



## Designing Recurrent Neural Networks for Explainability

Pattarawat Chormai | Technische Universität Berlin | 02/12/2018

---



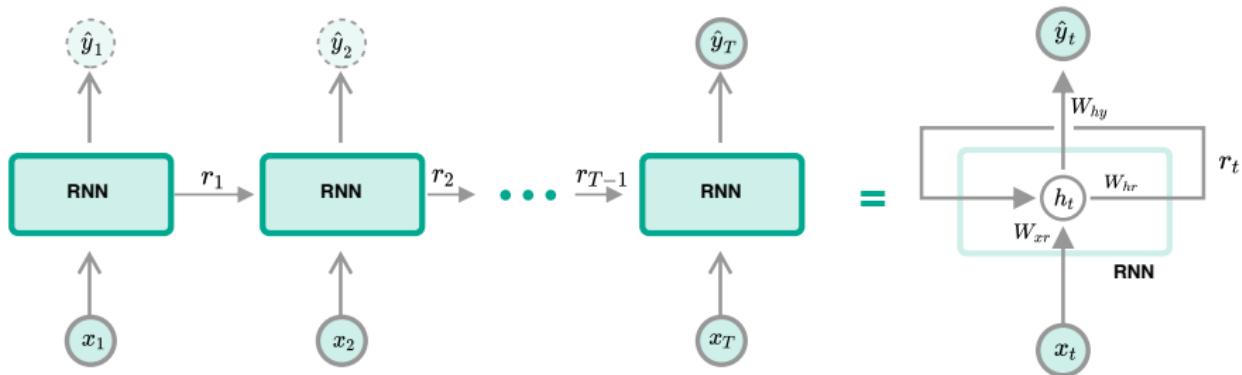
## Motivation

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

## Recurrent Neural Networks (RNNs)

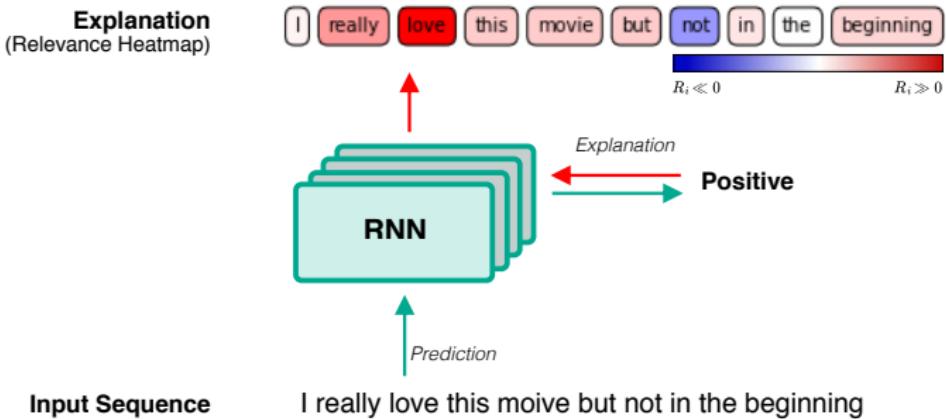
- Successful applications:**

▷ machine translation [WSC<sup>+</sup>16], image captioning [VTBE15, VSP<sup>+</sup>17]



## 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?





## Explanation Methods I

### 1. Sensitivity Analysis [SVZ13]

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

### 2. Guided Backprop [SDBR15]

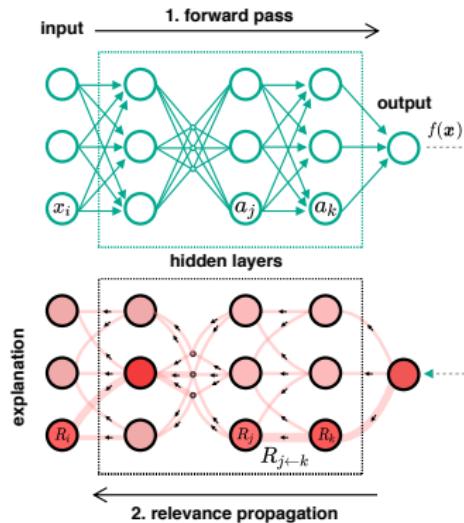
- Developed for explaining **ReLU**-based CNNs

