ETH zürich

Granger-causal Attentive Mixtures of Experts: Learning Important Features with Neural Networks

Patrick Schwab[1] (@schwabpa), Djordje Miladinovic[2], Walter Karlen[1] (@mhsl_ethz)

[1]Institute of Robotics and Intelligent Systems, ETH Zurich

[2]Department of Computer Science, ETH Zurich

1 Introduction

Knowledge of **the importance of input features** towards decisions made by machine-learning **models** is essential to **increase our understanding** of both the models and the underlying data.

Here, we present a new approach to learning to **produce** (1) **accurate predictions** and **(2) estimates of feature importance** in a **single model** in order to to improve our ability to **understand** its **predictions**.

2 Attentive Mixtures of Experts

Based on **neural soft attention** [1,2,3], we introduce a new model structure with the aim to ensure that

- each expert's contribution c_i can *only* be based on their respective input feature x_i
- the importance of c_i towards the final prediction y can *only* be increased by increasing the associated attention factor a_i

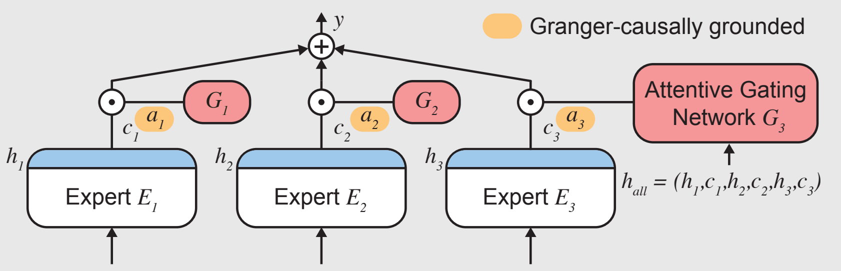


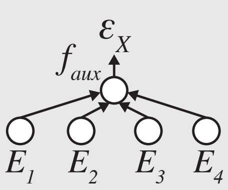
Figure 1. An overview of attentive mixtures of experts (AMEs). The attentive gating networks G_i (red) attend to the combined hidden state hall (blue). Each expert's G_i assigns an attentive factor a_i to opportunistically control its contribution c_i to the final prediction y.

3 Granger-causal Objective

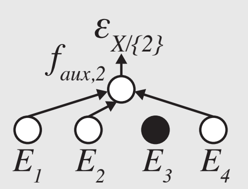
A **fundamental issue** of neural soft attention mechanisms is that they provide **no incentive** to learn feature representations that **accurately reflect feature importance**.

To address this issue, we introduce a **secondary Granger-causal objective** that **estimates the importance** of inputs, and **penalises** learning representations that do not **accurately reflect importance**.

The core idea of the Granger-causal objective is to **define feature importance** as the **reduction in prediction error** associated with adding that feature. We leverage the structure of AMEs to calculate the Granger-causal error at training time with auxiliary outputs f_{aux}.



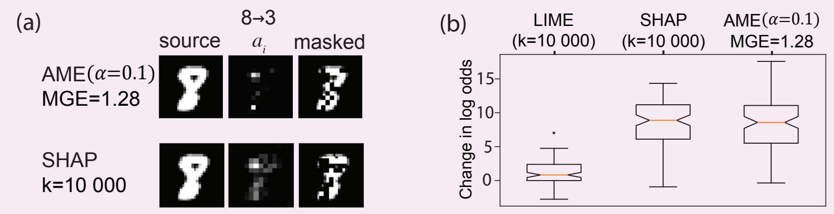
Error when considering all information



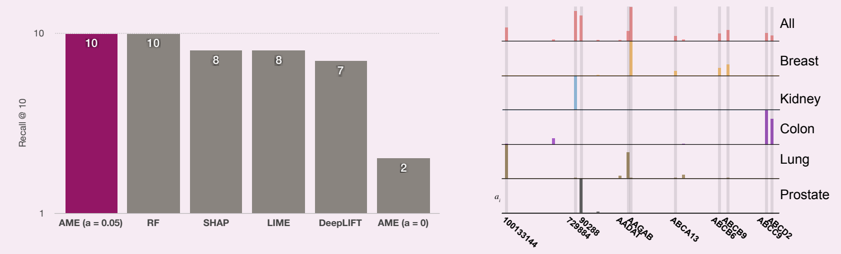
Error when considering all information apart from E_3

4 Results

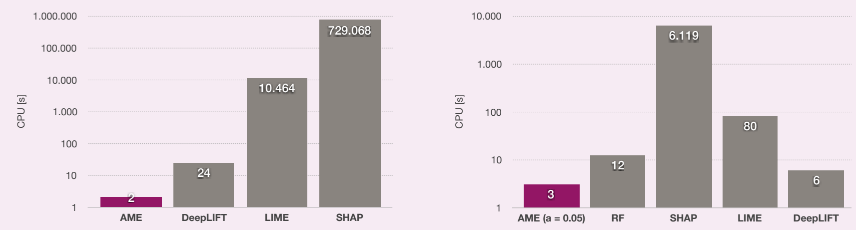
Feature importance estimation accuracy: Comparable to state-of-the-art methods on MNIST benchmark [4].



Discovered associations between genes and several cancer types are **consistent** with those reported by **domain experts**.



Computational performance: Orders of magnitude faster than existing methods at estimating feature importance.



5 Conclusion

We present a feature importance estimation approach that …

- **learns to estimate feature importance** from labelled data
- produces **predictions** and **importance scores** in a **single model**
- is **orders of magnitude faster** at estimating importance than perturbation-based approaches
- is **consistent with** associations reported by **domain experts**

6 References

1. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. ICML.
2. Rocktäschel, T.; Grefenstette, E.; Hermann, K. M.; Kocisky, T.; and Blunsom, P. 2016. Reasoning about entailment with neural attention. ICLR.
3. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A. J.; and Hovy, E. H. 2016. Hierarchical Attention Networks for Document Classification. In NAACL HLT.
4. Shrikumar, A.; Greenside, P.; Shcherbina, A.; and Kundaje, A. 2017. Learning important features through propagating activation differences. ICML.
5. Schwab, P., Miladinovic, D., and Karlen, W. 2019. Granger-causal Attentive Mixtures of Experts: Learning Important Features with Neural Networks. AAAI Conference on Artificial Intelligence.

AAAI | 2019

FNSNF Swiss National Science Foundation

75 NRP Big Data National Research Programme