Unsupervised Disentanglement without Autoencoding: Pitfalls and Future Directions
Andrea Burns, Aaron Sarna, Dilip Krishnan, Aaron Maschinot

**Motivation: Disentanglement without VAEs**
We would like to learn disentangled visual representations for improved interpretability, data efficiency, and generalization. Prior work has focused on generative methods for disentanglement, which do not scale well to large datasets.

Thus, we explore regularization methods with contrastive learning, which could result in disentangled representations that are powerful enough for large scale datasets and downstream applications.

**Standard Contrastive [1]: Maximize Mutual Information (MI) between views:**
\[
\mathcal{L}_{\text{InfoMax}} = - \sum_{i \in I} \log \left( \sum_{a \in A} \exp \left( \frac{z_i \cdot z_j}{\tau} \right) \right)
\]

Extend to maximizing information between views of subembeddings, where the \(k\)th subembedding maps to a slice of the representation. In all of our experiments, \(K = 2\).

\[
\mathcal{L}_{\text{SubInfoMax}} = \sum_{k \in K} \left( - \sum_{i \in I} \log \left( \sum_{a \in A} \exp \left( \frac{z_{i,k} \cdot z_{j,k}}{\tau} \right) \right) \right)
\]

We add a regularization term \(R\) to this to obtain our final objective
\[
\mathcal{L}_{\text{Disentanglement}} = \mathcal{L}_{\text{SubInfoMax}} + \lambda \mathcal{R}
\]

**Approach: Minimize MI between subembeddings of the same view**
\[
\mathcal{R}_{\text{InfoMin}} = \sum_{k \neq k'} \sum_{i \in I} \log \left( \sum_{a \in A} \exp \left( \frac{z_{i,k} \cdot z_{i,k'}}{\tau} \right) \right)
\]

**Approach: Enforce orthogonality to approximate linear independence**
\[
\mathcal{R}_{\text{Ortho}} = \sum_{k \neq k'} \sum_{i} \sum_{j} \left| \frac{P(z_{i,k} \cdot z_{i,k'})}{\|z_{i,k}\| \|z_{i,k'}\|} \right|
\]

**Approach: Minimize element-wise dependencies between subembeddings using the Hessian of the loss**
\[
\mathcal{R}_{\text{Hess}} = \sum_{i \in I} \sum_{j \neq j'} \left| \frac{\partial \mathcal{L}_{\text{InfoMax}}}{\partial z_i} \cdot \frac{\partial \mathcal{L}_{\text{InfoMax}}}{\partial z_j} \right|^2
\]

**Experimental Setup**
1. Train encoder + projection network via contrastive loss with regularizer
2. Use resulting subembeddings to train different linear classifiers

**MNIST/STL-10 Dataset**
Overlay MNIST digit on STL-10 images &
Vary one factor in a view pair while two are fixed:
- Digit Class (DC)
- Digit Location (DL)
- Background Class (BC)

Want to disentangle the two fixed factors of variation in a view pair (DC-BC)
Then perform classification on those two tasks

**Ideal Disentanglement**

<table>
<thead>
<tr>
<th>Classification Input</th>
<th>DC</th>
<th>BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r)</td>
<td>97.3</td>
<td>64.5</td>
</tr>
<tr>
<td>(r_0)</td>
<td>97.3</td>
<td>10.1</td>
</tr>
<tr>
<td>(r_1)</td>
<td>11.7</td>
<td>64.5</td>
</tr>
</tbody>
</table>

\(|C(r_0) - C(r_1)| = 85.6\)

**References**