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

## Explanation Methods II

### 3. Layer-wise Relevance Propagation (LRP) [BML<sup>+</sup>16]

- Distributing relevance based on **neurons' activities**



LRP- $\alpha\beta$  rule:

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

## Explanation Methods III

### 4. Deep Taylor Decomposition [MLB<sup>+</sup>17]

- LRP's theoretical view for explaining **ReLU-based** architectures

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

- Two important propagation rules:

–  $z^+$  for  $a_j \in \mathbb{R}^+$  (LRP- $\alpha_1\beta_0$ )

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

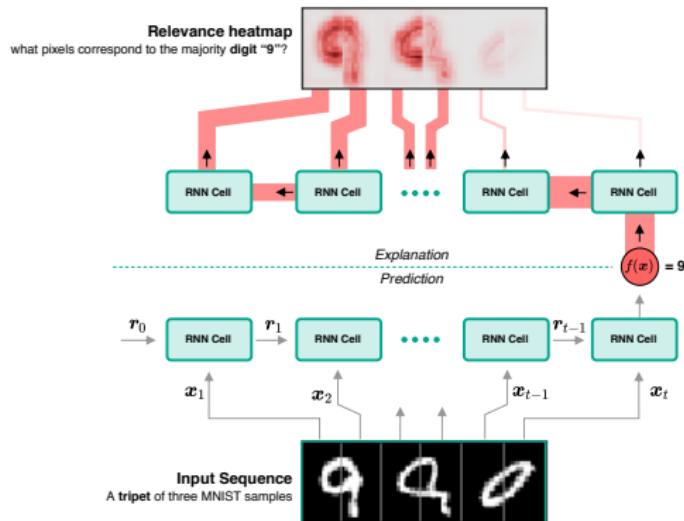
–  $z^B$  for  $a_j \in [l_j, h_j]$  where  $l_j \leq 0 < h_j$

