Designing RNN for Explainability

This work is a part of master's thesis.

Pattarawat Chormai
Supervised by Dr. Grégoire Montavon, Prof. Klaus-Robert Müller
Technical University Berlin, Department of Machine Learning
p.chormai@campus.tu-berlin.de

Abstract

Sed fringilla tempus hendrerit. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Etiam ut elit sit amet metus lobortis consequat sit amet in libero. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus vel sem magna. Nunc at convallis urna. isus ante. Pellentesque condimentum dui. Etiam sagittis purus non tellus tempor volutpat. Donec et dui non massa tristique adipiscing.

Introduction

Aliquam non lacus dolor, *a aliquam quam* [**?**]. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Nulla in nibh mauris. Donec vel ligula nisi, a lacinia arcu. Sed mi dui, malesuada vel consectetur et, egestas porta nisi. Sed eleifend pharetra dolor, et dapibus est vulputate eu. **Integer faucibus elementum felis vitae fringilla.** In hac habitasse platea dictumst. Duis tristique rutrum nisl, nec vulputate elit porta ut. Donec sodales sollicitudin turpis sed convallis. Etiam mauris ligula, blandit adipiscing condimentum eu, dapibus pellentesque risus.

Sensitivity Analysis

[3], an modified version **Guided Backprop**[4]

Deep Taylor Decomposition

[2]

Layer-Wise Relevance Propagation

[1]

Experimental Setup

Dataset, problem training ... procedures ...

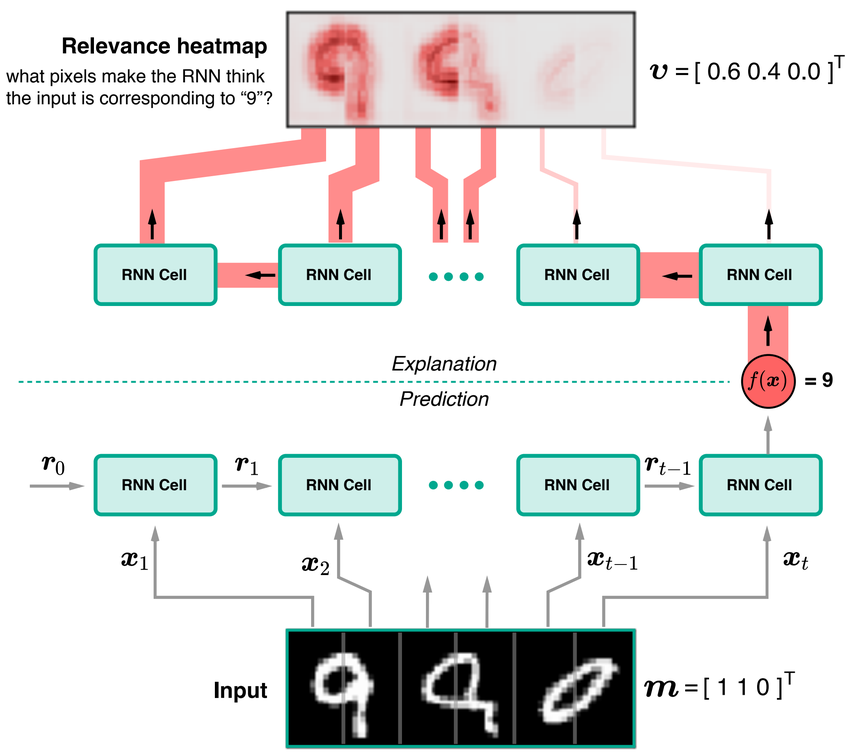

Figure 1: Figure caption

Evaluation methodology

Our quantitative evaluation is based on *cosine similarity* between a binary vector $\boldsymbol{m} \in \mathbb{R}^3$, whose entry indicates whether the item belongs to the majority group, and a vector $\boldsymbol{v} \in \mathbb{R}^3$ representing percentage of relevance distributed to the corresponding item.

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

Architectures

Figure X shows RNN architectures considered in this study. Standard RNN

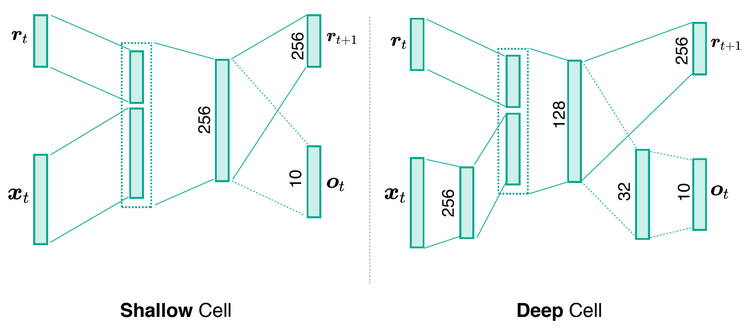

Figure 2: Figure caption

LSTM ...

Results

Donec faucibus purus at tortor egestas eu fermentum dolor facilisis. Maecenas tempor dui eu neque fringilla rutrum. Mauris *lobortis* nisl accumsan. Aenean vitae risus ante. Phasellus imperdiet, tortor vitae congue bibendum, felis enim sagittis lorem, et volutpat ante orci sagittis mi. Morbi rutrum laoreet semper. Morbi accumsan enim nec tortor consectetur non commodo nisi sollicitudin. Proin sollicitudin. Pellentesque eget orci eros. Fusce ultricies, tellus et pellentesque fringilla, ante massa luctus libero, quis tristique purus urna nec nibh.

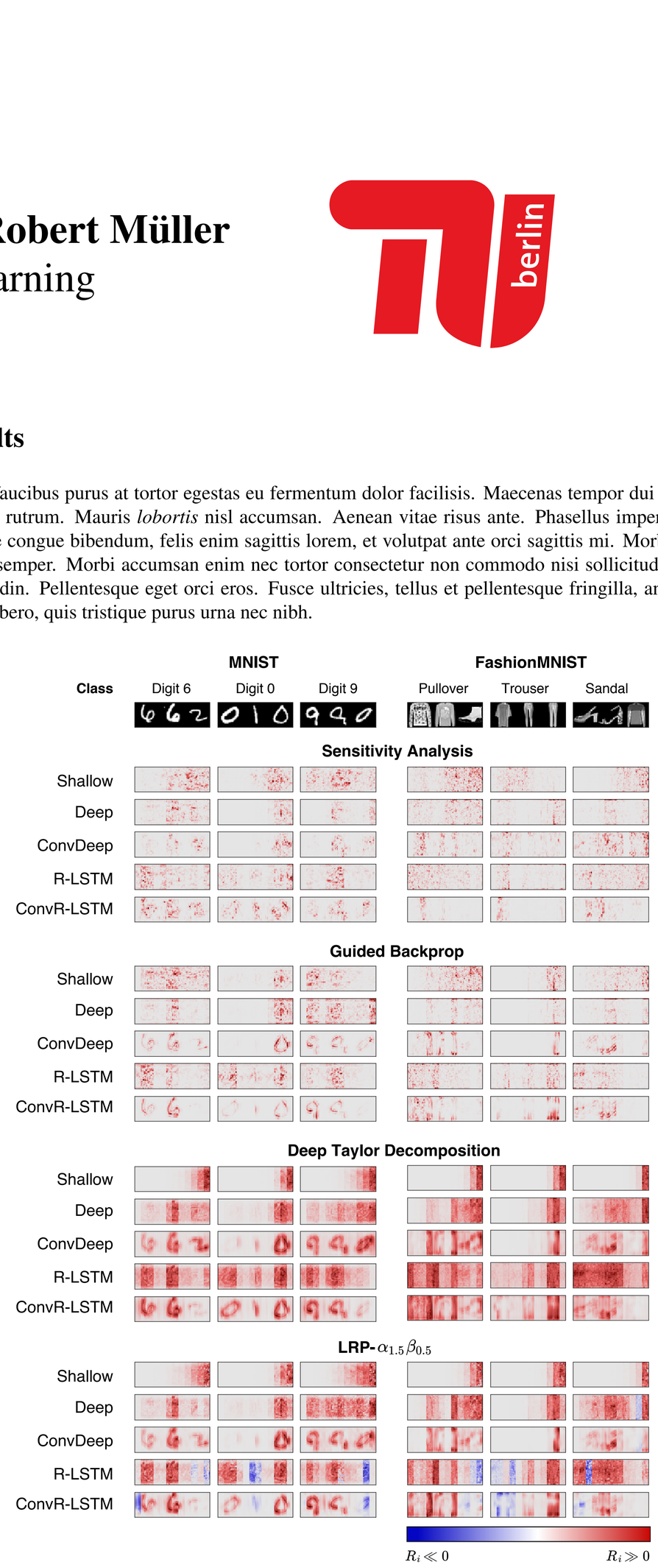

Figure 3: Figure caption

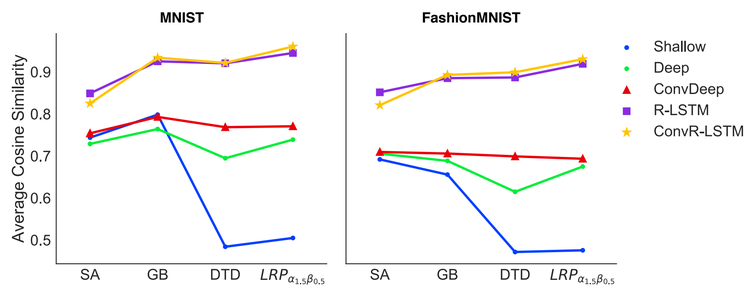

Figure 4: Figure caption

Conclusions

- Pellentesque eget orci eros. Fusce ultricies, tellus et pellentesque fringilla, ante massa luctus libero, quis tristique purus urna nec nibh. Phasellus fermentum rutrum elementum. Nam quis justo lectus.
- Vestibulum sem ante, hendrerit a gravida ac, blandit quis magna.
- Donec sem metus, facilisis at condimentum eget, vehicula ut massa. Morbi consequat, diam sed convallis tincidunt, arcu nunc.
- Nunc at convallis urna. isus ante. Pellentesque condimentum dui. Etiam sagittis purus non tellus tempor volutpat. Donec et dui non massa tristique adipiscing.

Future Work

Vivamus molestie, risus tempor vehicula mattis, libero arcu volutpat purus, sed blandit sem nibh eget turpis. Maecenas rutrum dui blandit lorem vulputate gravida. Praesent venenatis mi vel lorem tempor at varius diam sagittis. Nam eu leo id turpis interdum luctus a sed augue. Nam tellus.

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. 10(7).

[2] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. 65:211–222.

[3] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.

[4] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net.

Acknowledgements

Etiam fermentum, arcu ut gravida fringilla, dolor arcu laoreet justo, ut imperdiet urna arcu a arcu. Donec nec ante a dui tempus consectetur. Cras nisi turpis, dapibus sit amet mattis sed, laoreet.