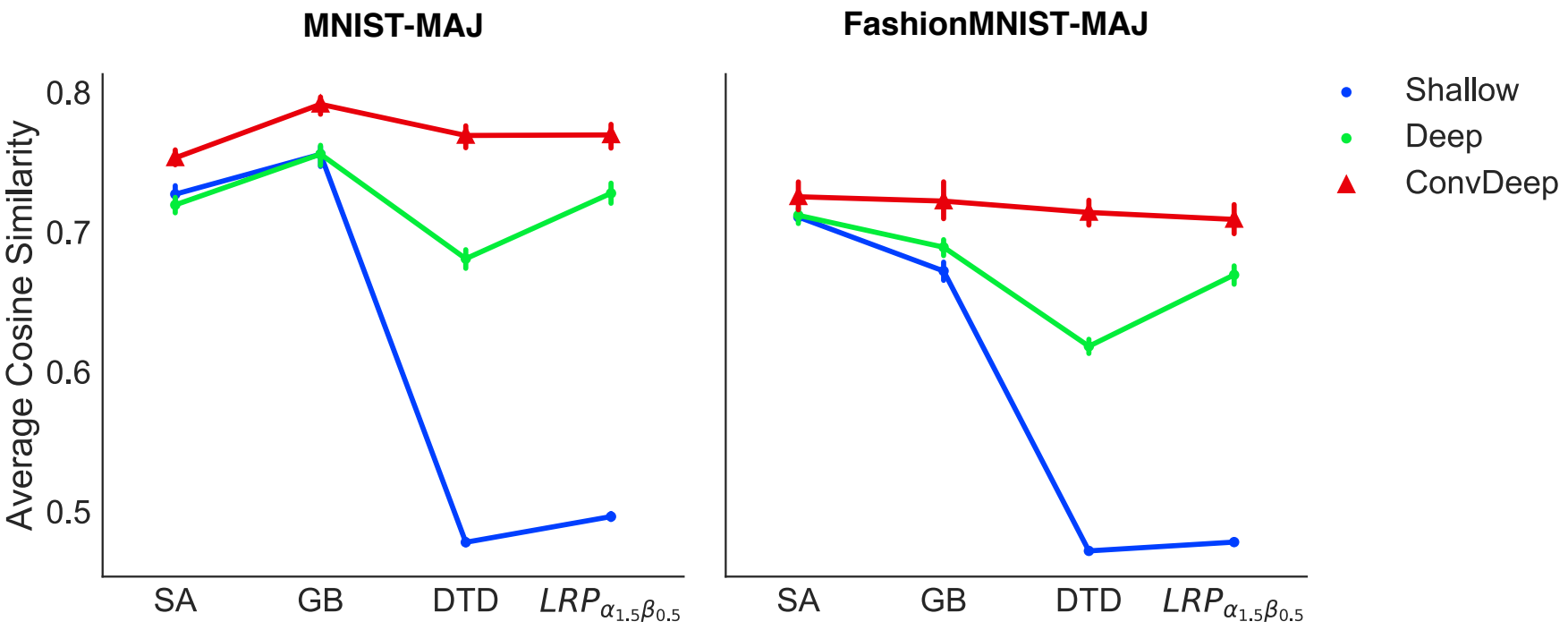
Sample Relevance Heatmaps

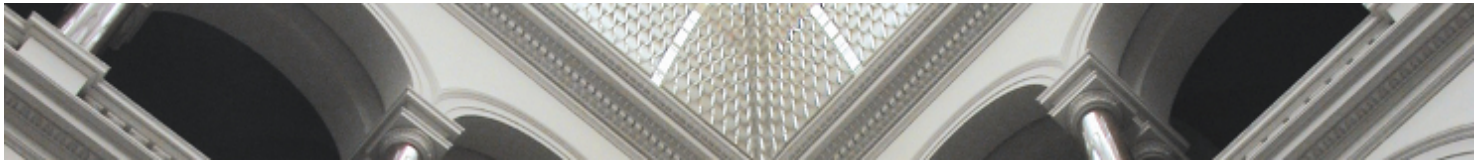




Experiment 1: Standard RNN Architectures

Cosine Similarity Evaluation

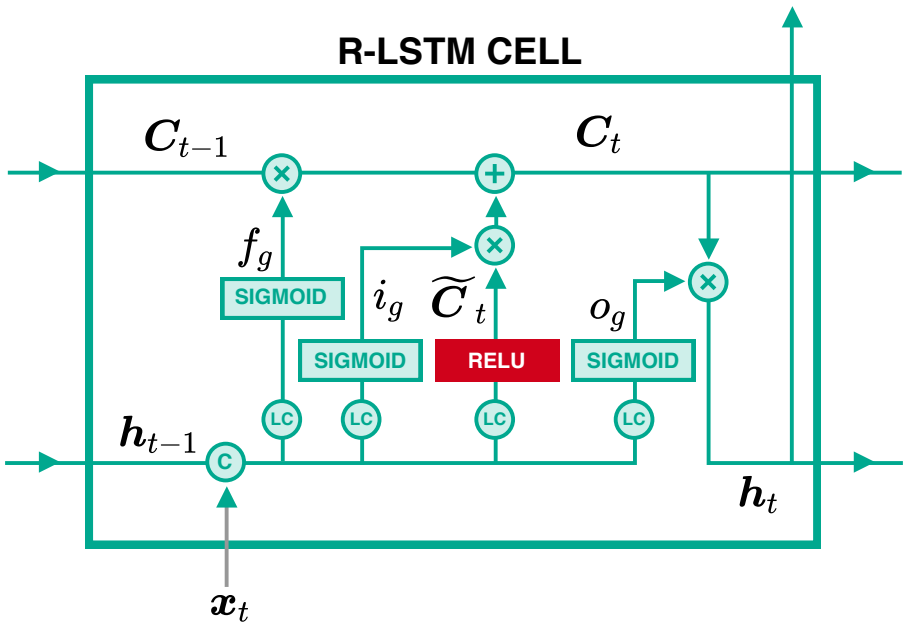




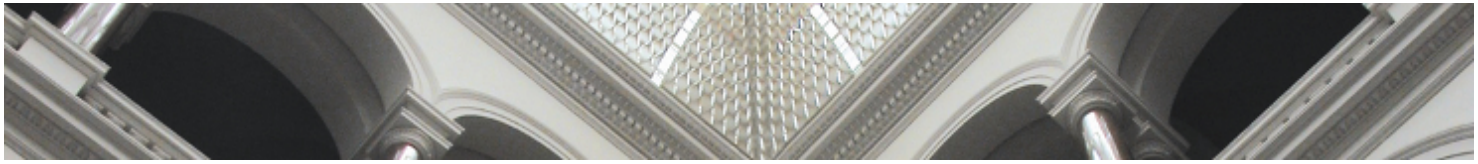
Technische Universität Berlin

Experiment 2: More Explainable Models

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



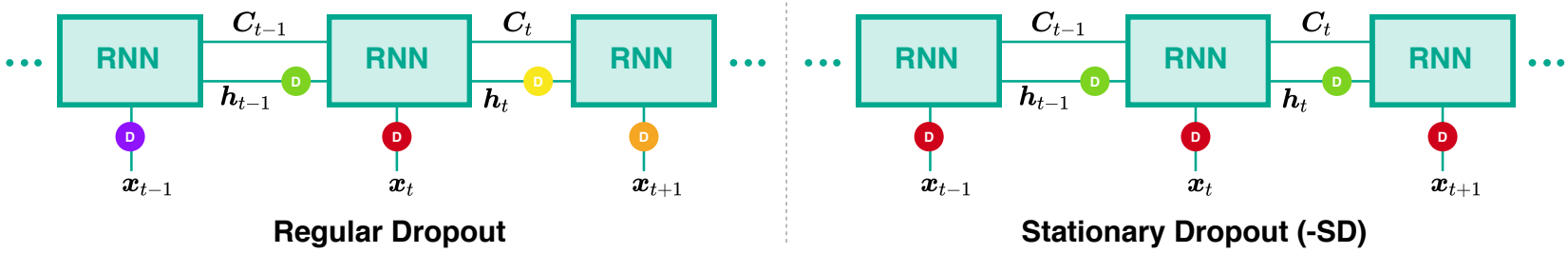
C Vector concatenation LC Linear combination ⊗ Element-wise multiplication ⊕ Element-wise addition

