Example: Projecting a World Model onto a latent space

- Take a variational autoencoder (VAE) that was pre-trained on MNIST.
- Take the world_place.csv that was used in Gurnee and Tegmark (2023), which maps locations/areas to their latitude and longitude. For each place, it also indicates the corresponding country.
- Take the subset that contains countries that are currently part of the top 10 FIFA world ranking.
- Assign the current rank to each country, i.e. Argentina gets 1, France gets 2, …
- Encode all features excluding the FIFA world rank indicator but including `longitude` and `latitude` as continuous variables: $X^{(n \times d)}$ where n is the number of samples and d is the number of resulting features.
- Initialize a small neural network, let's call it a *projector*, that maps from X to a single small hidden layer with sigmoid activation and then form there to the latent space $\mathcal{Z} \subset \mathbb{R}$ of the pre-trained VAE.
- Without performing any training on the *projector*, simply compute a forward pass of the continuously encoded data X through the *projector* and store the output: $\mathbf{Z}^{(n \times k)}$ where k is the dimension of the latent space of the VAE.
- Next, we perform the linear probe on $\mathbf{Z}$ through Ridge regression: $\mathbf{W} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I})(\mathbf{Z}'\mathbf{Y})^{-1}$ where $\mathbf{Y}$ is the $(n \times 2)$ matrix containing the longitude and latitude for each sample.
- Finally, compute the predicted coordinate for each samples as $\widehat{\mathbf{Y}} = \mathbf{ZW}$
- Plot the results on a world map and be amazed.

  I think what's going on here is that we essentially randomly project a small set of features into a higher-dimensional space ($d < k$). The linear probe is then having an easy time separating features in the high-dimensional space, even though the projection is completely random. Even though one might be momentarily amazed by this, there is really nothing surprising at all going on here. We can still draw a very interesting parallel to Gurnee and Tegmark (2023): we know that geographical coordinates are in the training corpus of Llama-2; that information gets passed through a massive latent space (the model); they run probes on a subsapce of the model (still massive) and manage to linearly separate their variable of interest (coordinates). I don't know how accurate that parallel really is but it just seems striking to me that ultimately what they is regressing a 2-dimensional output variable on this massive feature matrix ("ranging from 4,096 to 8,192 features" for roughly 40,000 total samples, can't figure out how big the train/test sets are, respectively.). I'm not good enough at theory but maybe there's really just no surprise at all here?

Relatedly, it is worth pointing out that

- a recent paper has shown that the double-descent phenomenon is not exclusive to deep learning but also applies to linear regression with many features [LINK].

Example: PCA

Example: Autoencoder

Example: LLM

References

Gurnee, Wes, and Max Tegmark. 2023. "Language Models Represent Space and Time." https://arxiv.org/abs/2310.02207.