As per our discussion in May, I implemented an auto-encoder with latent state regularization given by the DrLIM loss. The motivation for doing this is to factorize the data into: (1) factor(s) implicitly conveyed by the 'similarity labels' (y) and (2) other information necessary for reconstruction. The loss is:

$$L = (1-y)\|U\left(G_w(X_1) - G_w(X_2)\right)\| + y\ max(0,\ m - \|U\left(G_w(X_1) - G_w(X_2)\right)\|) \\ + \alpha\left[\|R_w(G_w(X_1)) - X_1\|^2 + \|R_w(G_w(X_2)) - X_2\|^2\right]$$

- $G_w()$ is the encoder
- $R_w()$ is the decoder
- I call matrix U the 'selector'
- Call $z = G_w(X)$ features
- Call $f = Uz$ factors
- I refer to the DrLIM portion of the loss as 'metric loss'

I hope that the reasoning for my terminology will become clear as you read on. I focused on data-sets where the similarity label can be obtained naturally (via temporal consistency for example), as suggested in the original DrLIM paper. Following that paper I used the NORB data-set of rotated objects. The images can be sorted by factor (e.g. azimuth & elevation angles) where neighboring images can be considered similar ($y = 0$).

- Assume the pair $\{X_1, X_2\}$ are similar. In order to satisfy the metric loss, the factors ($f = UG_w(X)$) must be insensitive (invariant) to any variations between X_1 and X_2 (these variations are not captured by the similarity label)
- However, the features $z = G_w(X)$ cannot be invariant to any variations in X because they are used by the decoder to reconstruct the input
- Thus the invariant factors (f) are obtained via a simple linear transformation U on z, implying that z is factorized into invariant components (f) plus other information necessary to reconstruct X. In effect U simply selects the necessary factors in order to satisfy the similarity relationship implied by y.

Another way to interpret the metric loss is to relate it to the definition of sensitivity: i.e. $\|f(X_1) - f(X_2)\| \leq m\|X_1 - X_2\|$ for some m, note that $UG_w(X)$ plays the role of $f(X)$. If X_1 and X_2 are similar, then the functional $\|f(X_1) - f(X_2)\|$ is minimized w.r.t. f. In other words, f is made *insensitive* to any variation that may exist between X_1 and X_2. If X_1 and X_2 are deemed dissimilar then $UG_w(X)$ is made more sensitive to the factor variation that exists between X_1 and X_2.

Experiments

I used all the plane images from NORB with a fixed lighting condition, i.e. all 5 instances, all azimuths and elevations. As a first test I set the reconstruction weight (α) to zero and produced a mapping which projected the images to two dimensions (just to test vanilla DrLIM). I used a two layer fully connected *relu*-network for $G_w()$. The network had 1200 units in the first layer and 600 in the second.

In the 1^{st} experiment the factors which are 'implicitly conveyed in the similarity labels' are the azimuth, elevation angles, and instance:

- Planes from the same instance with neighboring azimuth and elevation angles are labeled as similar
- All other random pairs are labeled dissimilar which includes all pairs from different instances, i.e. inter-instance relationships are specified as dissimilar

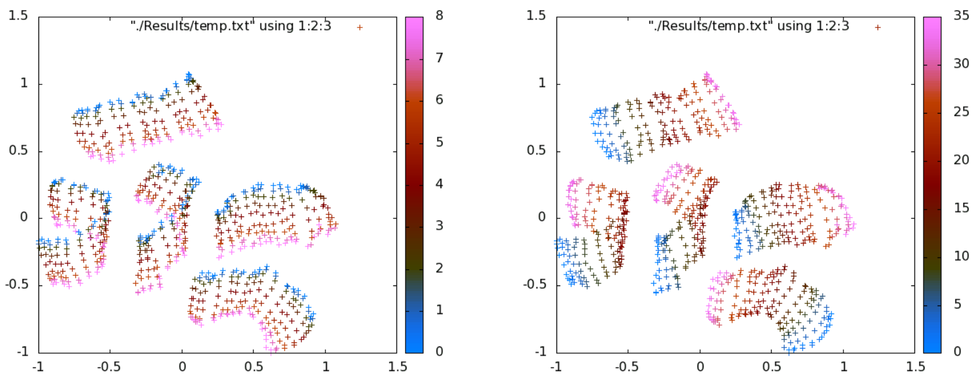

Figure 1: Left: output factor space colorized by elevation angle, Right: colorized by azimuth

Note since inter-instances relationships were specified, there are 5 distinct manifolds, each corresponding to an object instance. *It is unlikely for features to be shared between instances since they must be distinctly represented in factor space.* Note that in my implementation examples with azimuth 0 and 35 are not specified as neighbors, thus instance manifolds are not encouraged to wrap back on themselves.