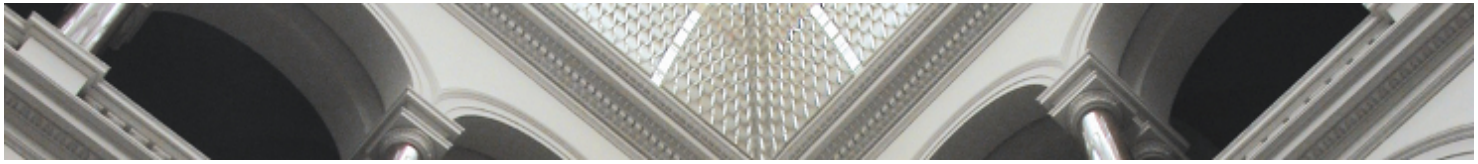


Technische Universität Berlin

Experiment 2: More Explainable Models

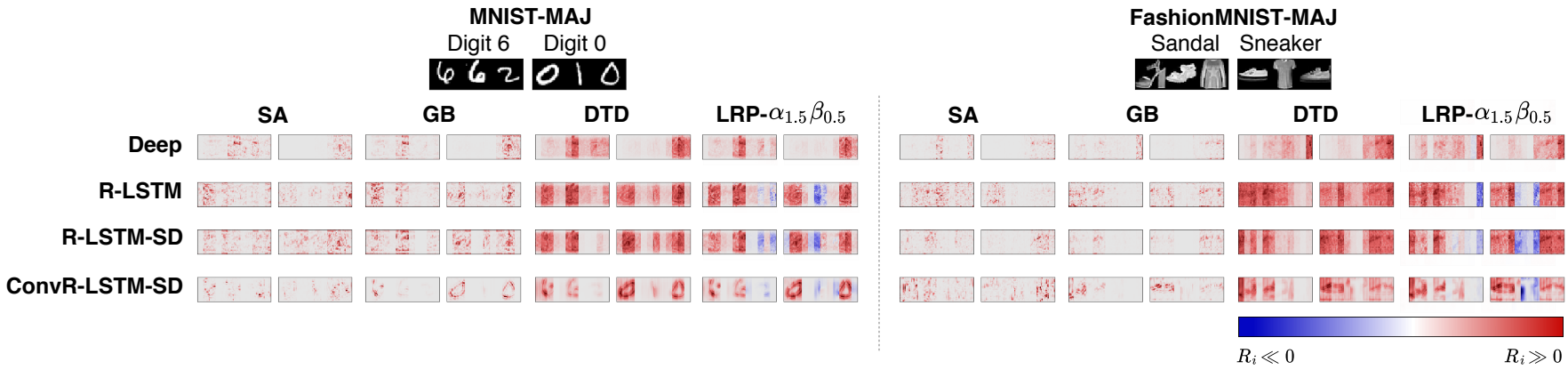
2. Stationary Dropout [GG16]

