Identifying Trends in Feature Attributions during Training of Neural Networks

Elena Terzieva, Maximilian Muschalik, Paul Hofman,
and Eyke Hüllermeier

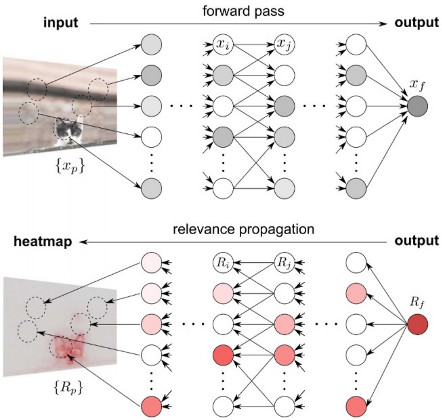
enterzieva@gmail.com
maximilian.muschalik@lmu.de
LMU Munich

Motivation

- Explainable artificial intelligence (XAI) techniques crucial in unraveling the inner workings of opaque machine learning models [1]
- XAI measures could also be used to explain how a model evolves during training
- How do the explanations of XAI methods change as the approximation uncertainty of a model is reduced?
- Study feature attribution methods as a function of training time!

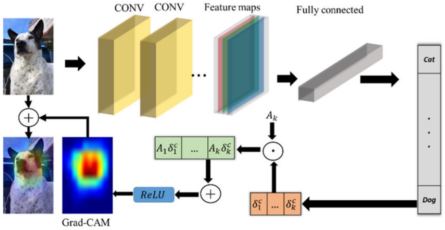
Backpropagation-Based XAI Approaches: LRP and Grad-CAM

LRP [2]



https://giorgiomorales.github.io/Layer-wise-Relevance-Propagation-in-Pytorch/

Grad-CAM [3]

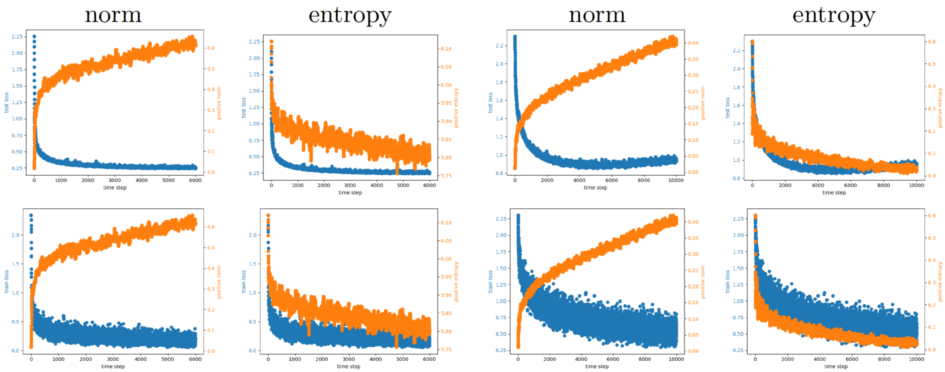


https://www.researchgate.net/publication/352795278/figure/fig3/AS:1039746730053633@1624906346684/Grad-CAM-architecture$_W$640.jpg

Methodology

- Model training: Train different models on FashionMNIST and CIFAR10
- Compute feature attributions: separate dataset of seen and unseen images
- Summarize: compute summary measures
  - Frobenius norm: $\|A\|_F = \sqrt{\sum_i^m \sum_j^n |a_{ij}^2|}$
  - Shannon entropy: $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$
- Analyze: descriptive and correlation analysis
  - Spearman's ρ

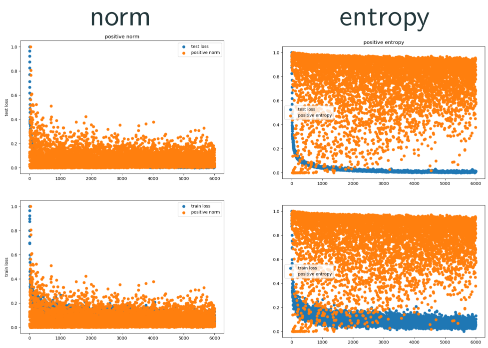
Descriptive Analysis Results



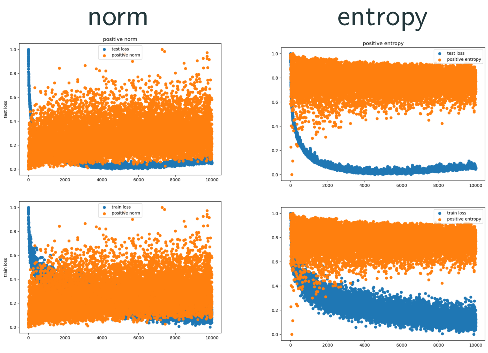
(a) FashionMNIST (good fit) (b) CIFAR10 (slight overfit)

Figure 1: Trend of the summary measures of the pos. LRP values and the model performance on FashionMNIST (a) and CIFAR10 (b).

Descriptive Analysis Results



(a) FashionMNIST (good fit)



(b) CIFAR10 (slight overfit)

Figure 2: Trend of the summary measures of the pos. Grad-CAM values and the model performance on FashionMNIST (a) and CIFAR10 (b).

Correlation Analysis Results

dataset	**FashionMNIST**				**CIFAR10**			
model fit	good fit		slight overfit		overfit		strong overfit	
summary	norm	entropy	norm	entropy	norm	entropy	norm	entropy
test loss	-0.875	0.806	0.226	-0.141	-0.363	0.391	0.345	-0.296
train loss	-0.657	0.591	-0.836	0.783	-0.832	0.808	-0.940	0.884

Figure 3: Correlation coefficients using LRP.

dataset	**FashionMNIST**				**CIFAR10**			
model fit	good fit		slight overfit		overfit		strong overfit	
summary	norm	entropy	norm	entropy	norm	entropy	norm	entropy
test loss	0.040	0.118	0.158	0.022	-0.086	0.055	0.035	-0.022
train loss	0.129	0.075	0.261	0.198	-0.108	0.050	-0.031	-0.020

Figure 4: Correlation coefficients using Grad-CAM.

Conclusion and Future Work

- **uncertainty** regarding the attribution values **decreases** during training
- **evolution** of the **norm** and **entropy** is **independent** of the **generalization** capabilities of the neural network
- **correlation** coefficients **differ** depending on whether the model is **overfitting**
- **Grad-CAM** values are very volatile and do not yield significant results

Future Work

- evaluate this behavior across model **architectures**, **datasets**, and **training times**
- possibly exploring the **double descent phenomenon**

Thank you for your attention!

References

[1] Amina Adadi and Mohammed Berrada. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[2] Sebastian Lapuschkin et al. “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”. In: *PLoS ONE* 10 (July 2015), e0130140. DOI: 10.1371/journal.pone.0130140.

[3] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *IEEE International Conference on Computer Vision, ICCV*. IEEE Computer Society, 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.