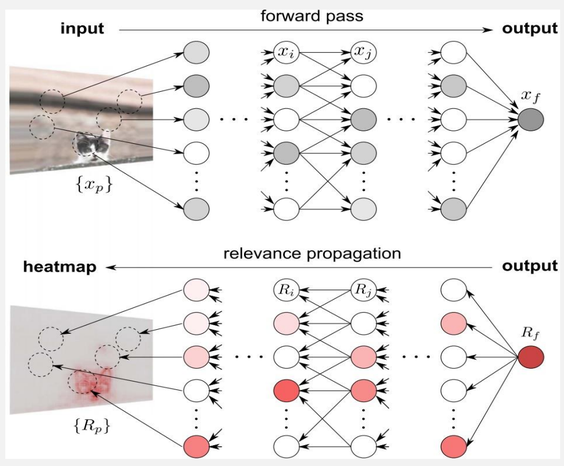
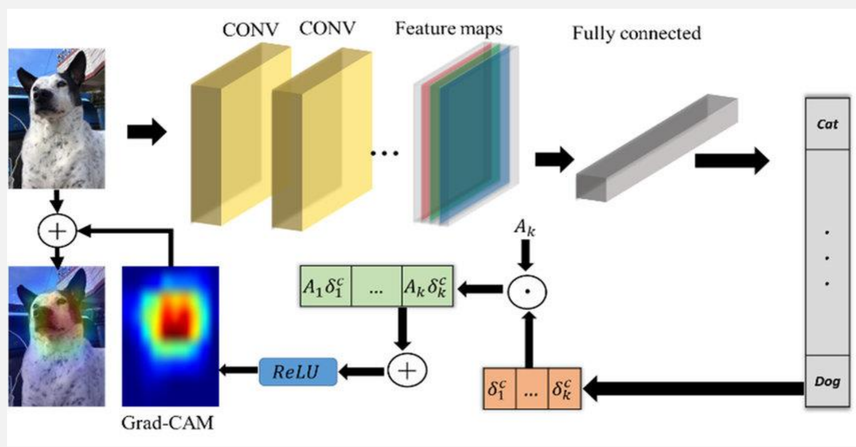
Identifying Trends in Feature Attributions during Training of Neural Networks

Elena Terzieva[1], Maximilian Muschalik[1], Paul Hofman[1], Eyke Hüllermeier[1]
(1) LMU Munich, Germany

Layer-wise Relevance Propagation (LRP)



