# Related Work

### INTERPRETABLE FEATURES

Our work is inspired by research exploring the features that naturally occur in neural networks. Many models form at least some interpretable features. Word embeddings have semantic directions (*see* [8]). There is evidence of interpretable neurons in RNNs (*e.g.* [11, 12]), convolutional neural networks (*see generally e.g.* [13, 14, 41, 19]; *individual neuron families* [6, 18]), and in some limited cases, transformer language models (*see detailed discussion in our previous paper*). However this work has also found many "polysemantic" neurons which are *not* interpretable as a single concept [21].

### SUPERPOSITION

We're aware of two separate origins of the idea of superposition in neural networks. The first is the superposition hypothesis explored in this paper. The existence of polysemantic neurons (described in the previous section) led to the superposition hypothesis as one of the most plausible seeming explanations [1]. This hypothesis is a kind of "feature level" superposition.

Separately, Cheung *et al.* [7] explore what one might describe as "model level" superposition: can neural network parameters represent multiple completely independent models? Their investigation is motivated by catastrophic forgetting, but seems quite related to the questions investigated in this paper.

### DISENTANGLEMENT

The goal of learning *disentangled representations* arises from Bengio *et al.*'s influential position paper on representation learning [5]: "we would like our representations to *disentangle the factors of variation*... to learn representations that separate the various explanatory sources." Since then, a literature has developed motivated by this goal, tending to focus on creating genderantive models which separate out major factors of variation in their latent spaces.

Concretely, disentanglement research often explores whether one can train a VAE or GAN where basis dimensions correspond to the major features one might use to describe the problem (e.g. rotation, lighting, gender... as relevant). In the language of this paper, the goal is to impose a strong privileged basis on the latent space of a generative model, which are often totally rotationally invariant by default. Early work often focused on semi-supervised approaches where the features were known in advance, but fully unsupervised approaches started to develop around 2016 [42, 43, 44]

How does superposition relate to disentanglement? Although our investigation was motivated primarily by different examples, we see no reason to think that superposition doesn't also occur in the latent spaces of generative models. If so, it may be that superposition is a major reason why disentanglement is difficult. Superposition may allow generative models to be much more effective than they would otherwise be without. Put another way, disentanglement often assumes a small number of important latent variables explain the data. There are clearly examples of such variables, like the orientation of objects – but what if a large number of sparse, rare, individually unimportant features are collectively very important? Superposition would be the natural way for models to represent this.[22]