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



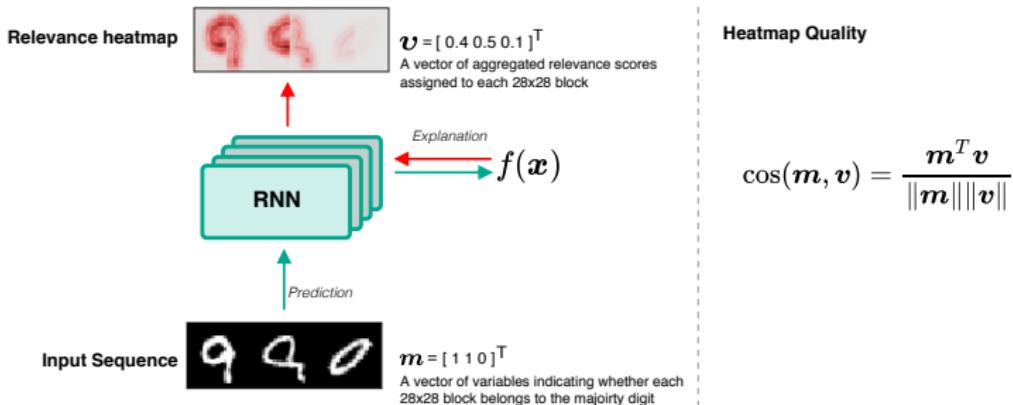
## Experimental Setup

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



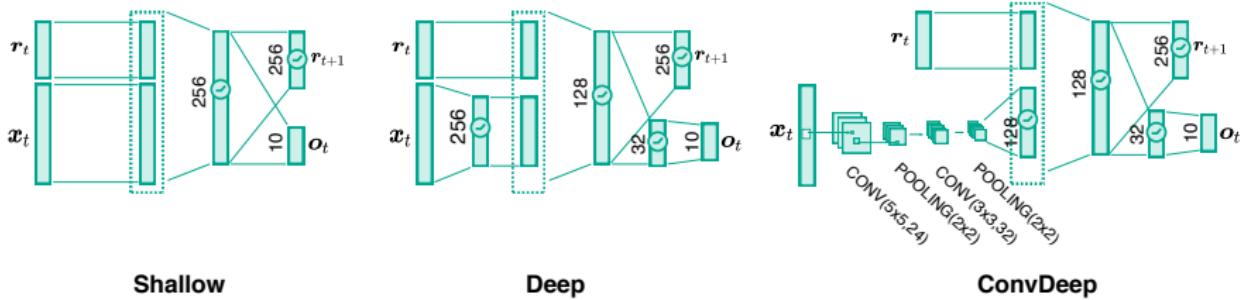
## Experimental Setup

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





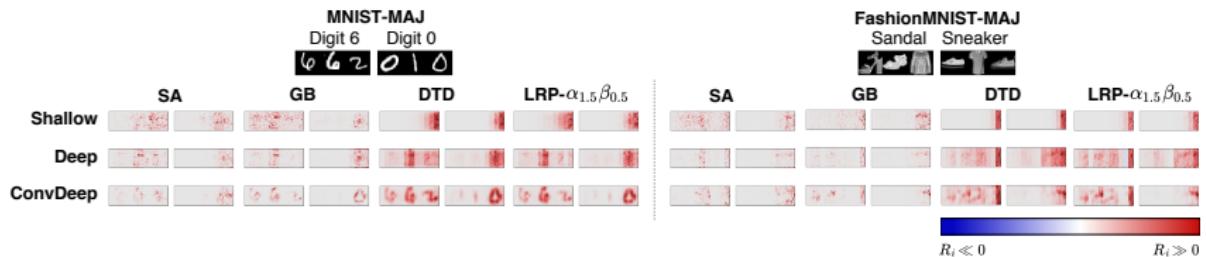
## Experiment 1: Standard RNN Architectures