In the 2^{nd} experiment, the factors which are 'implicitly conveyed in the similarity labels' are only the azimuth and elevation angles:

- No inter-instance relationships are specified, i.e. pairs from different instances are never presented to the network
- Samples from the same instance are labeled as similar if they are neighbors in azimuth and elevation

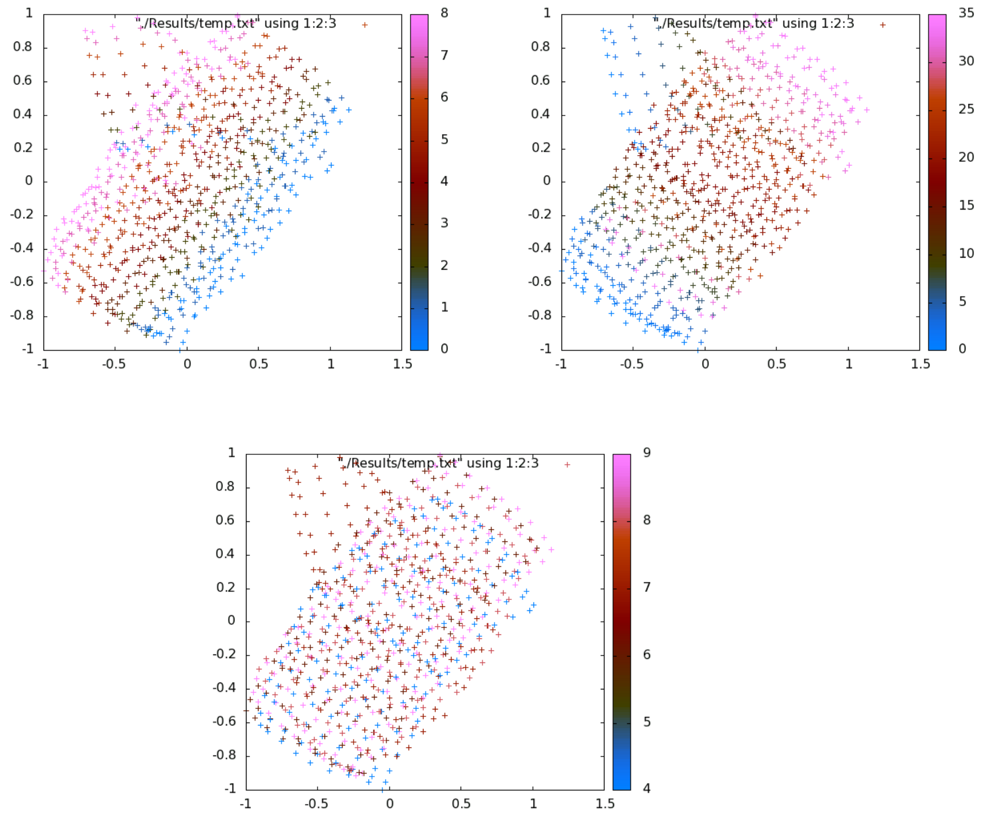

Figure 2: Factor space of the network trained with no reconstruction criterion ($\alpha = 0$). Left: elevation, Right: azimuth, Bottom: instance (note the invariance to instance)

Although instance invariance is not explicitly encouraged, the network is not penalized for producing a feature representation invariant to instance. In other words, there is no penalty for sharing features between instances.

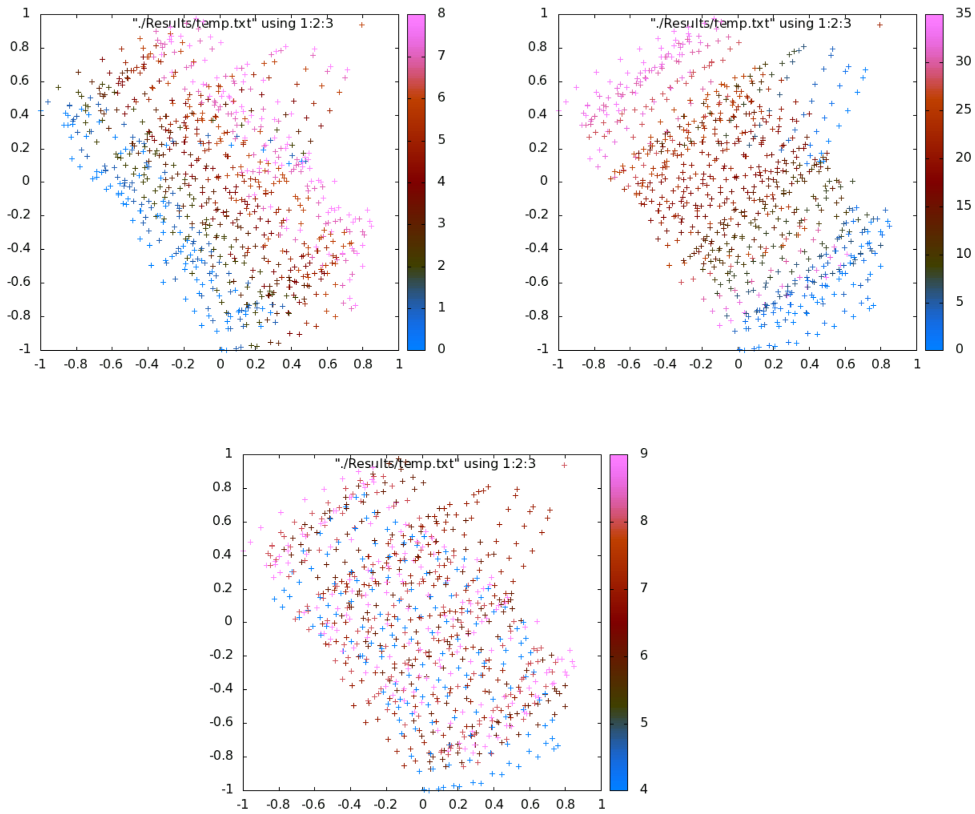

Figure 3: Factor space of the network trained with reconstruction criterion ($\alpha > 0$). Left: elevation, Right: azimuth, Bottom: instance (note the invariance to instance)

Note that both networks (trained with and without reconstruction) produce roughly the same factor representation, i.e. they can extract factors, $UG_w()$, which roughly correspond to azimuth and elevation (Figures 2 & 3) and are largely insensitive (almost invariant) to instance.

I then trained a linear decoder on top of the features i.e. the output of $G_w()$ which produces 600 outputs. Accurate reconstruction can't be (easily) obtained using the features produced by the network trained without a reconstruction criterion. Obviously, accurate reconstruction can be obtained using the features produced by the network trained with the reconstruction criterion.
The conclusion is:

- The features (and thus also the factors) produced by the network trained with no reconstruction loss turn to be invariant to everything (including instance) but azimuth and elevation
- The features produced by the network trained with reconstruction loss are not invariant and thus can be used to reconstruct the input
- The factors, $UG_w()$, are invariant to instance even in the network trained with reconstruction (see Figure 3).
- Invariance to instance in the former is achieved by multiplication by U ('pooling'). **Thus these features make it possible to separate the instance information from the rotation information using a linear transformation.**

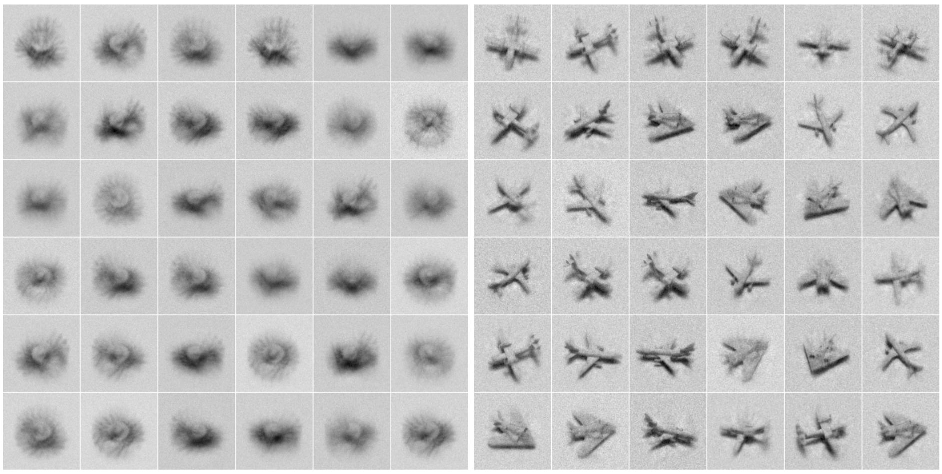

Figure 4: Left: reconstructions from the features learned from only metric loss (DrLIM) Right: reconstructions from features learned from metric and reconstruction loss