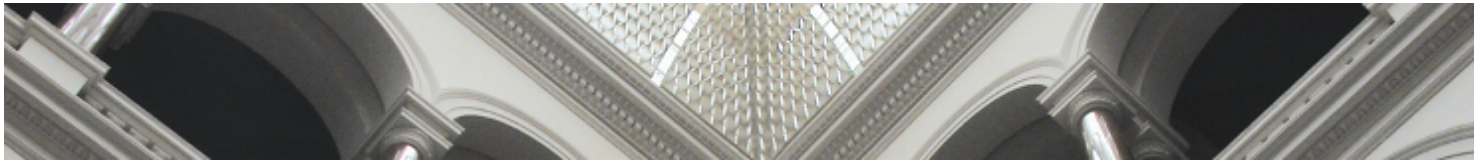




Experiment 2: More Explainable Models

Sample Relevance Heatmaps

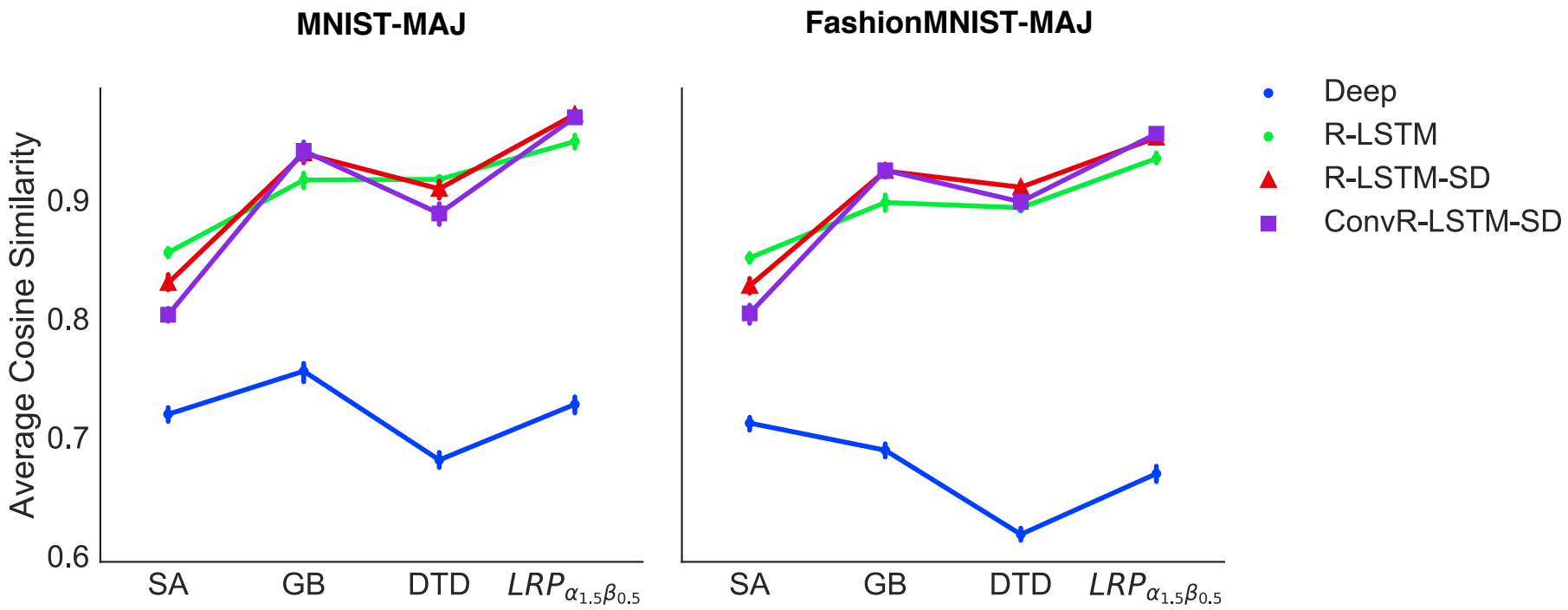


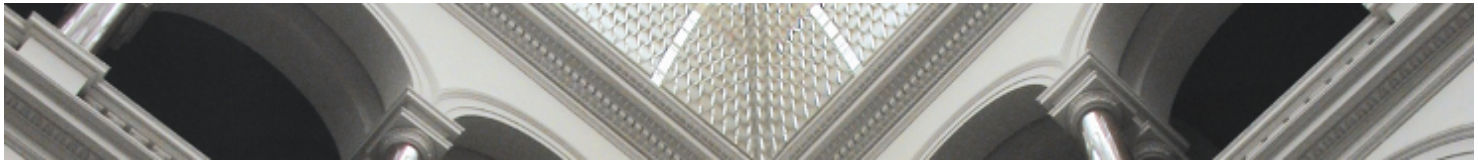


Technische Universität Berlin

Experiment 2: More Explainable Models

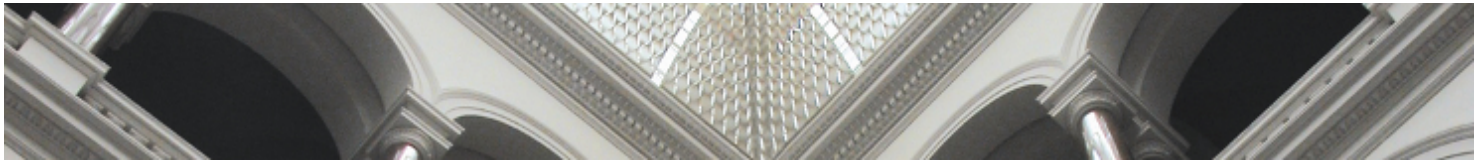
Cosine Similarity Evaluation





Technische Universität Berlin

Conclusions

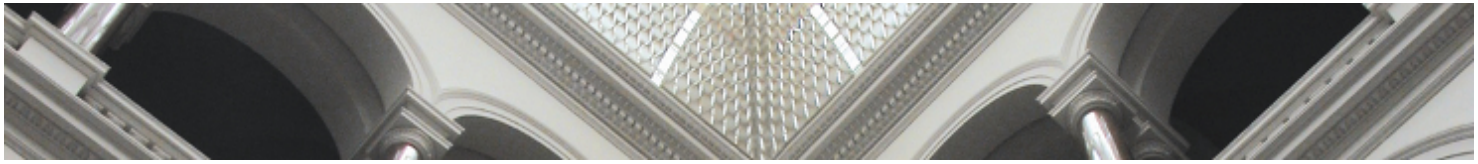


Technische Universität Berlin

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).