## Experiment 1: Standard RNN Architectures

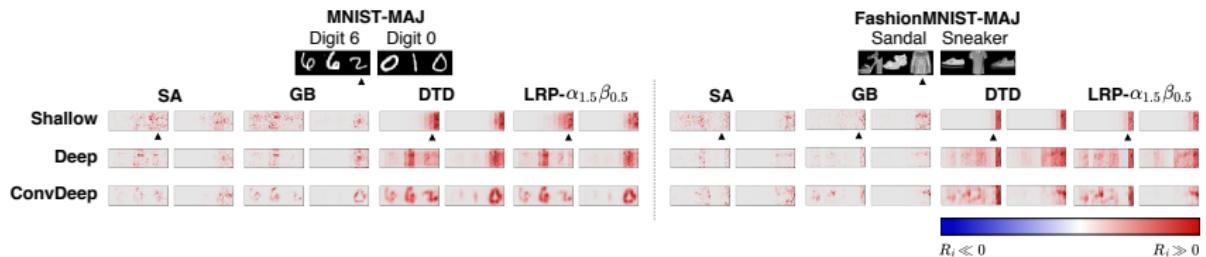
### Sample Relevance Heatmaps





## Experiment 1: Standard RNN Architectures

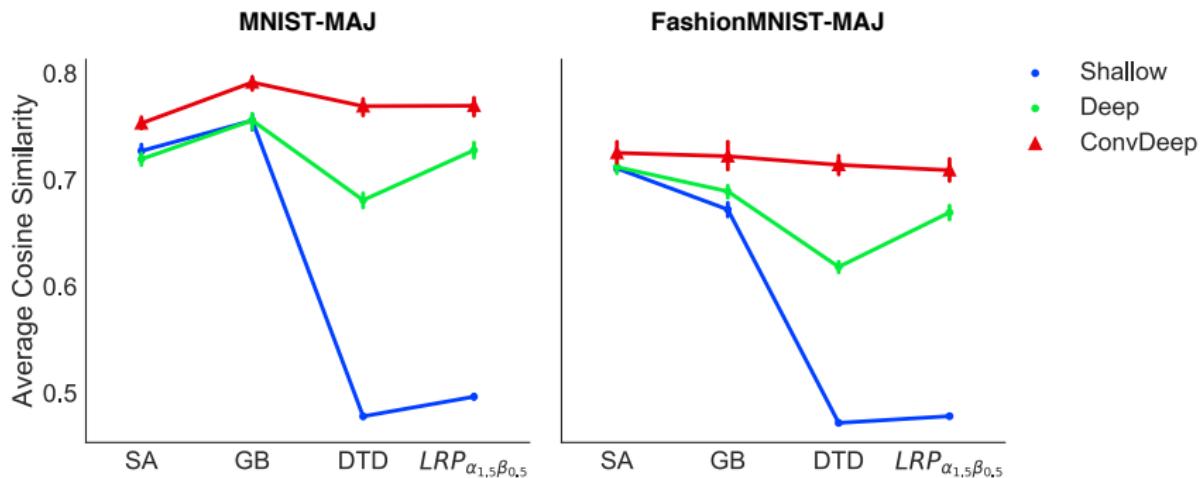
### Sample Relevance Heatmaps





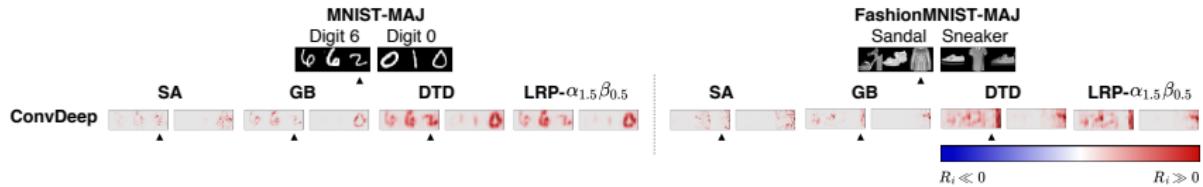
## Experiment 1: Standard RNN Architectures

### Cosine Similarity Evaluation



## Experiment 2: More Explainable Models

**Motivation:** addressing the improper relevance assignment

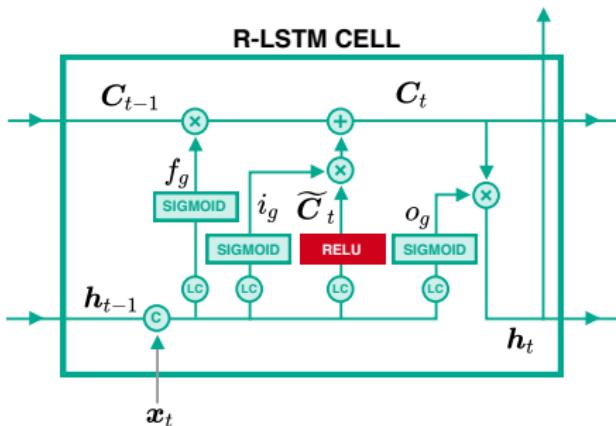




## Experiment 2: More Explainable Models

### 1. LSTM [HS97]

- Ignore gating units during explaining [AMMS17]
- Replace tanh by ReLU (R-LSTM)



