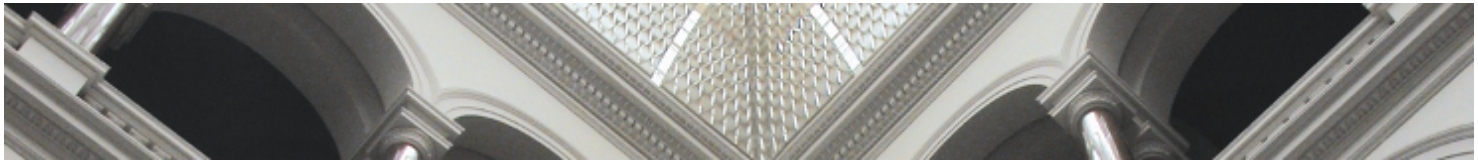
Technische Universität Berlin

Designing Recurrent Neural Networks for Explainability

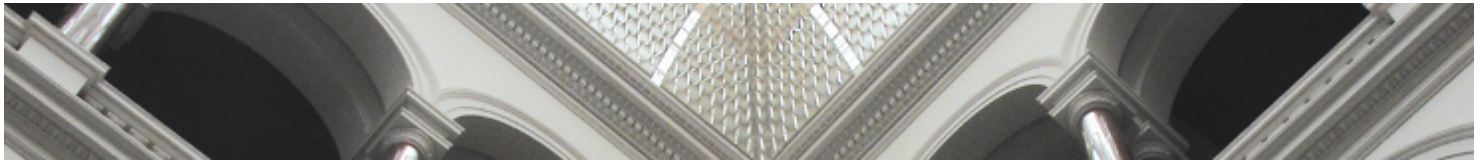
Pattarawat Chormai | Master's thesis supervised by Prof. Klaus-Robert Müller & Dr. Grégoire Montavon



Technische Universität Berlin

Motivation

- Impact of the architecture of RNNs on explainability
- Do deep RNNs have more explainable predictions?
- Are there ways to make RNNs more explainable?

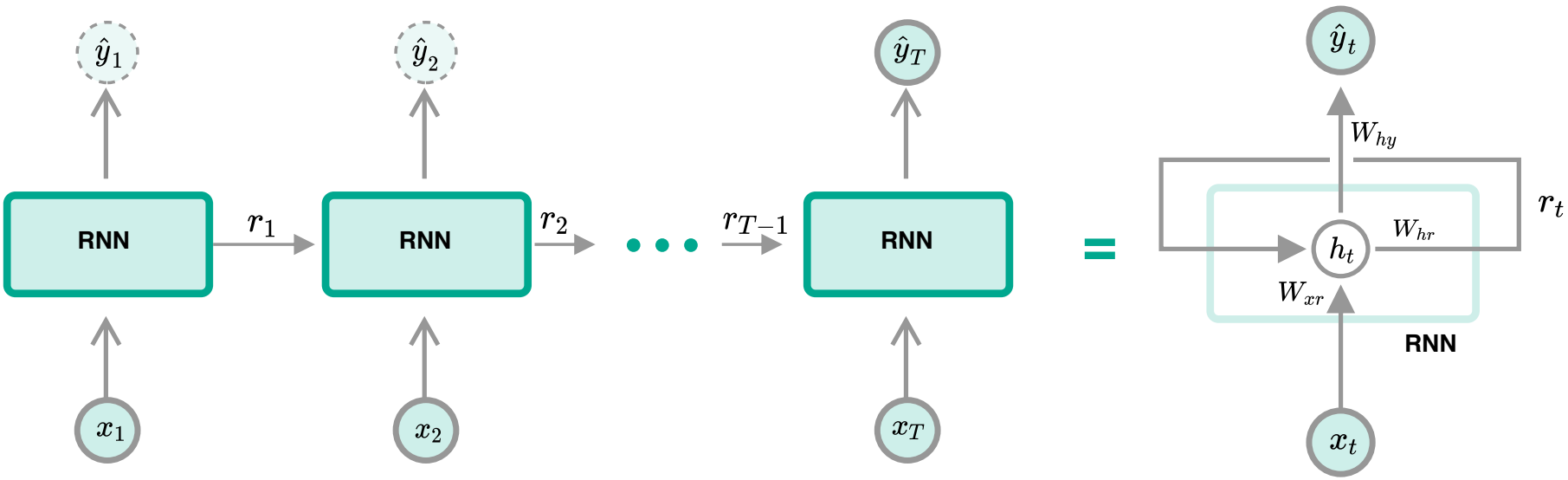


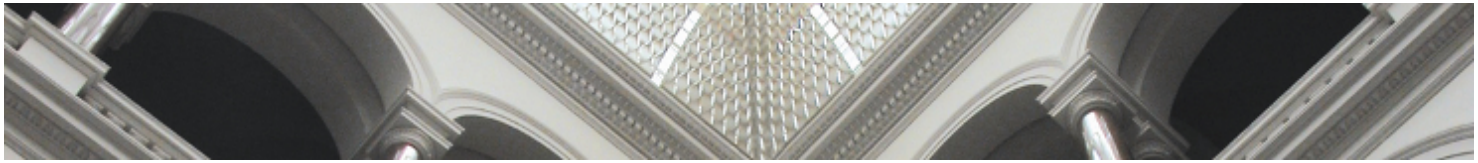
Technische Universität Berlin

Recurrent Neural Networks (RNNs)

– **Successful applications:**

▷ machine translation [WSC+16], image captioning [VTBE15, VSP+17]

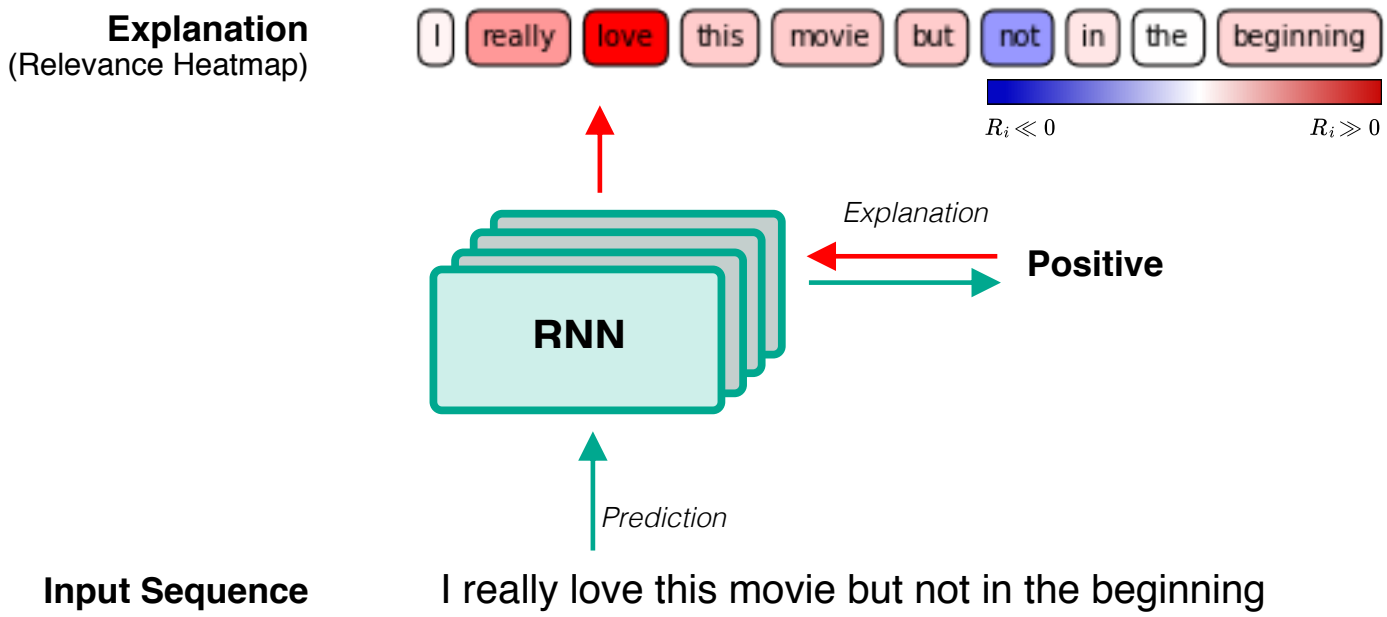


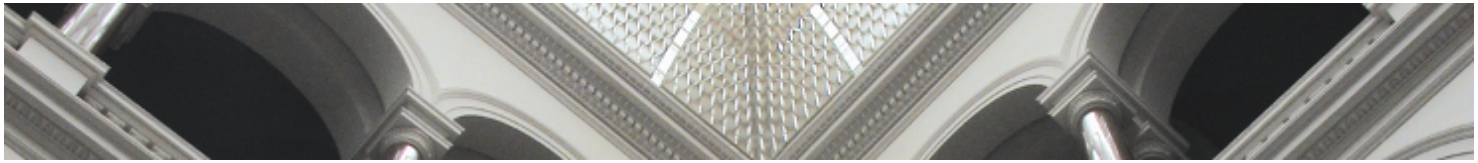


Technische Universität Berlin

Explainability

- Ability to provide **sensible explanation** towards how input associates to a certain prediction
- **Analogy:** why does the RNN classify this text as a positive review?





