Notes: "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations"*

March 2nd

1 Notes:

1.1 Summary:

1. Unsupervised learning of disentangled representations is fundamentally impossible without inherint or explicit biases.
2. Good random initialization on bad model architectures frequently outperform bad random initializations on good model architectures.
3. Current disentanglement methods enforce independence between latent variables well (when measured by the μ parameter of the Gaussian), but this does not mean when *sampled* the dimensions stay factorized.
4. There is no concrete measure on how much disentanglement actually helps downstream tasks; the authors found it was not highly correlated when measuring disentanglement via certain metrics.

*Since this paper was mostly empirical and proved little theoretically, a "top down" approach on the notes does not work well here.

1.2 Impossibility theorem:

Theorem 1 *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits density $p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i)$. Then, there exists an infinite family of bijective functions $f : \text{supp}(\mathbf{z}) \to \text{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e.: $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \text{supp}(\mathbf{z})$ (i.e.: they have the same marginal distribution).*

We are defining "disentangled" in this following section to mean "a single change in the latent variable leads to a single high-level change in the representation".

Theorem 1 in more digestible words: given a multidimensional vector $p(\mathbf{z})$ with independent (factorized) dimensions, a generative model $p(\mathbf{x}|\mathbf{z})$, and a representation (that is the mean of the approximate posterior distribution) $Q(\mathbf{z}|\mathbf{x}) \equiv r(\mathbf{x})$ that is disentangled with respect to $\mathbf{z}$ in the generative model, Theorem 1 implies that there is an equivalent latent variable $\hat{\mathbf{z}} \equiv f(\mathbf{z})$ that is *completely* entangled with respect to $\mathbf{z}$ (every dimension of $\hat{\mathbf{z}}$ depends on every dimension of $\mathbf{z}$) and as a result is entangled with $r(\mathbf{x})$ as well. Due to the marginalization $P(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \equiv \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}$, the generative models are the same.

Since the disentanglement method only has access to $\mathbf{x}$, not $\mathbf{z}$, it cannot 'tell' between these two, and as such, is entangled with one.

This result is self evident. The main novelty is *complete* entanglement. Theorem 1 does not say disentanglement is impossible in practice, just that it is aided by bias about the underlying distribution of $\mathbf{x}$.

1.3 Empirical results:

The meat of this paper was the fact that they created over 12,000 models (>2.5 P100 GPU *years*). They found some pretty negative results on all things tested:

1. Different disentanglement metrics did not always have high correlation.
2. High disentanglement according to metrics did not have high correlation with increased downstream performance
3. Random initialization commonly mattered more than model architecture
4. All metrics tested required many ground-truth label or the true generative distribution to be known, which isn't applicable in practice

However, the paper was not perfect, as is noted by the authors results must be drawn with care. Some issues with the methodology were:

1. All datasets tested on were low dimension.
2. All models were purely unsupervised.
3. Their conclusion was 4 paragraphs.

They also provide a number of tips for how to select model architecture and hyperparameters:

1. End disentanglement was not highly correlated to annealing hyperparameters
2. KL / Reconstruction loss was also not highly correlated with disentanglement
3. There *was* correlation between disentanglement score of a model on similar datasets. If one has labels on a similar enough dataset to the end dataset, one can optimize hyperparameters for the one with labels then transfer them.
4. There was no way to distinguish a good architecture and a bad seed from a bad architecture with a good seed.