Vector concatenation

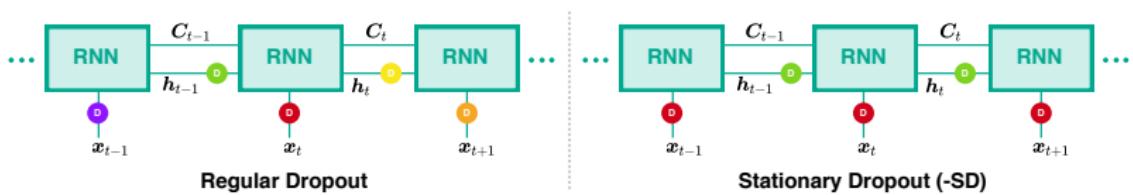
Linear combination

Element-wise multiplication

Element-wise addition

## Experiment 2: More Explainable Models

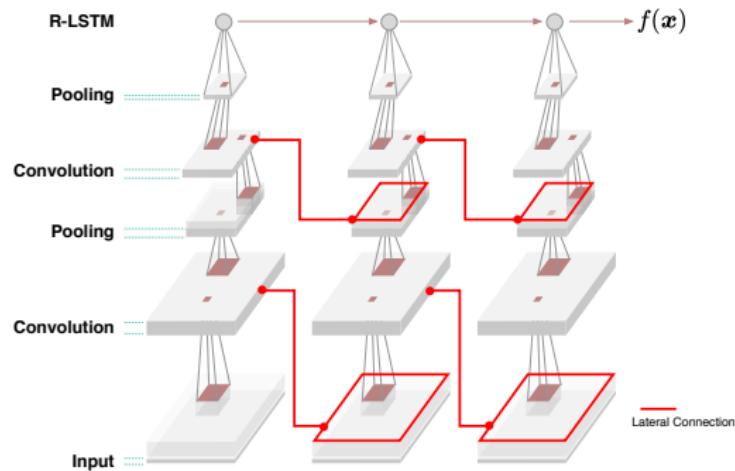
1. LSTM [HS97]
2. Stationary Dropout [GG16] (-SD)





## Experiment 2: More Explainable Models

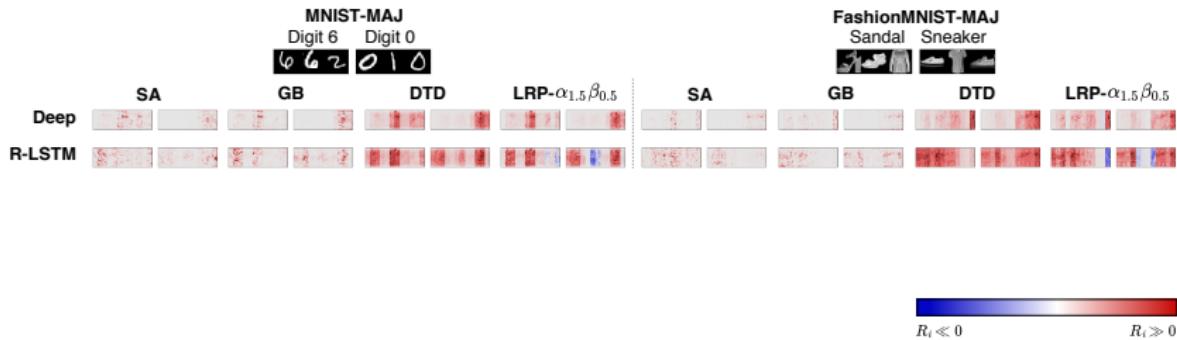
1. LSTM [HS97]
2. Stationary Dropout [GG16] (-SD)
3. Lateral connections for convolutional layers ( $\text{Conv}^+$ )





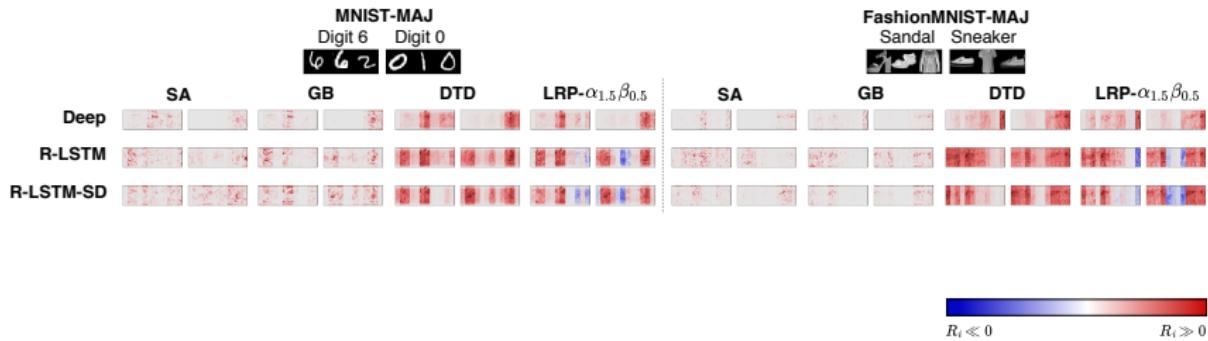
## Experiment 2: More Explainable Models