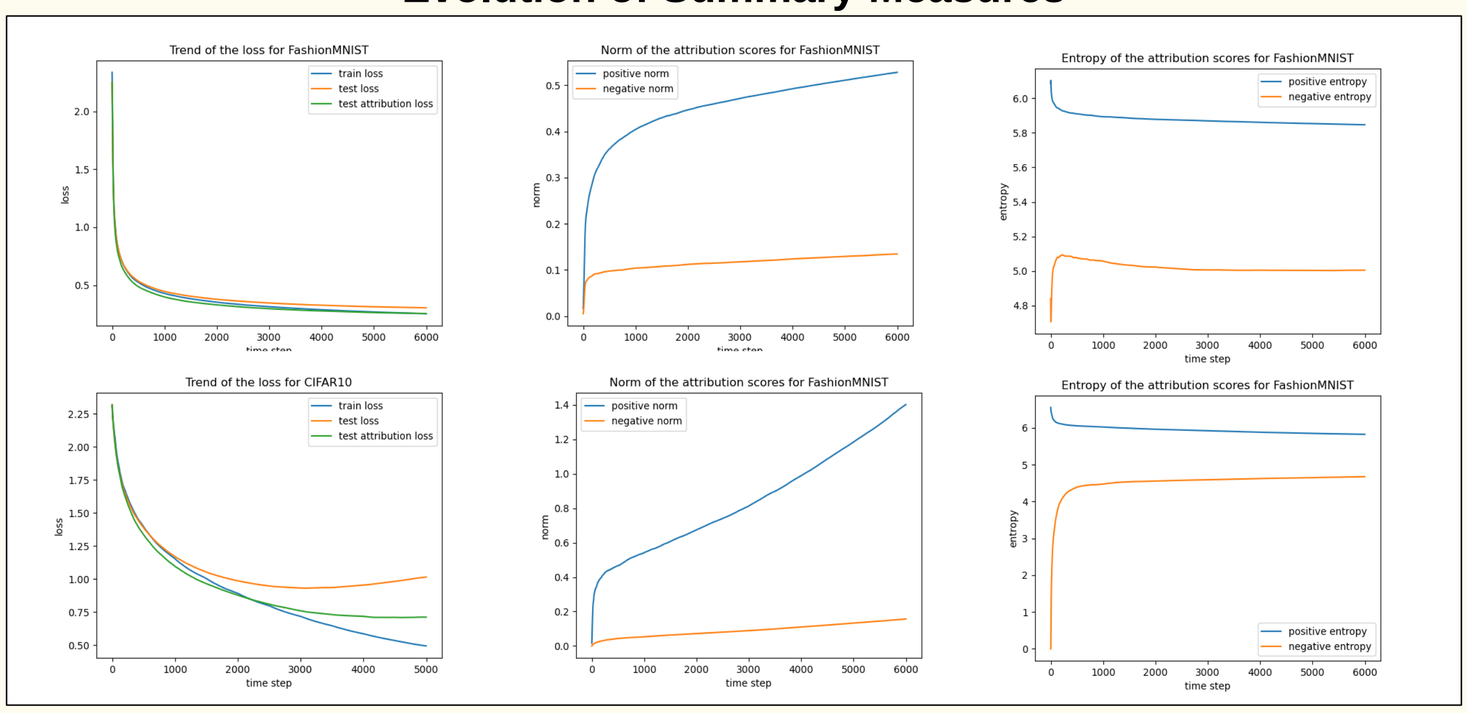
Gradient-weighted Class Activation Mapping (Grad-CAM)



