Øster Farimagsgade 9, 4 TH.
1353 København K
Denmark
+45 44 12 18 06
ivar.cashin.eriksson@gmail.com

Ivar Cashin Eriksson

Research Statement

I am interested in exploring how sparse and interpretable internal representations emerge in large language models, and how such representations can be used to better understand and control model behaviour.

In earlier work, I used autoencoders to extract low-dimensional anatomical features from 3D medical images in order to build distance metrics between human anatomy. These would then be used in the design of automatic radiotherapy planning tools. In addition, I have used autoencoders to extract latent representations from invoicing agreements in order to be able to use them for downstream classification tasks, and to enable transfer learning between invoices based on different agreements. It has always been difficult to *know* that the latent representations contain the type of information I have wished it did. Because of this, when models have not behaved as expected, they have been exceedingly difficult to debug. This sparked an ongoing interest in how meaningful latent structure arises in deep models and how that structure might be made more transparent and actionable. I would be excited to pursue similar questions in the context of natural language processing, particularly in relation to concept bottleneck models and sparse probing techniques.

One possible direction would be to investigate whether sparse concept representations, discovered from model internals or guided by weak supervision, can act as useful intermediate decision layers in NLP models. This could involve identifying interpretable dimensions in pre-trained models, enforcing sparsity constraints to disentangle them, or evaluating how faithful such concept representations are to the model's actual reasoning. I am especially interested in methods of unsupervised structure discovery.

This is just one idea, and I am very open to refining it or exploring other directions depending on the needs and interests of the research group. My main goal is to contribute meaningfully to a longer-term research agenda focused on interpretability and structured understanding in NLP.

Thank you for considering my application.

Ivar Cashin Eriksson