### Sample Relevance Heatmaps



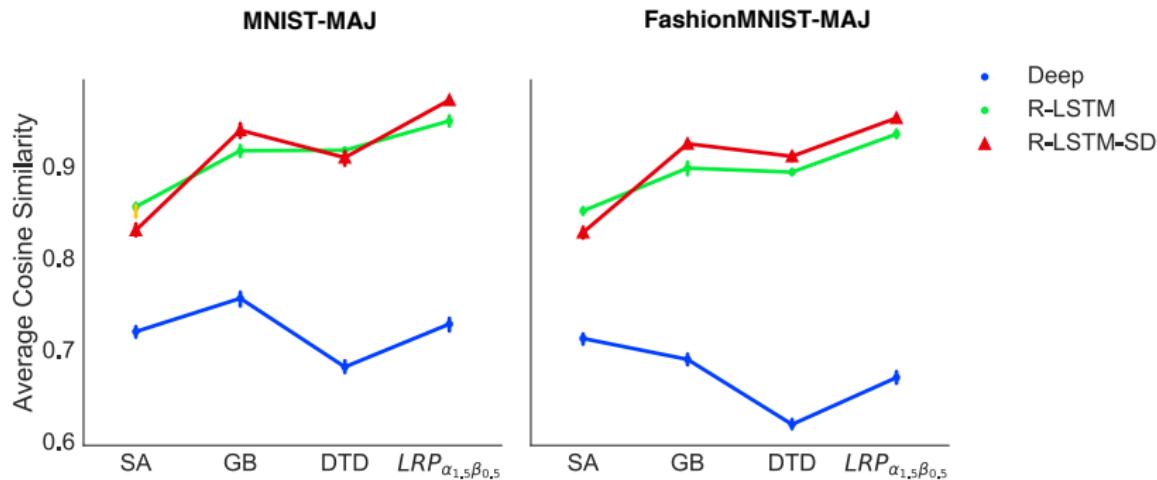
## Experiment 2: More Explainable Models

### Sample Relevance Heatmaps



## Experiment 2: More Explainable Models

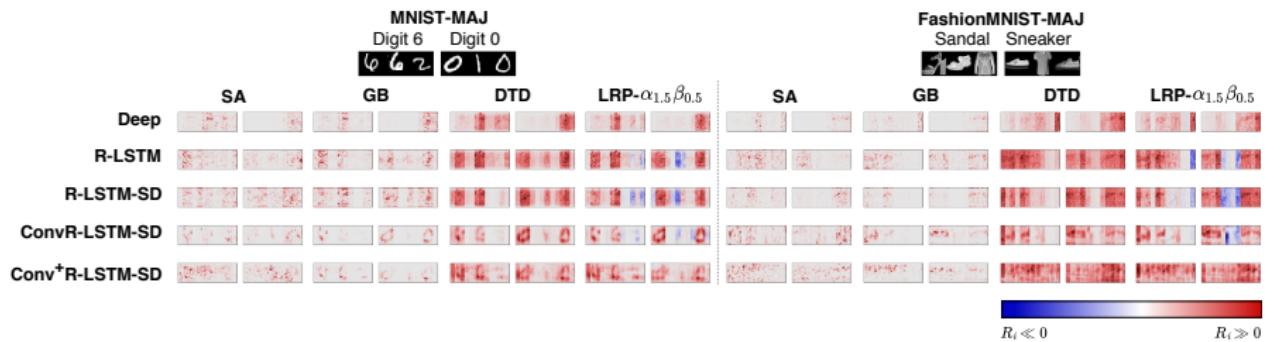
### Cosine Similarity Evaluation





## Experiment 2: More Explainable Models

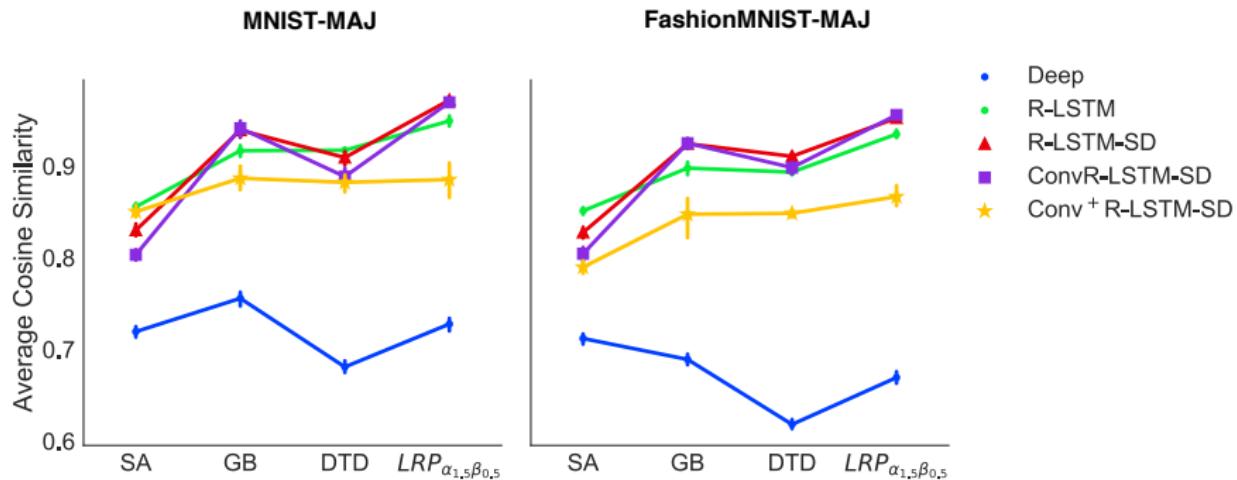
### Sample Relevance Heatmaps





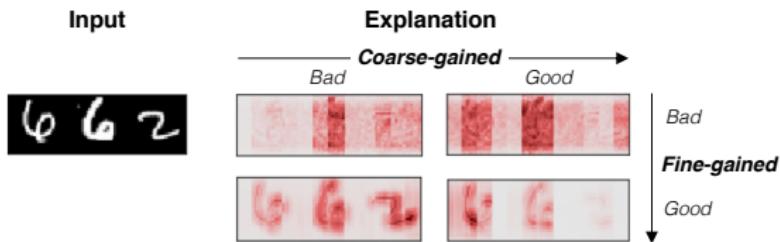
## Experiment 2: More Explainable Models

### Cosine Similarity Evaluation



## Conclusion I

- 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 adequately propagated to the 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)



## Conclusion II

- Part of this work will appear in:

Rieger, L., Chormai, P., Montavon, G., Hansen, L. K., & Müller, K.-R. (2018). **Structuring Neural Networks for More Explainable Predictions.** In H. Jair Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. A. J. van Gerven (Eds.), Explainable and Interpretable Models in Computer Vision and Machine Learning (1st ed., p. 299). Springer International Publishing.



## Q&A



## 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<sup>+</sup>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.

## 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<sup>+</sup>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.



## 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](https://arxiv.org/abs/1312.6034) (2013).
- [VSP<sup>+</sup>17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz 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.



## References IV

- [WSC<sup>+</sup>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](https://arxiv.org/abs/1609.08144) (2016).