Technische Universität Berlin

Explanation Methods I

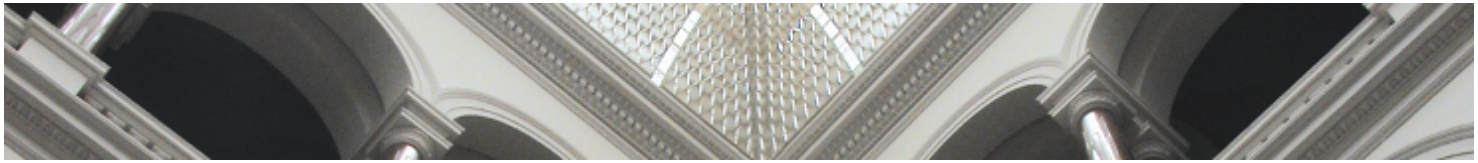
1. Sensitivity Analysis [SVZ13] (SA)

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

2. Guided Backprop [SDBR15] (GB)

   - Developed for explaining **ReLU**-based CNNs

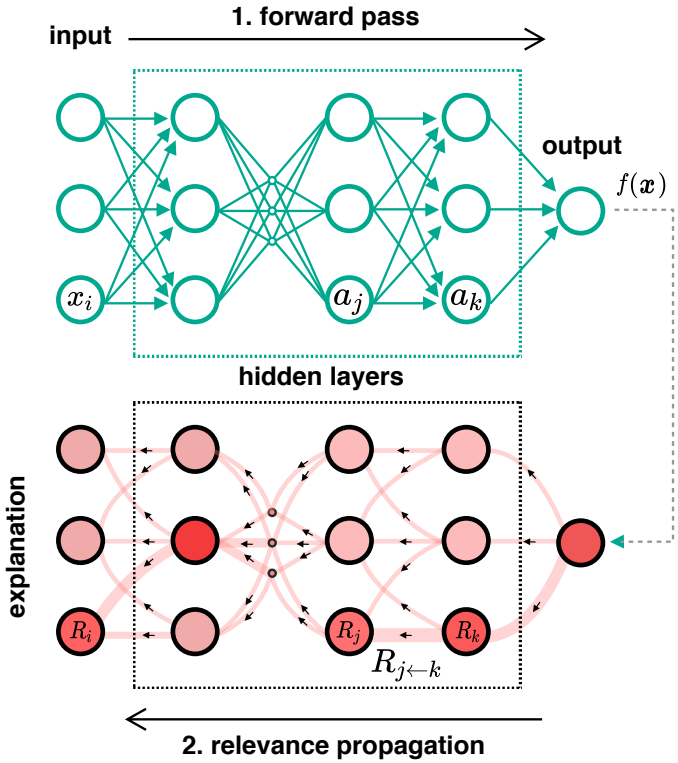
$$\frac{\partial_* f(\boldsymbol{x})}{\partial h_j} = \mathbb{1}\left[h_j > 0\right] \max\left(0, \frac{\partial_* f(\boldsymbol{x})}{\partial a_j}\right); \qquad R_i(\boldsymbol{x}) = \left(\frac{\partial_* f(\boldsymbol{x})}{\partial x_i}\right)^2$$



Technische Universität Berlin

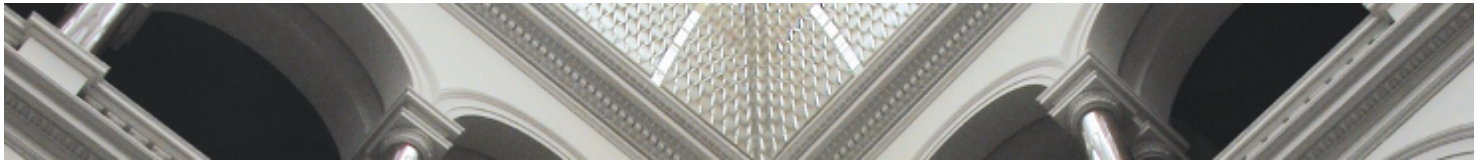
Explanation Methods II

3. Layer-wise Relevance Propagation [BML+16] (LRP)
   - Distributing relevance based on **neurons' activities**



LRP-$\alpha\beta$ rule:

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



Technische Universität Berlin

Explanation Methods III

4. Deep Taylor Decomposition [MLB$^+$17] (DTD)
   - LRP's theoretical view for explaining **ReLU-based** architectures

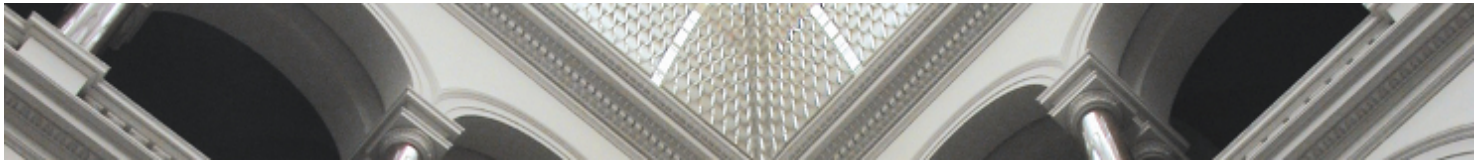
$$R_k = R_k\Big|_{\{\tilde{a}_j\}_j} + \sum_j \frac{\partial R_k}{\partial a_j}\Big|_{\{\tilde{a}_j\}_j} (a_j - \tilde{a}_j) + \xi_k \qquad \text{(Taylor expansion)}$$

   - Two important propagation rules:
     – z^+ for $a_j \in \mathbb{R}^+$ (LRP-$\alpha_1\beta_0$)

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

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

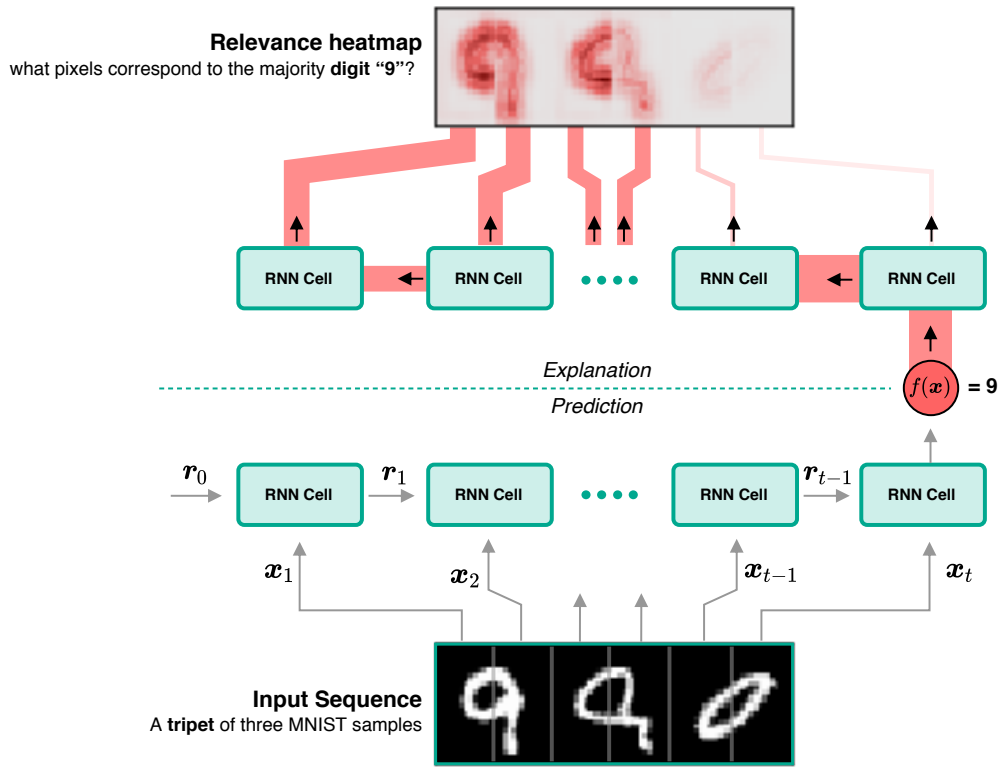
$$R_j = \sum_k \frac{a_j w_{jk} - l_j w_{jk}^+ - h_j w_{jk}^-}{\sum_{j'} a_{j'} w_{j'k} - l_{j'} w_{j'k}^+ - h_{j'} w_{j'k}^-} R_k$$

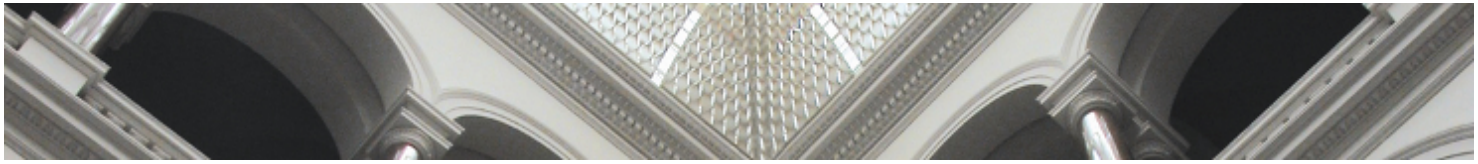


Technische Universität Berlin

Experimental Setup

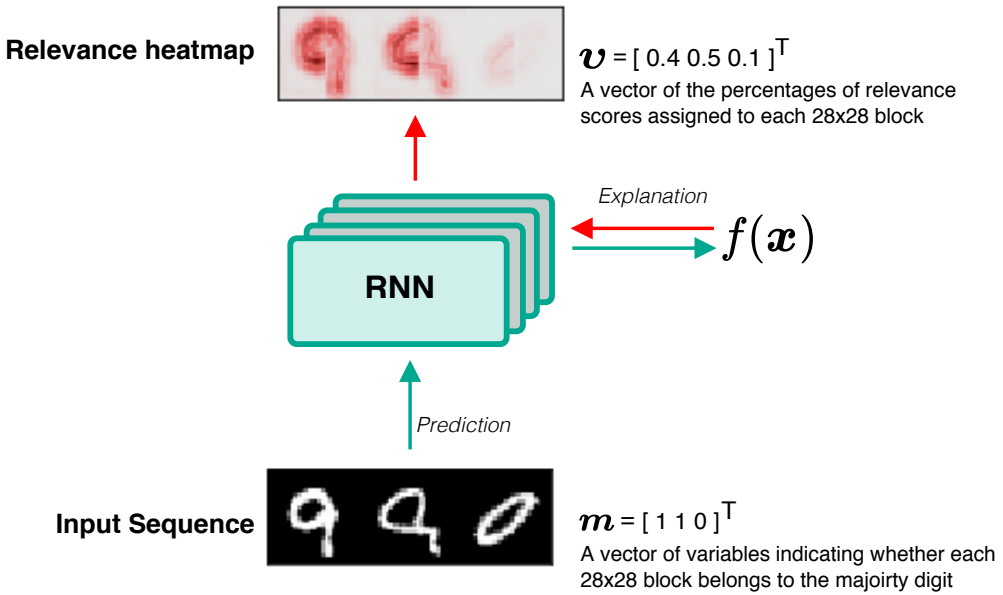
- **Problem:** Majority-Sample Sequence Classification (MNIST-MAJ, FashionMNIST-MAJ)
  - ▷ Sequence length: **12** $\left(\{\boldsymbol{x}_t \in \mathbb{R}^{28\times 7}\}_{t=1}^{12}\right)$
  - ▷ Minimum accuracy 98% for MNIST-MAJ and 89% for FashionMNIST-MAJ





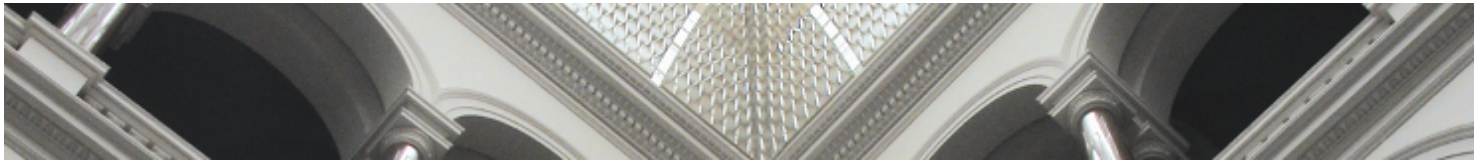
Experimental Setup

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

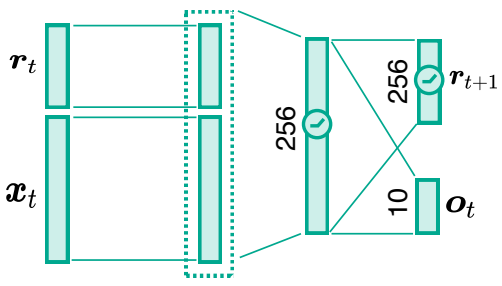


Heatmap Quality

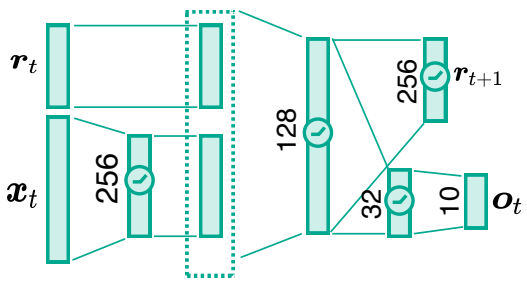
$$\cos(\boldsymbol{m}, \boldsymbol{v}) = \frac{\boldsymbol{m}^T \boldsymbol{v}}{\|\boldsymbol{m}\| \|\boldsymbol{v}\|}$$



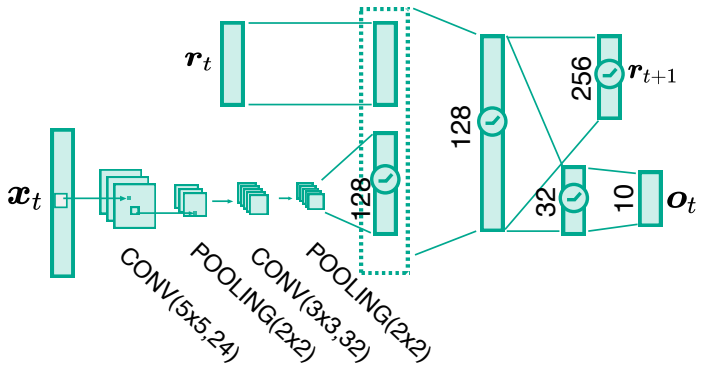
Experiment 1: Standard RNN Architectures



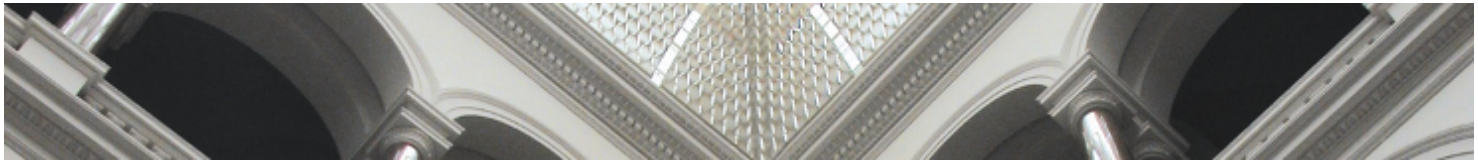
Shallow



Deep



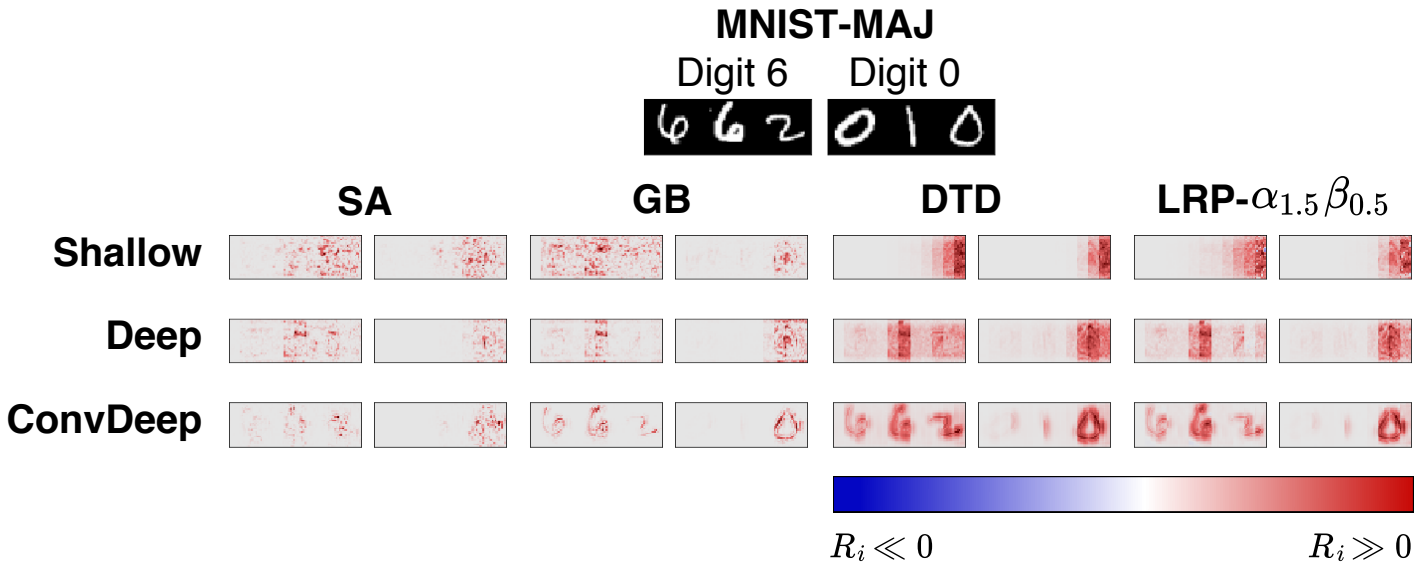
ConvDeep

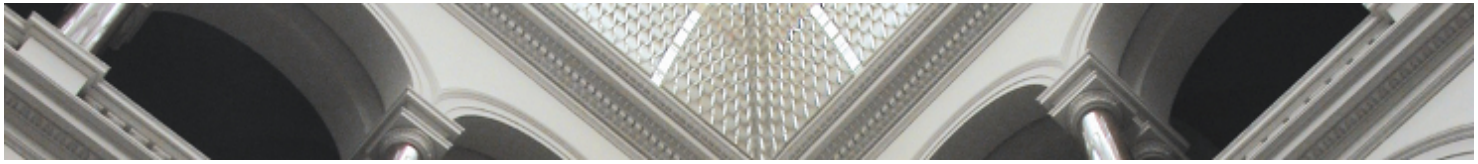


Technische Universität Berlin

Experiment 1: Standard RNN Architectures

Sample Relevance Heatmaps

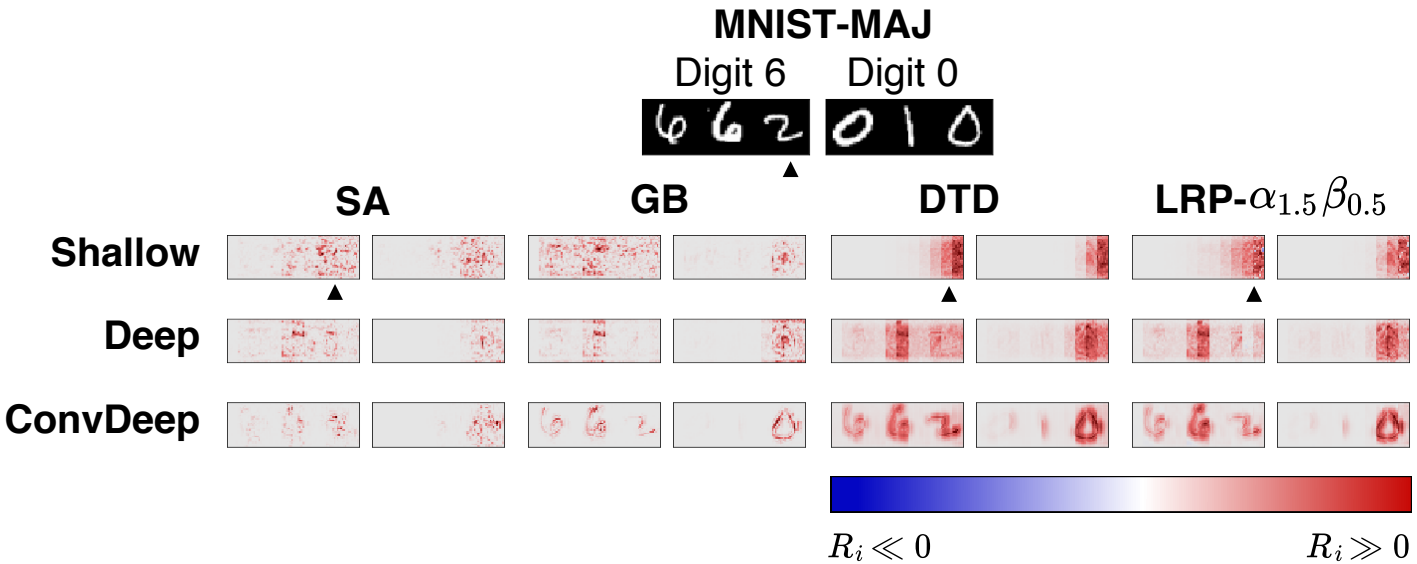


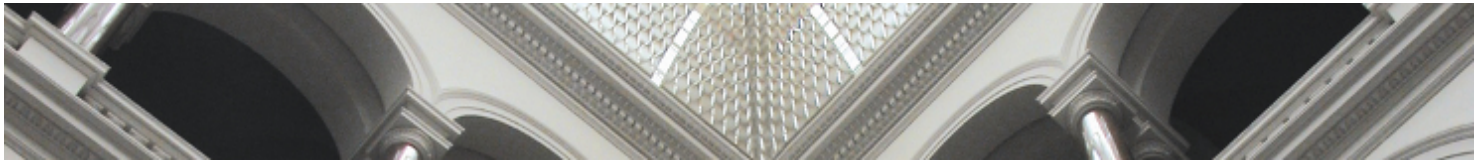


Technische Universität Berlin

Experiment 1: Standard RNN Architectures

Sample Relevance Heatmaps

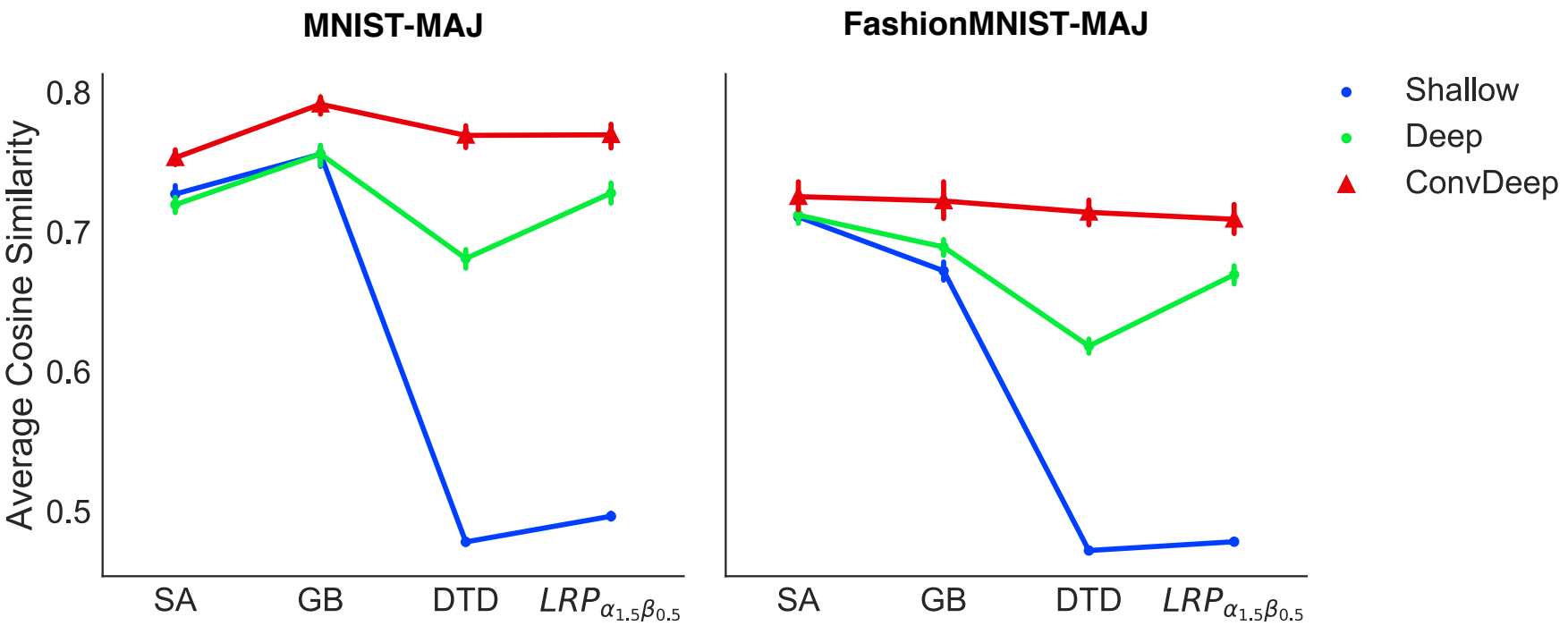


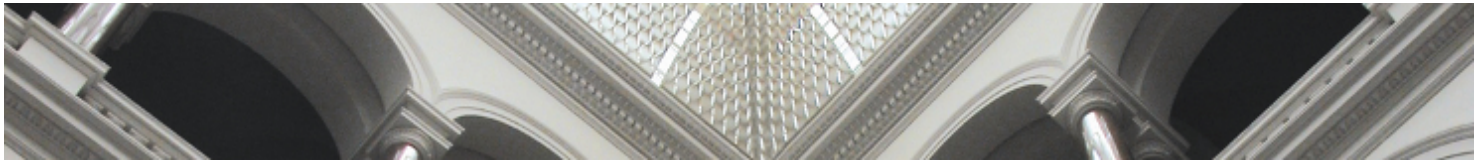


Technische Universität Berlin

Experiment 1: Standard RNN Architectures

Cosine Similarity Evaluation

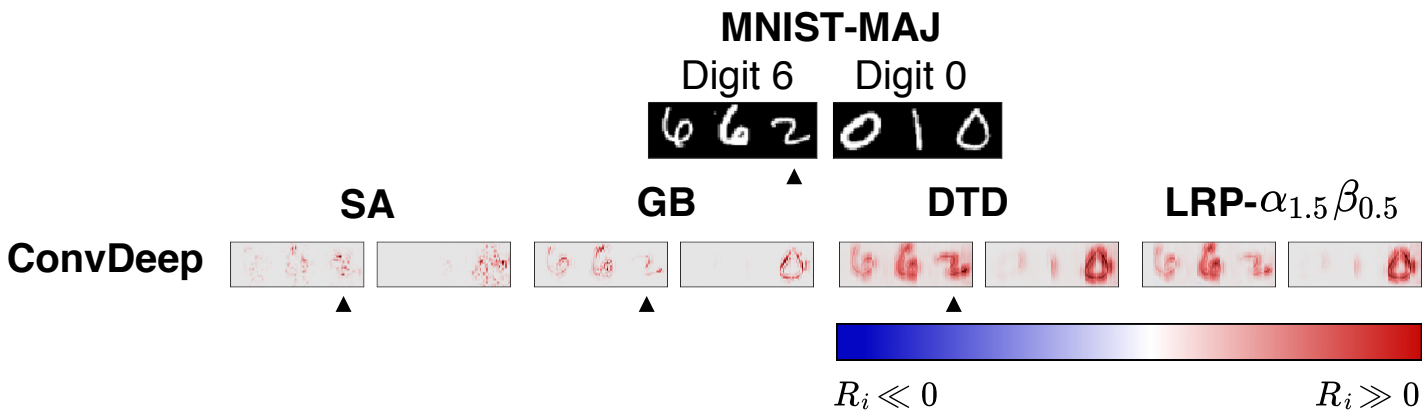




Technische Universität Berlin

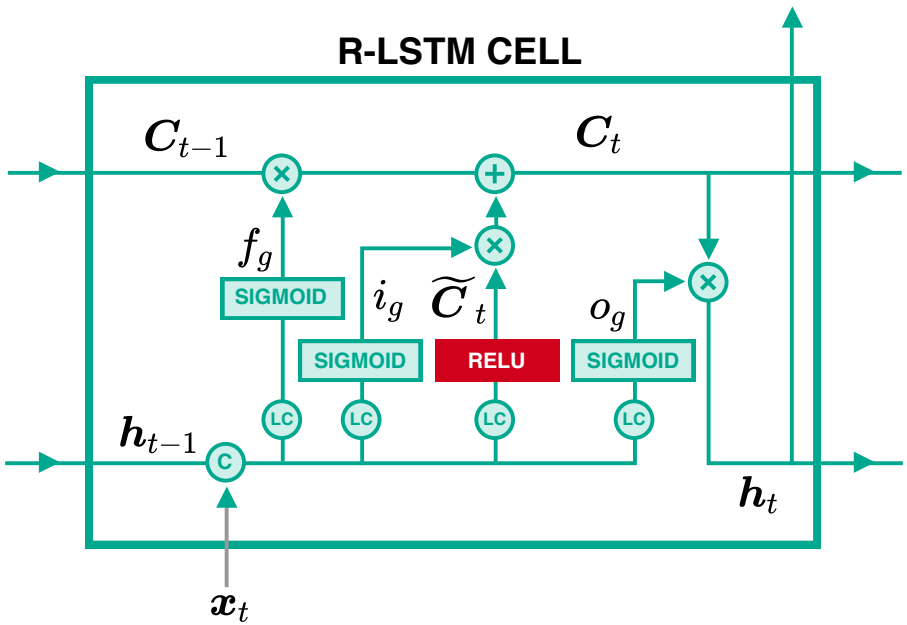
Experiment 2: More Explainable Models

Motivation: addressing the improper relevance assignment

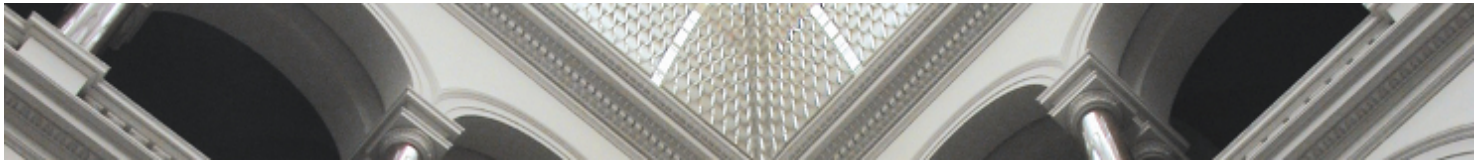


Experiment 2: More Explainable Models

1. LSTM [HS97]
   — Ignore gating neurons during explaining [AMMS17]
   — Replace tanh by ReLU (R-LSTM)

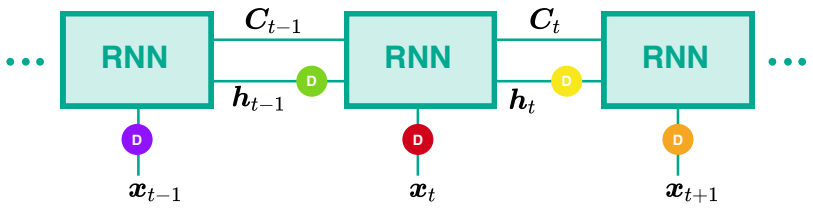


C Vector concatenation LC Linear combination ⊗ Element-wise multiplication ⊕ Element-wise addition

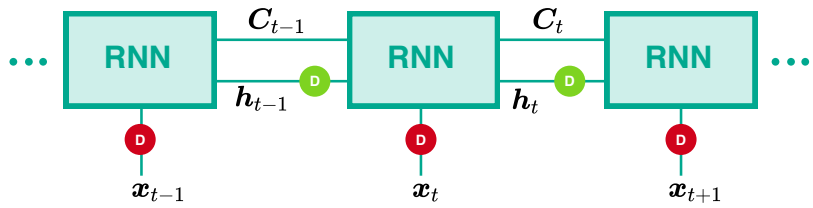


Experiment 2: More Explainable Models

1. LSTM [HS97]
2. Stationary Dropout (*Variational Recurrent* [GG16])

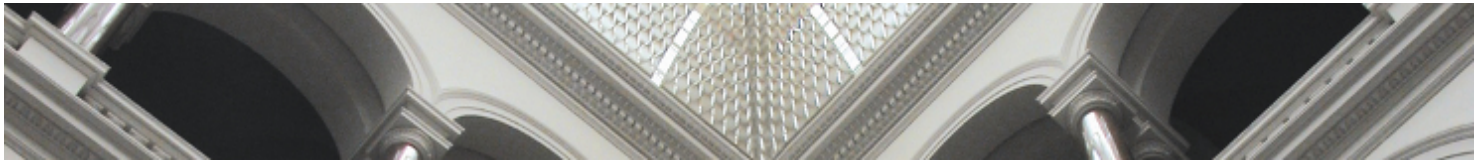


Naive Dropout



Stationary Dropout (-SD)

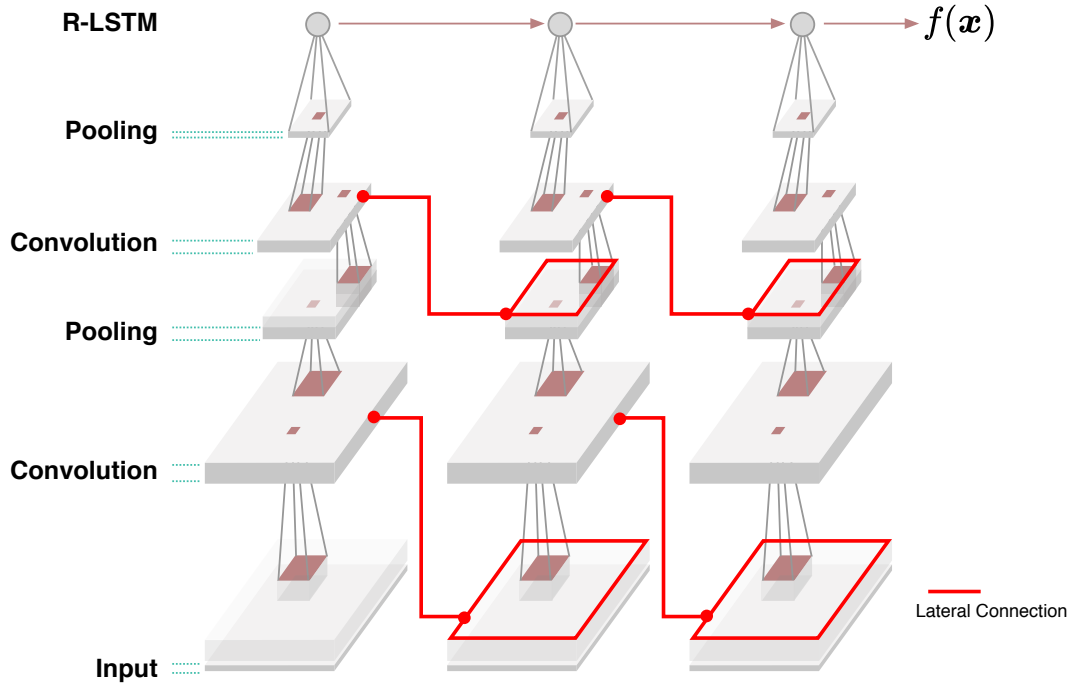
- Keras(2.2.4) : default behvariour for recurrent layers
- TensorFlow(r1.12): need to specify **variational_recurrent=True**

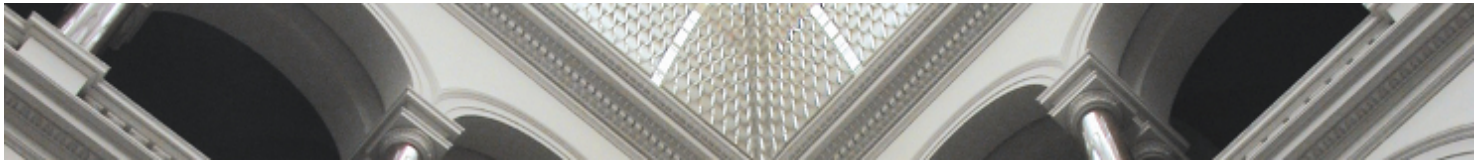


Technische Universität Berlin

Experiment 2: More Explainable Models

1. LSTM [HS97]
2. Stationary Dropout (*Variational Recurrent* [GG16])
3. Lateral connections for convolutional layers (Conv^{+})

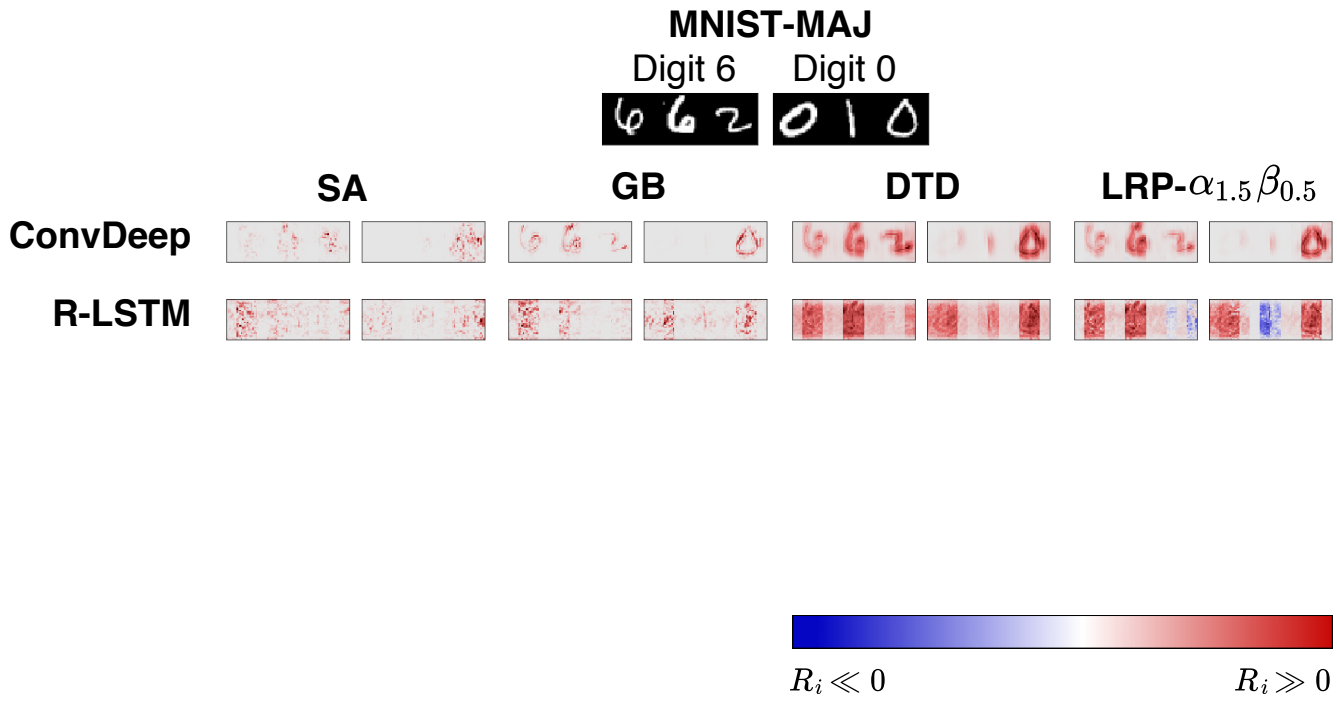


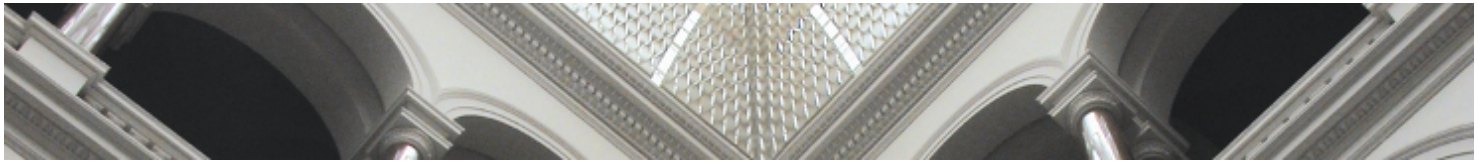


Technische Universität Berlin

Experiment 2: More Explainable Models

Sample Relevance Heatmaps

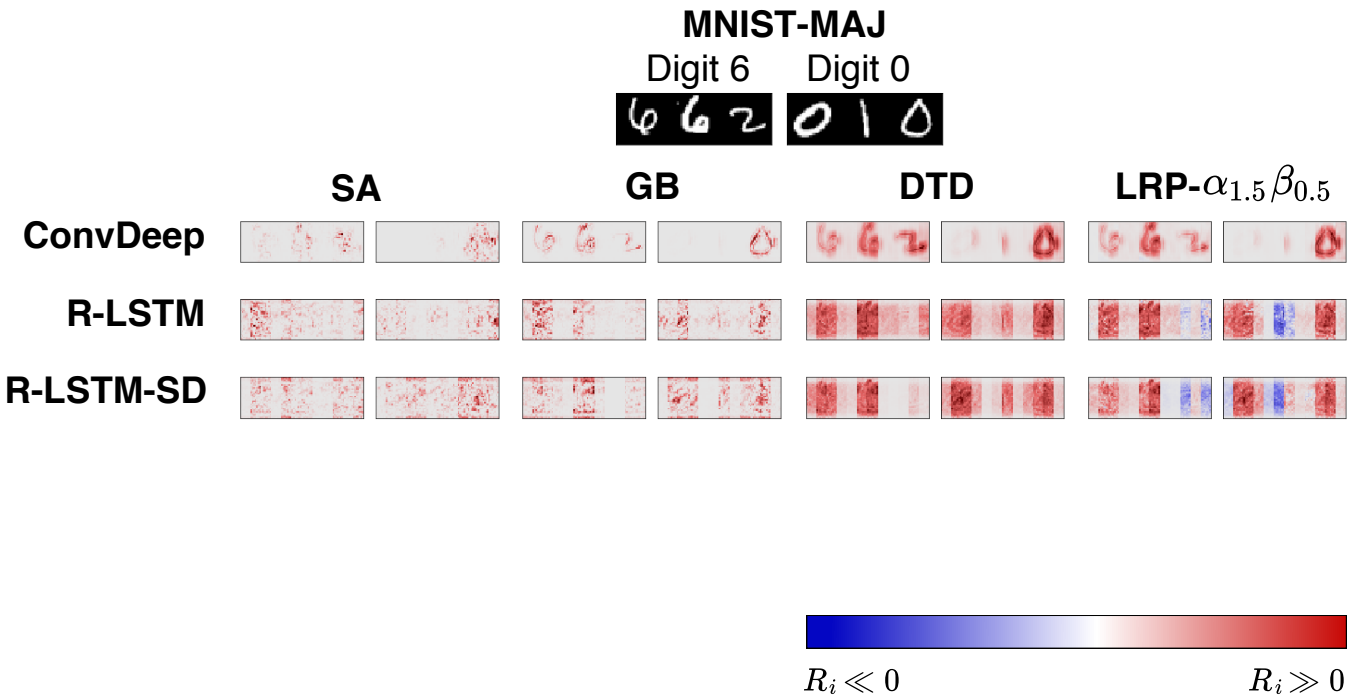


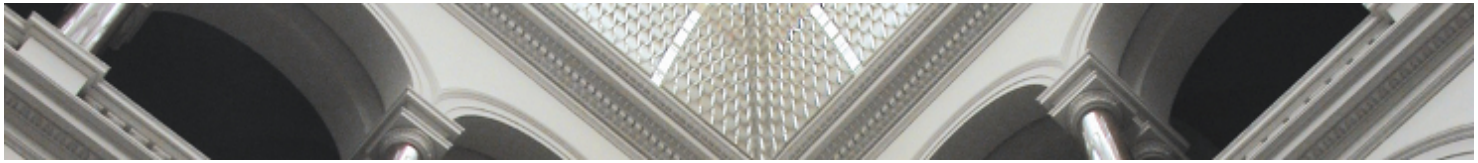


Technische Universität Berlin

Experiment 2: More Explainable Models

Sample Relevance Heatmaps

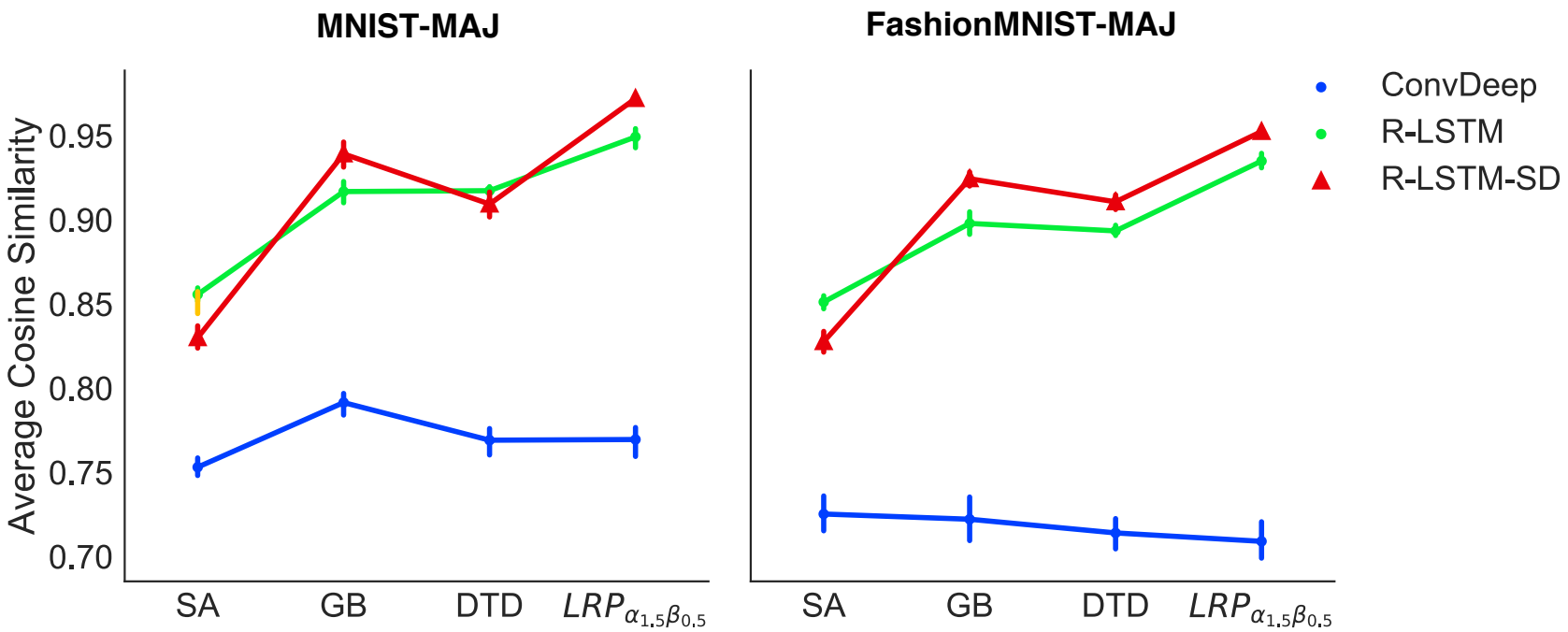


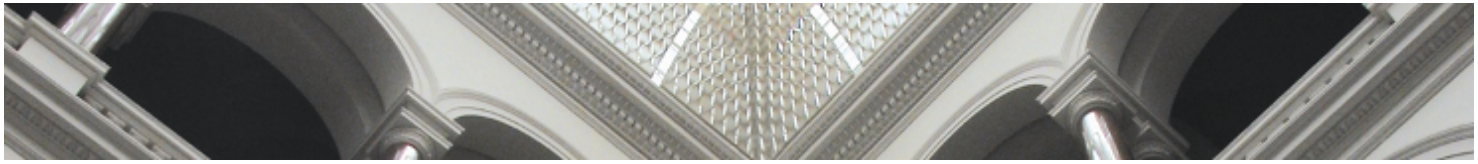


Technische Universität Berlin

Experiment 2: More Explainable Models

Cosine Similarity Evaluation

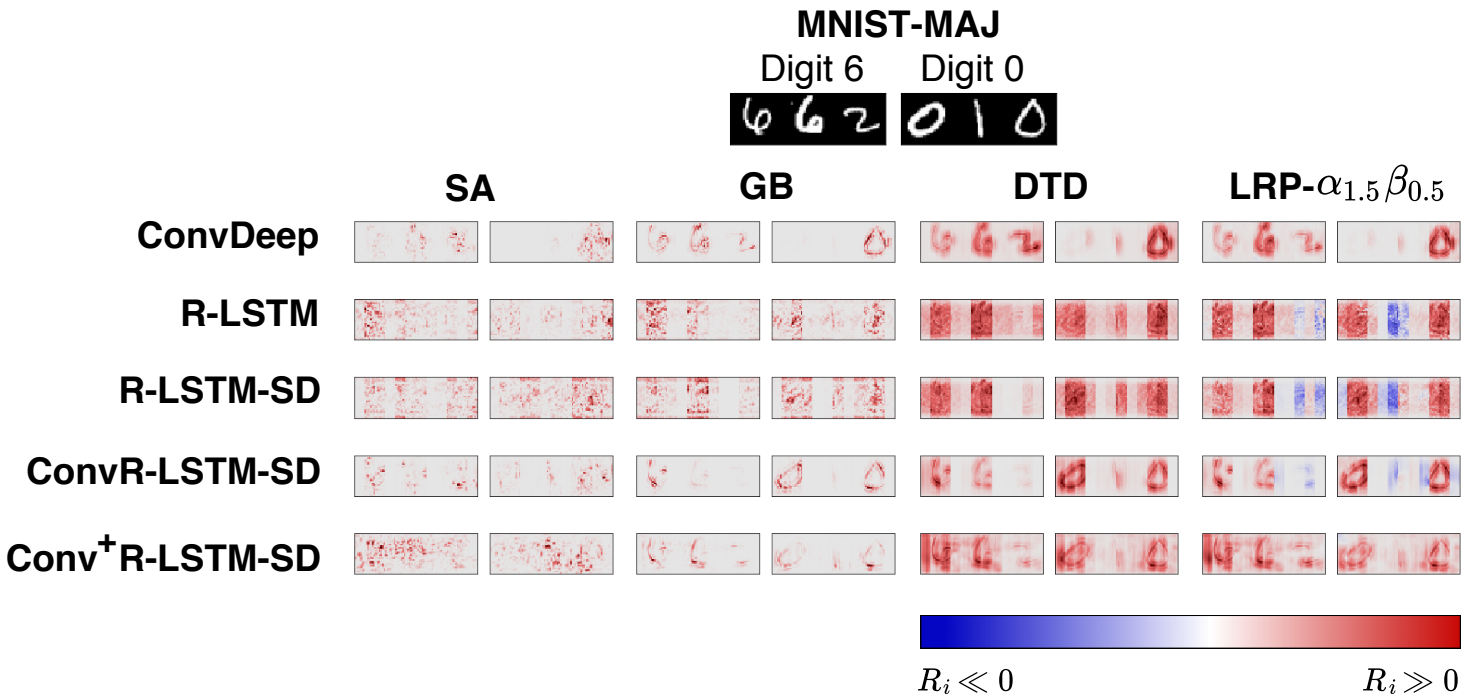


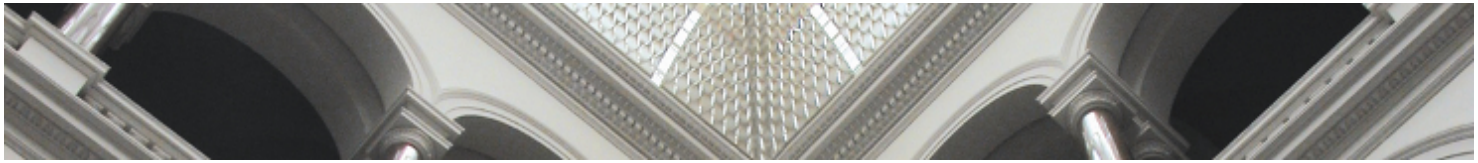


Technische Universität Berlin

Experiment 2: More Explainable Models

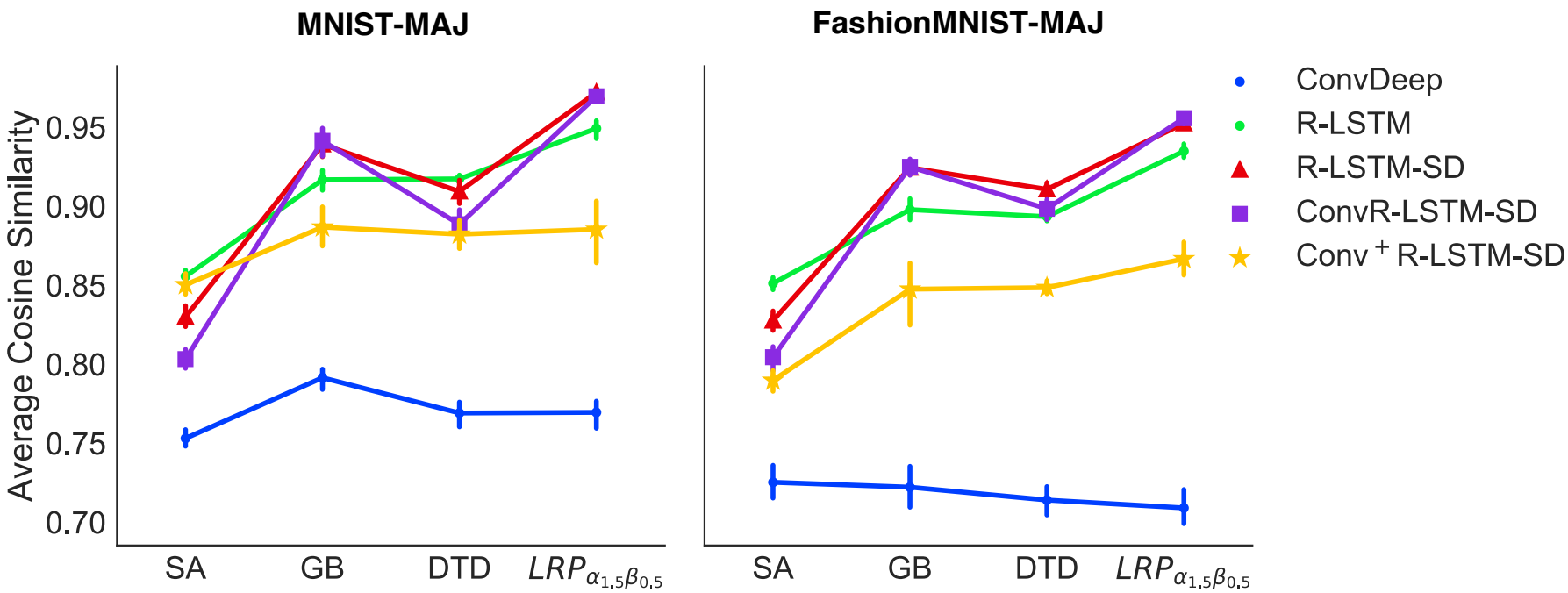
Sample Relevance Heatmaps

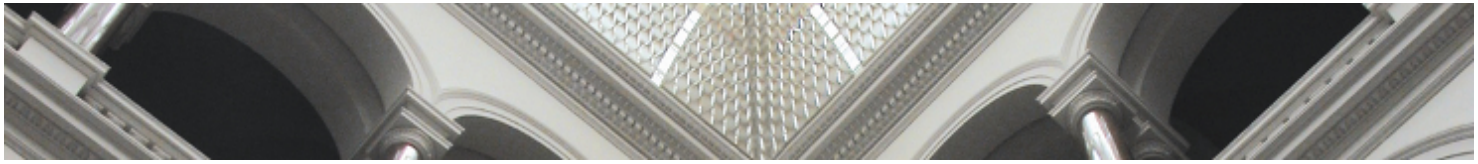




Experiment 2: More Explainable Models

Cosine Similarity Evaluation

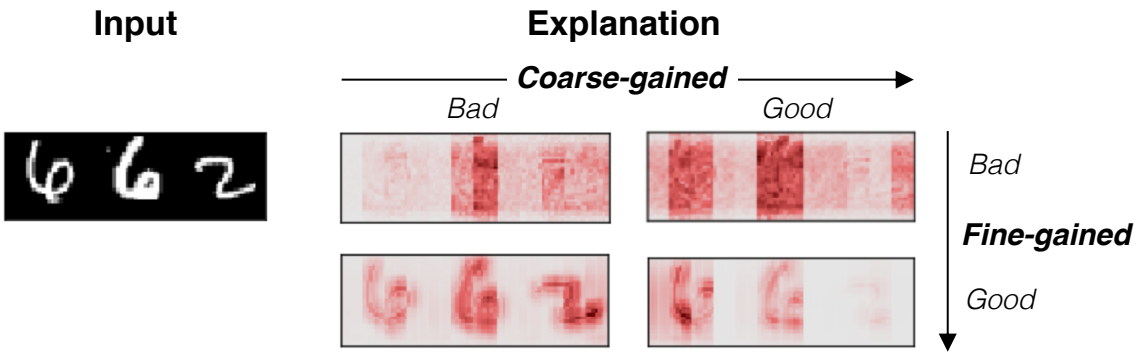




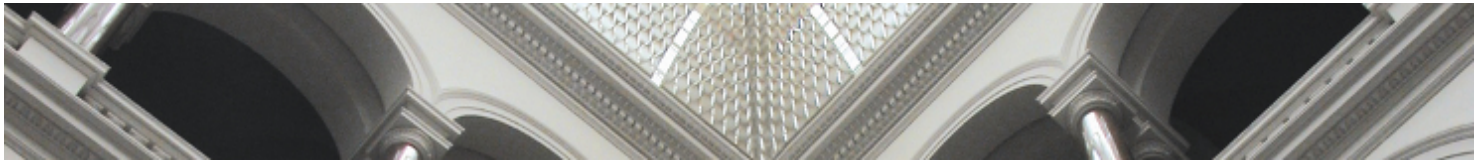
Technische Universität Berlin

Conclusion

- **Deep and LSTM-type RNNs** have more explainable predictions.
- **Stationary dropout** could improve model's explainability.
- Explainability of RNNs should be considered in two aspects:



- **Coarse-gained**: relevance quantities adequately propagated to relevant steps
  — solution: better recurrent mechanism (i.e. LSTMs)
- **Fine-gained**: soundness of each step's explanation
  — solution: choice of input layers (i.e. convolution and pooling layers for image data)



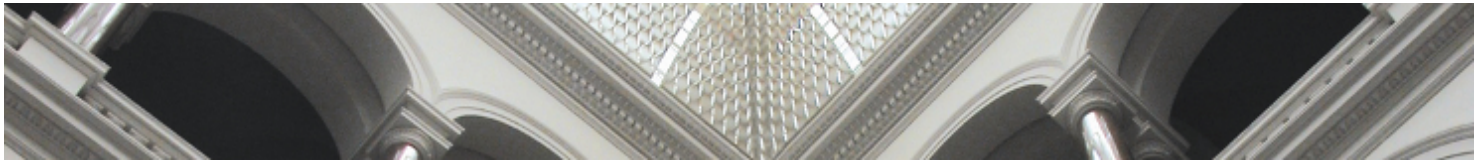
Technische Universität Berlin

Q&A

- Part of this work is in

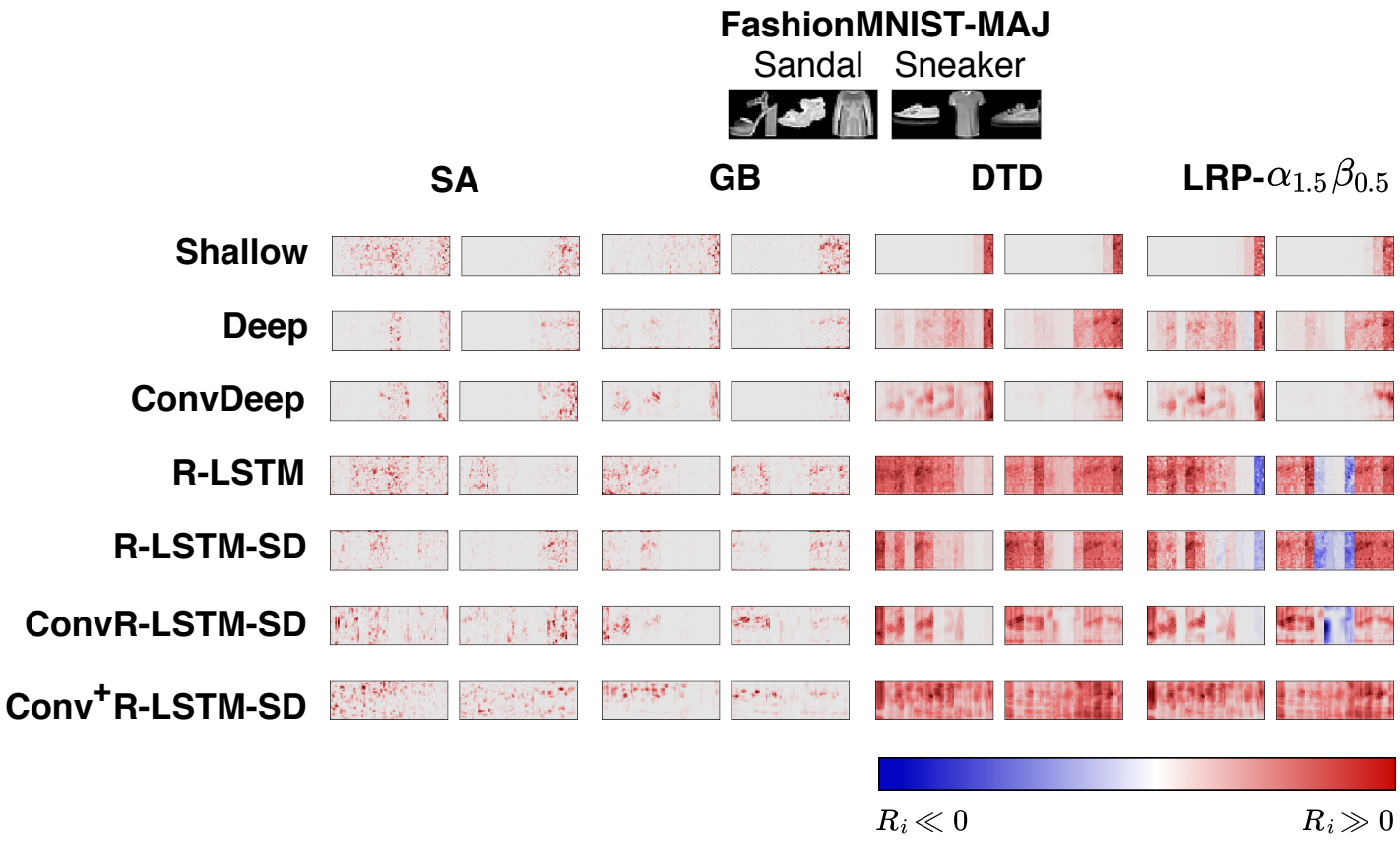
  Rieger, L., Chormai, P., Montavon, G., Hansen, L. K., & Müller, K. R. (2018). **Structuring Neural Networks for More Explainable Predictions.** In *Explainable and Interpretable Models in Computer Vision and Machine Learning* (pp. 115-131). Springer, Cham.

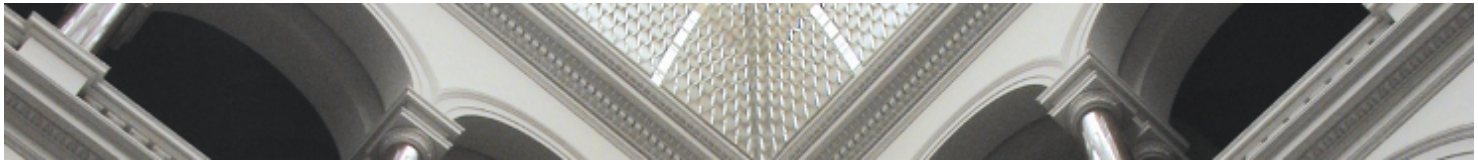
- Code can be found at **bit.ly/pat-thesis-repo**.



Technische Universität Berlin

FashionMNIST Heatmaps

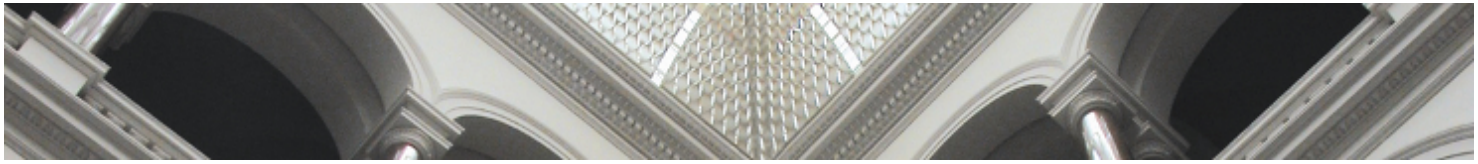




Technische Universität Berlin

Model Accuracy

dataset	architecture	count	avg_acc	std
mnist-maj	shallow	7	98.35	0.2393
mnist-maj	deep	7	98.43	0.0926
mnist-maj	convdeep	7	99.28	0.0737
mnist-maj	rlstm	7	98.77	0.0603
mnist-maj	rlstm_sd	7	98.75	0.1248
mnist-maj	convrlstm_sd	7	99.60	0.0467
mnist-maj	convlateral_rlstm_sd	7	98.30	0.1842
fashion-maj	shallow	7	92.30	0.2550
fashion-maj	deep	7	91.35	0.3053
fashion-maj	convdeep	7	94.19	0.2128
fashion-maj	rlstm	7	93.40	0.2714
fashion-maj	rlstm_sd	7	93.42	0.2513
fashion-maj	convrlstm_sd	7	95.44	0.1356
fashion-maj	convlateral_rlstm_sd	7	89.71	0.4215

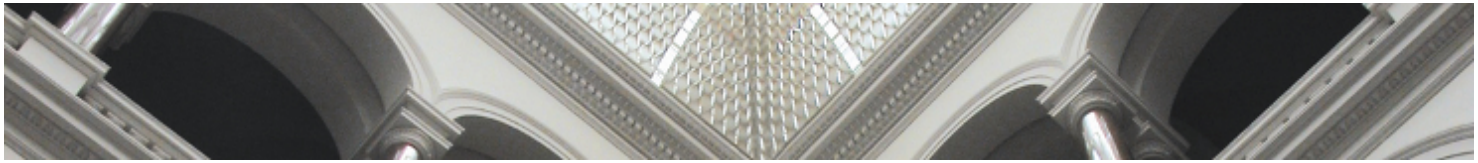


Technische Universität Berlin

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

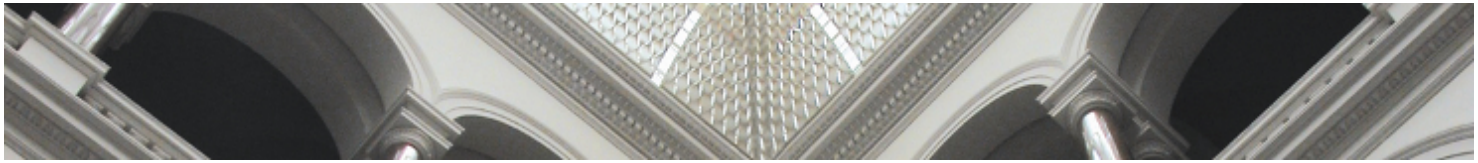
References II

[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.



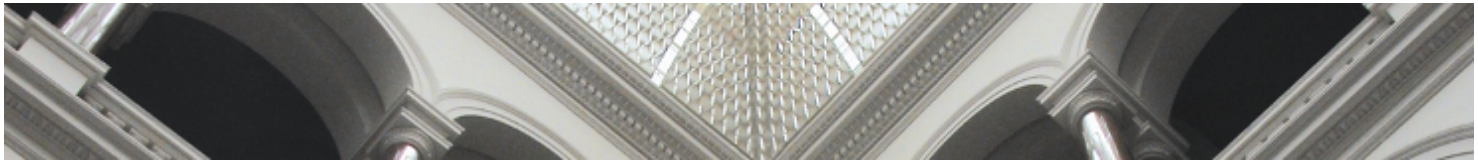
Technische Universität Berlin

References III

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

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, **Attention is all you need**, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017, pp. 6000–6010.

[VTBE15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, **Show and tell: A neural image caption generator**, IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, 2015, pp. 3156–3164.



Technische Universität Berlin

References IV

[WSC+16] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean, **Google's neural machine translation system: Bridging the gap between human and machine translation**, CoRR **abs/1609.08144** (2016).