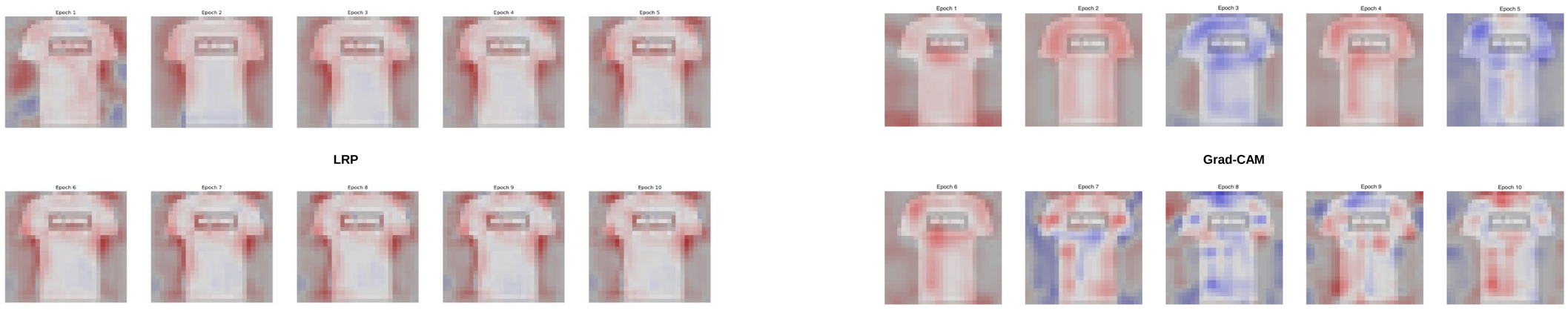
Methodology

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

Evolution of Summary Measures



Computed Explanations during Training



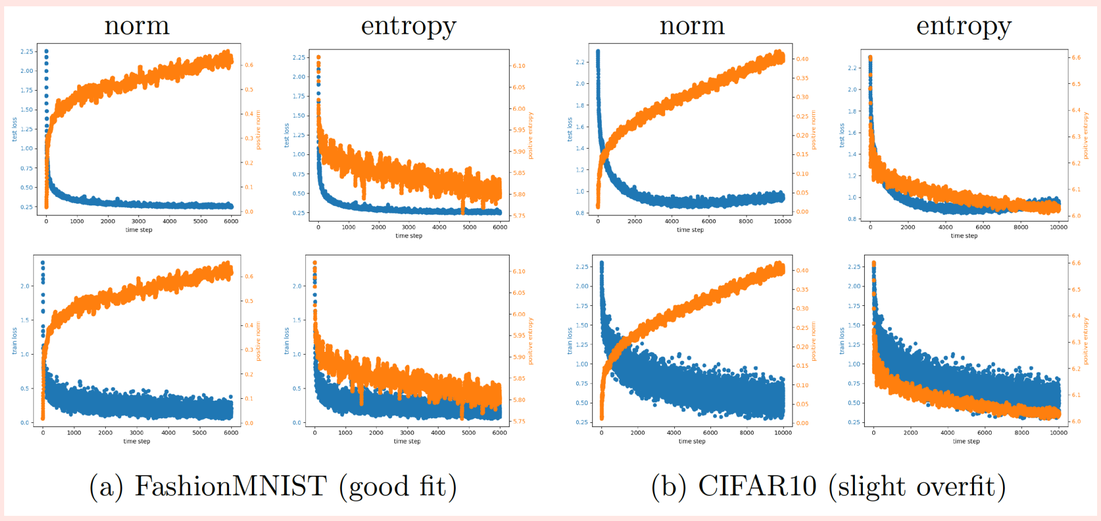
Results

- evolution of the norm and entropy is independent of the generalization capabilities of the neural network
- stronger correlation between the summary measures and the performance measures if the model is stable
- weak relationship between the summary measures and the test loss if the model is overfitting
- Grad-CAM values are very volatile und do not yield significant 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

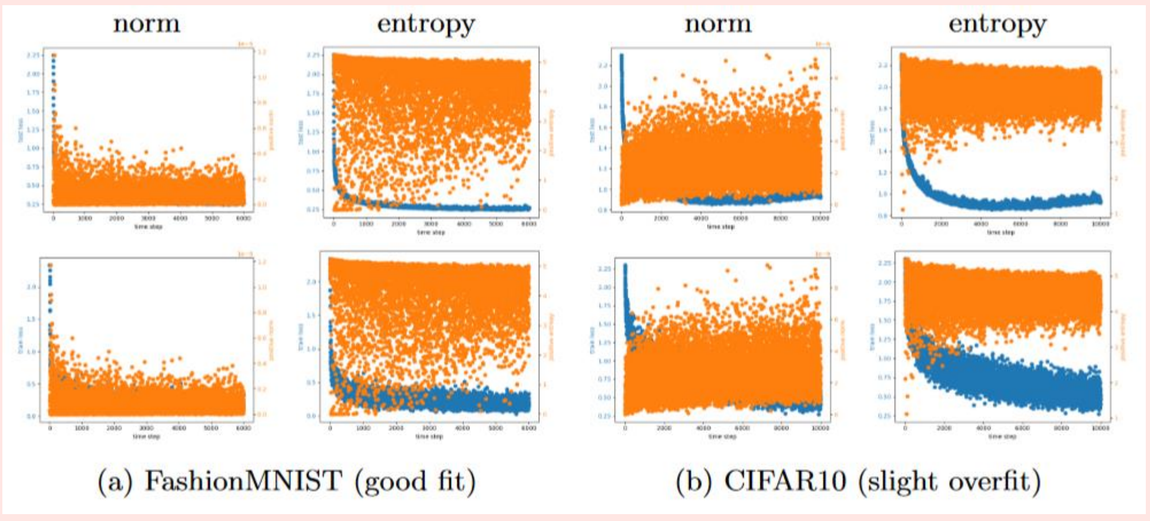
dataset	FashionMNIST				CIFAR10			
model fit	good fit		slight overfit		overfit		strong overfit	
summary	norm	entropy	norm	entropy	norm	entropy	norm	entropy
test loss	0.040	0.118	0.158	0.022	-0.086	0.055	0.035	-0.022
train loss	0.129	0.075	0.261	0.198	-0.108	0.050	-0.031	-0.020

LRP



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

GradCAM



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

References
Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLOS ONE, 10(7).
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.
Hüllermeier, E., Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning 110(3), 457–506

TRR 318 CONSTRUCTING EXPLAINABILITY
Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation