Non-Synergistic Variational Autoencoders

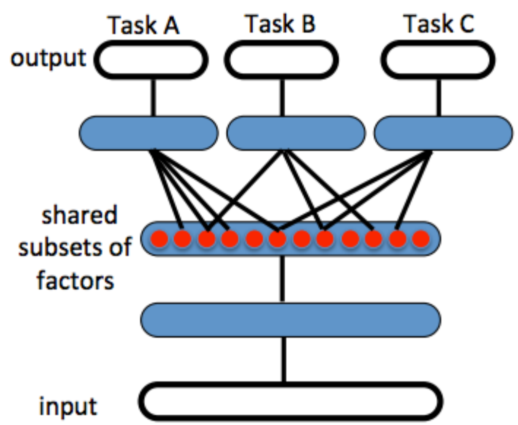
Gonzalo Barrientos

MSc. Machine Learning
University College London

LXAI Workshop @ NeurIPS 2018, Montréal, Canada

Representation Learning

- Why is it important to learn robust representations?
- Learning multiple levels of abstraction (Bengio and LeCun, 2007). Deep Learning allows higher layers of abstraction, which disentangle the factors of variation. Usually it overfits to a particular training task.
- Disentangle this factors explicitly allows a better generalization and domain transfer.

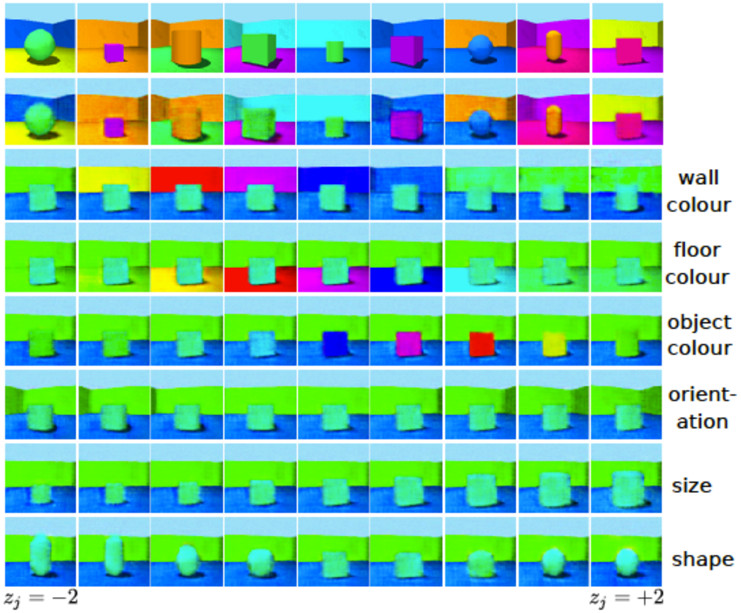


(Bengio, 2014)

Disentangled Representations

- Bengio (2013) described that representations should be factorized and interpretable.

- It allows the model to learn the structure of the world without any supervision
- Generalize knowledge between different tasks.
- Data efficiency.
- Latent manipulation.
- Compositionality



(Kim and Mnih, 2017)

Mutual Information Decomposition

Information Theory

- Entropy H(X):

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

- Mutual Information I(X;Y):

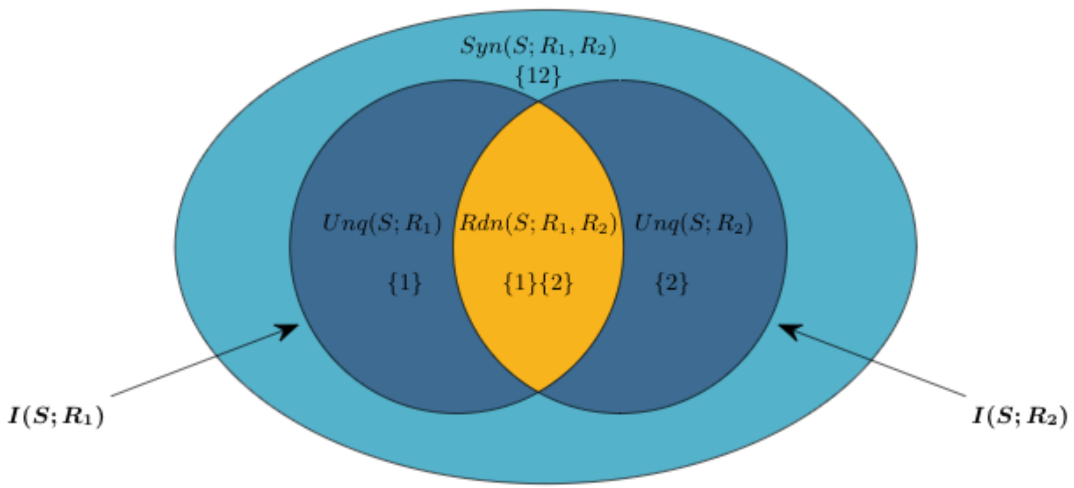
$$I(X;Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- Mutual Information & KL:

$$I(X;Y) = \mathbb{E}_{p(x,y)} \left[\log \frac{p(x,y}{p(x)p(y)} \right] = D_{KL} \left[p(x,y) \parallel p(x)p(y) \right]$$

MI decomposition: 2 variables

- Random variable S and a random vector R = {R1, R2, …Rn}.
- Williams and Beer (2010) introduced the PI-diagram: {12} needs both R1 and R2, {1}{2} means that the information provided by R1 is the same as R2.

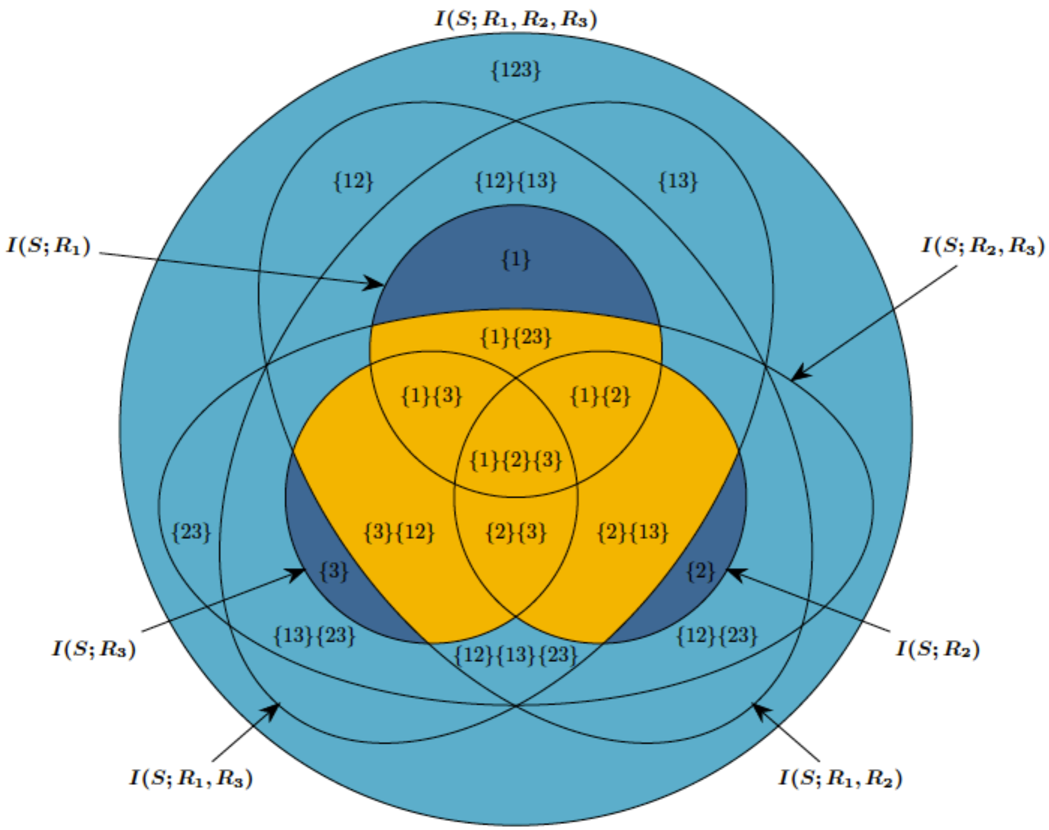


$$I(S;R_1) = \underbrace{Rdn(S;R_1,R_2)}_{\text{Redundant}} + \underbrace{Unq(S;R_1 \setminus R_2)}_{\text{Unique}}$$

$$I(S;R_2) = \underbrace{Rdn(S;R_1,R_2)}_{\text{Redundant}} + \underbrace{Unq(S;R_2 \setminus R_1)}_{\text{Unique}}$$

$$I(S;R_1,R_2) = \underbrace{Rdn(S;R_1,R_2)}_{\text{Redundant}} + \underbrace{Unq(S;R_1 \setminus R_2)}_{\text{Unique}} + \underbrace{Unq(S;R_2 \setminus R_1)}_{\text{Unique}} + \underbrace{Syn(S;R_1,R_2)}_{\text{Synergistic}}$$

MI decomposition: 3 variables



- Difficult to compute the synergy as we increase the number of predictors Ri.

- Yellow: Redundant MI
- Blue: Unique MI
- Light blue: Synergistic MI

Synergy

Definition

- Canonical example: XOR gate. We need X1 and X2 to fully specified the value of Y.

$X1$	$X2$	Y
0	0	0
0	1	1
1	0	1
1	1	0

$$I(X_1 X_2; Y) = H(Y) = 1 \text{ bit}$$

$$I(X_1; Y) = I(X_2; Y) = 0 \text{ bit}$$

- Useful notation:

- Ai: Subset of individual predictors: X_1, X_2, .. Xn
- {X_1, X_2, .., Xn} : set of all individual predictors
- Y: random variable to predict
- y: particular outcome of Y.

Metric: Imax Synergy

- Whole beyond the maximum of its parts

$$
\begin{aligned}
S_{max}(\{X_1, X_2, ..., X_n\}; Y) &= I(\boldsymbol{X}; Y) - I_{max}(\{A_1, A_2..A_n\}; Y) \\
&= I(\boldsymbol{X}; Y) - \sum_{y \in Y} p(Y = y) \max_i I(A_i; Y = y)
\end{aligned}
$$

The specific mutual information between Ai and the outcome “y” could be expressed as a KL term:

$$
S_{max}(\{X_1, X_2, ..., X_n\}; Y) = I(\boldsymbol{X}; Y) - \sum_{y \in Y} p(Y = y) \max_i D_{KL}\left[P(A_i \mid y) \parallel P(A_i)\right]
$$

This metric is bounded between the total mutual information I(X,Y) and 0.

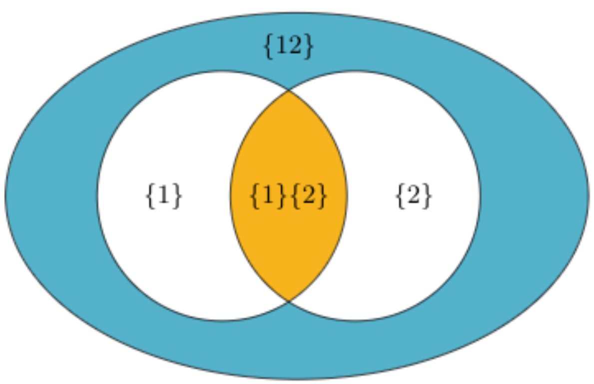
$$
0 \leq S_{max}(\{X_1, X_2, ..., X_n\}; Y) \leq I(\boldsymbol{X}; Y)
$$

Metric: Whole Minus Sum Synergy

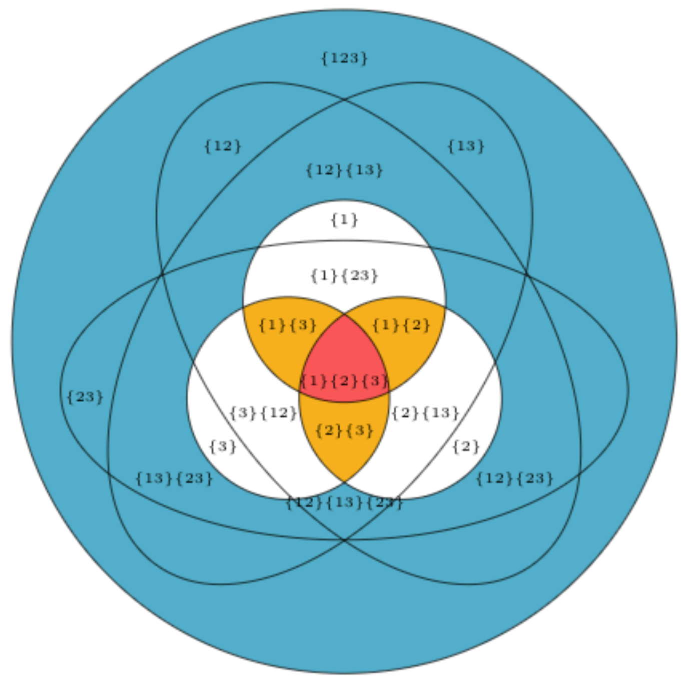
- Whole minus the sum of all mutual information of the individual predictors.

$$WMS(\{X_1, X_2, ..., X_n\}; Y) = I(\boldsymbol{X}; Y) - \sum_{i}^{n} I(X_i; Y)$$

Drawback: Counts the redundant information many times.



(Griffith and Koch, 2014)



Metric: S_{VK} Whole Minus Union Synergy

- Try to solve the issue with the multiple counts by preserving only the pairwise interactions P (Xi, Y) between each predictor Xi and Y.

$$S_{VK}(\{X_1, X_2, ..., X_n\}; Y) = I(\boldsymbol{X}; Y) - I_{VK}(\{X_1, X_2, ..., X_n\}; Y)$$

$$\begin{aligned} I_{VK}(\{X_1, X_2, ..., X_n\}; Y) = \underset{P^*(X_1, X_2, .., X_n, Y)}{\text{minimise}} \quad & I^*(\boldsymbol{X}; Y) \\ \text{subject to} \quad & P^*(X_i, Y) = P(X_i, Y)\ \forall i \end{aligned}$$

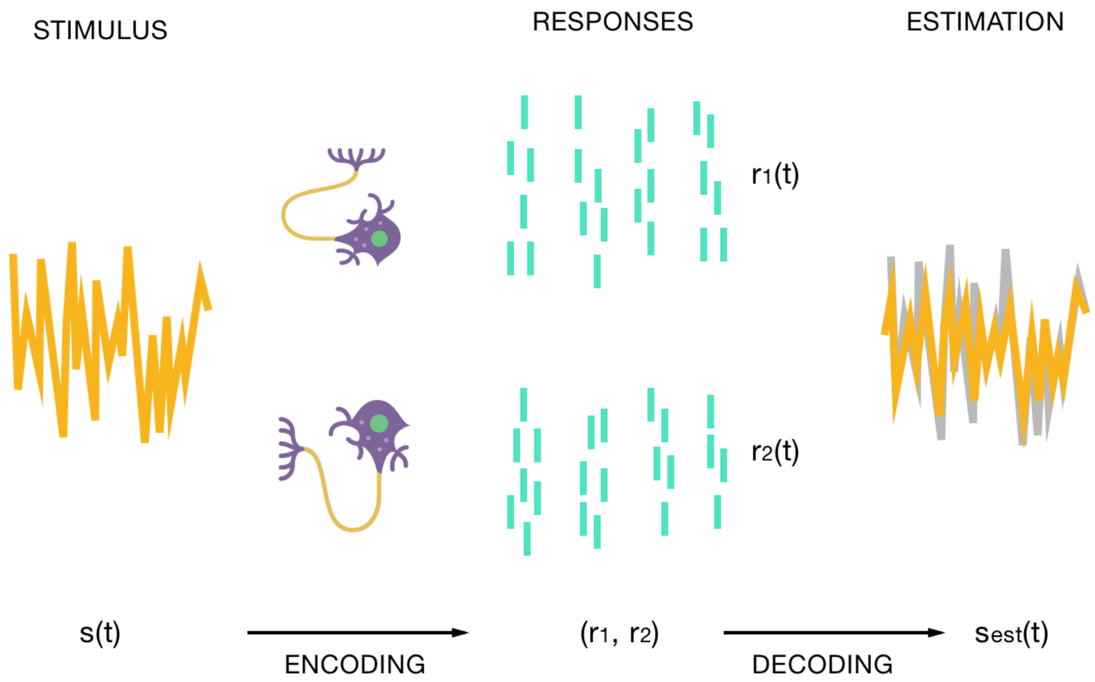
$$\text{where,} \quad I^*(\boldsymbol{X}; Y) = D_{KL}\big[P^*(\boldsymbol{X}, Y) \parallel P^*(\boldsymbol{X})P^*(Y))\big]$$

- The following bound can be proved:

$$\max[\ 0, WMS(\{X_1, X_2, ..., X_n\}; Y)]\ \leq S_{VK}(\{X_1, X_2, ..., X_n\}; Y) \leq S_{max}(\{X_1, X_2, ..., X_n\}; Y)$$

Synergy - Neuroscience

- MI is used to measure the correlation between the stimulus and the responses (spike trains). If the responses convey more information together than separate, there is a synergistic component.



Model

Related work

- B-VAE: Constraints the capacity of the latent information channel (Higgins et al. , 2017)

$$\mathbb{E}_{q_\phi(z|x)}\left[\ \log p_\theta(x|z)\right]\ -\beta D_{KL}\left[q_\phi(z|x)\parallel p(z)\right]$$

- Factor VAE: Minimizes total correlation in the latent z (Kim and Mnih, 2017)

$$\sum_{i=1}^{N}\mathbb{E}_{q_\phi(z|x^{(i)})}\left[\ \log p_\theta(x^{(i)}|z)\right]\ -D_{KL}\left[q_\phi(z|x^{(i)})\parallel p(z)\right]-\gamma D_{KL}\left[q(z)\parallel\prod_{j}^{D}q(z_j)\right]$$

- InfoGAN: Maximizes the mutual information between the latent code c and the output of the Generator that receives incompressible noise z and the latent c, I (c; G(z,c)). (Chen et al. 2016)

$$\min_{G,Q}\max_{D}V_{InfoGAN}(D,G,Q)=V(D,G)-\lambda L_I(G,Q)$$

Intuition

- Synergy not desirable for the task of disentanglement. We want the latents to be independently informative as possible about the data.

- We hypothesize that by minimizing the synergistic mutual information within the latents, we encourage the disentanglement of the factors of variation.

- Since our objective is to minimize the synergy, it makes sense to use the overestimate of the synergy, Smax, as the metric.

$$S_{max}(\{Z_1, Z_2, ..., Z_d\}; X) = I(\boldsymbol{Z}; X) - \sum_{x \in X} p(X = x) \max_i D_{KL}\left[q_\phi(\mathbb{A}_i \mid x) \parallel p(\mathbb{A}_i)\right]$$

Non-Synergistic VAE

- We decided to use the ELBO loss as the related methods. We subtract the synergy term, where α is hyper-parameter.

$$\mathcal{L}_{new}(\theta, \phi; x, z, \alpha) = \mathcal{L}_{elbo}(\theta, \phi; x, z) - \alpha\Big(I(z; x) - \sum_{x \in X} p(X = x) \max_i D_{KL}\big[q_\phi(\mathbb{A}_i \mid x) \parallel p(\mathbb{A}_i)\big]\Big)$$

- Hoffman (2016) proposed a different way to express to ELBO using the empirical data distribution:

$$\frac{1}{N} \sum_{n=1}^{N} D_{KL}\big[q_\phi(z_n \mid x_n) \parallel p(z_n)\big] = D_{KL}\big[q_\phi(z_n) \parallel p(z_n)\big] + I(x_n; z)$$

Non-Synergistic VAE

- Putting all together:

$$\mathcal{L}_{new}(\theta,\phi;x,z,\alpha) = \frac{1}{N}\sum_{i=1}^{N}\Big[\ \mathbb{E}_{q_\phi(z|x)}\big[\ \log p_\theta(x^{(i)} \mid z)\big]\ \Big] - D_{KL}\big[q_\phi(z_n) \parallel p(z_n)\big] - I(x_n;z)$$
$$\underbrace{-\alpha I(x_n;z)}_{\text{Penalise}} \underbrace{+\alpha \sum_{x\in X} p(X=x)\max_i D_{KL}\big[q_\phi(\mathbb{A}_i \mid x) \parallel p(\mathbb{A}_i)\big]}_{\text{Imax}}$$

- Kim and Mnih (2017) showed that penalizing the MI even more is not desirable.

$$\mathcal{L}_{new}(\theta,\phi,x) = \mathcal{L}_{elbo}(\theta,\phi,x) + \alpha \sum_{x\in X} p(X=x)\max_i D_{KL}\big[q_\phi(\mathbb{A}_i \mid x) \parallel p(\mathbb{A}_i)\big])$$

Non-Synergistic VAE

- Not a guaranteed lower bound on the likelihood. We changed the function:

$$\mathcal{L}_{new}(\theta, \phi, x) = \mathcal{L}_{elbo}(\theta, \phi, x) - \alpha \sum_{x \in X} p(X = x) \min_i D_{KL}\big[q_\phi(\mathbb{A}_i \mid x) \parallel p(\mathbb{A}_i)\big])$$

- Computing per outcome was too expensive, we used a mini-batch approximation, where Aw is the set of latent dimensions that provide the least amount of information about a batch of training examples:

$$\mathcal{L}(\theta, \phi; x, z, \alpha) = \underbrace{\mathbb{E}_{q_\phi(z|x)}\big[\ \log p_\theta(x|z)\big]\ - D_{KL}\big[q_\phi(z|x) \parallel p(z)\big]}_{\mathcal{L}_{elbo}} - \underbrace{\alpha D_{KL}\big[q_\phi(\mathbb{A}_w|x) \parallel p(\mathbb{A}_w)\big]}_{\mathcal{L}_{syn}}$$

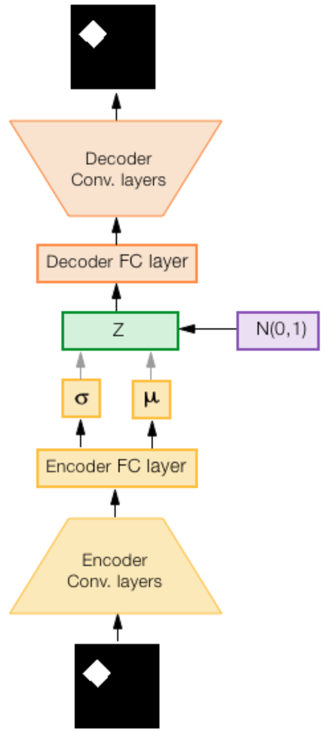
Training

- Two – step optimization. First the ELBO and then the synergy loss.

$$A_i = \arg\min_i D_{KL}\left[q_\phi(\mathbb{A}_i \mid x) \parallel p(\mathbb{A}_i)\right])$$

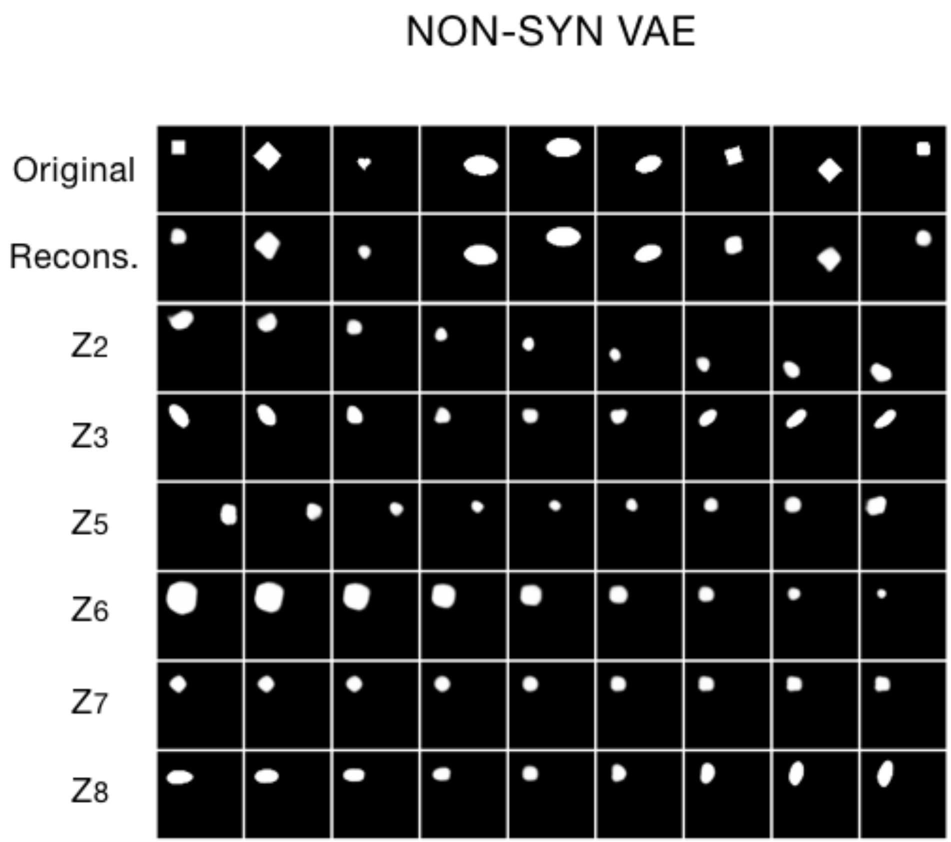
- Greedy approximation to compute the equation above to get the set of latent dimensions.

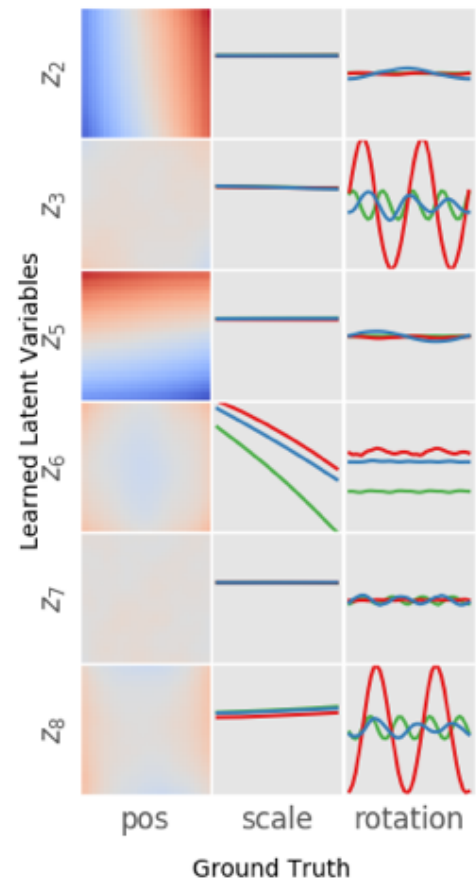
- Sample the values of mu and log var for the 2nd step (synergy loss)



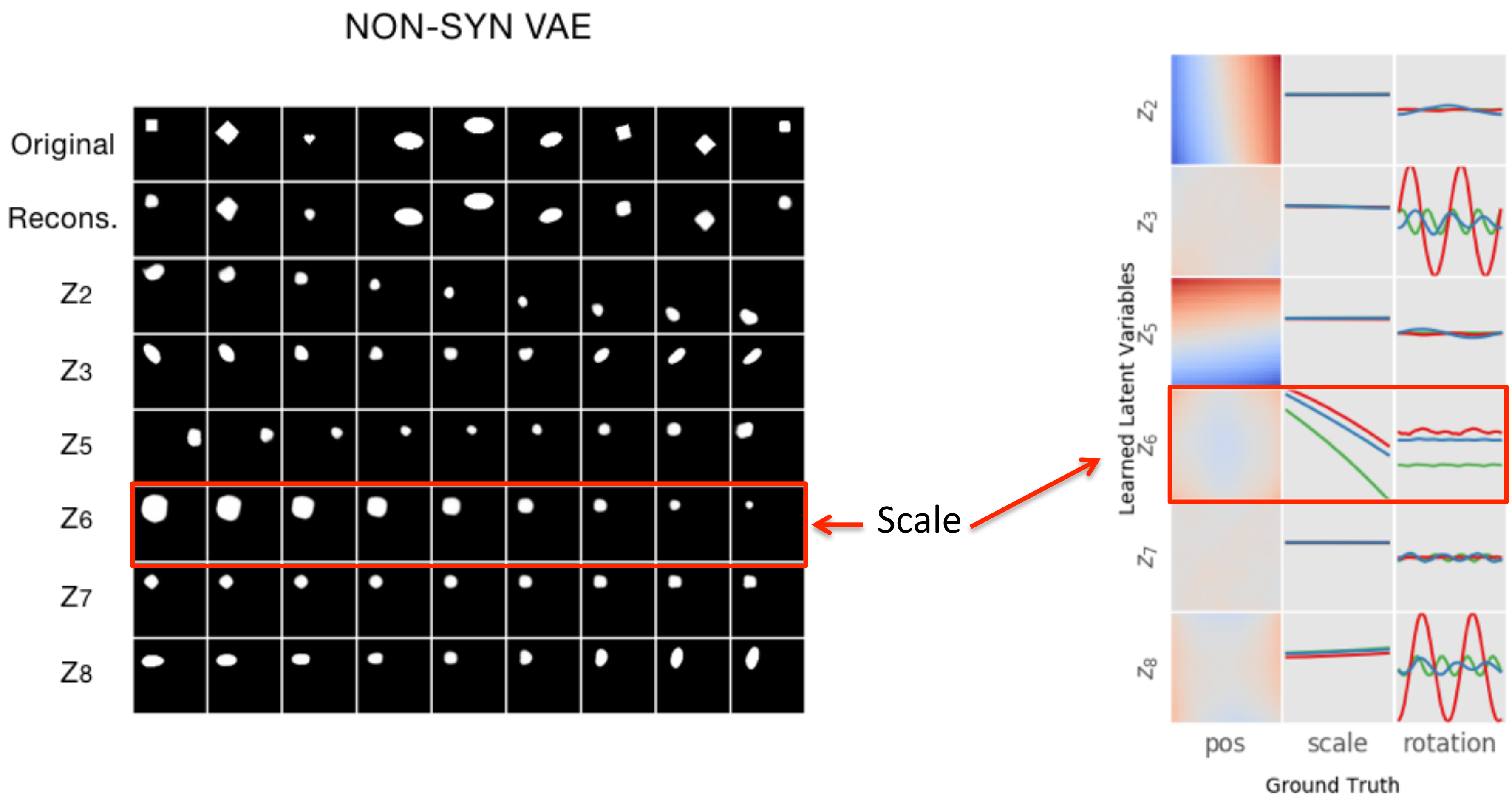
Experiments

Latent traversal and Mean activation

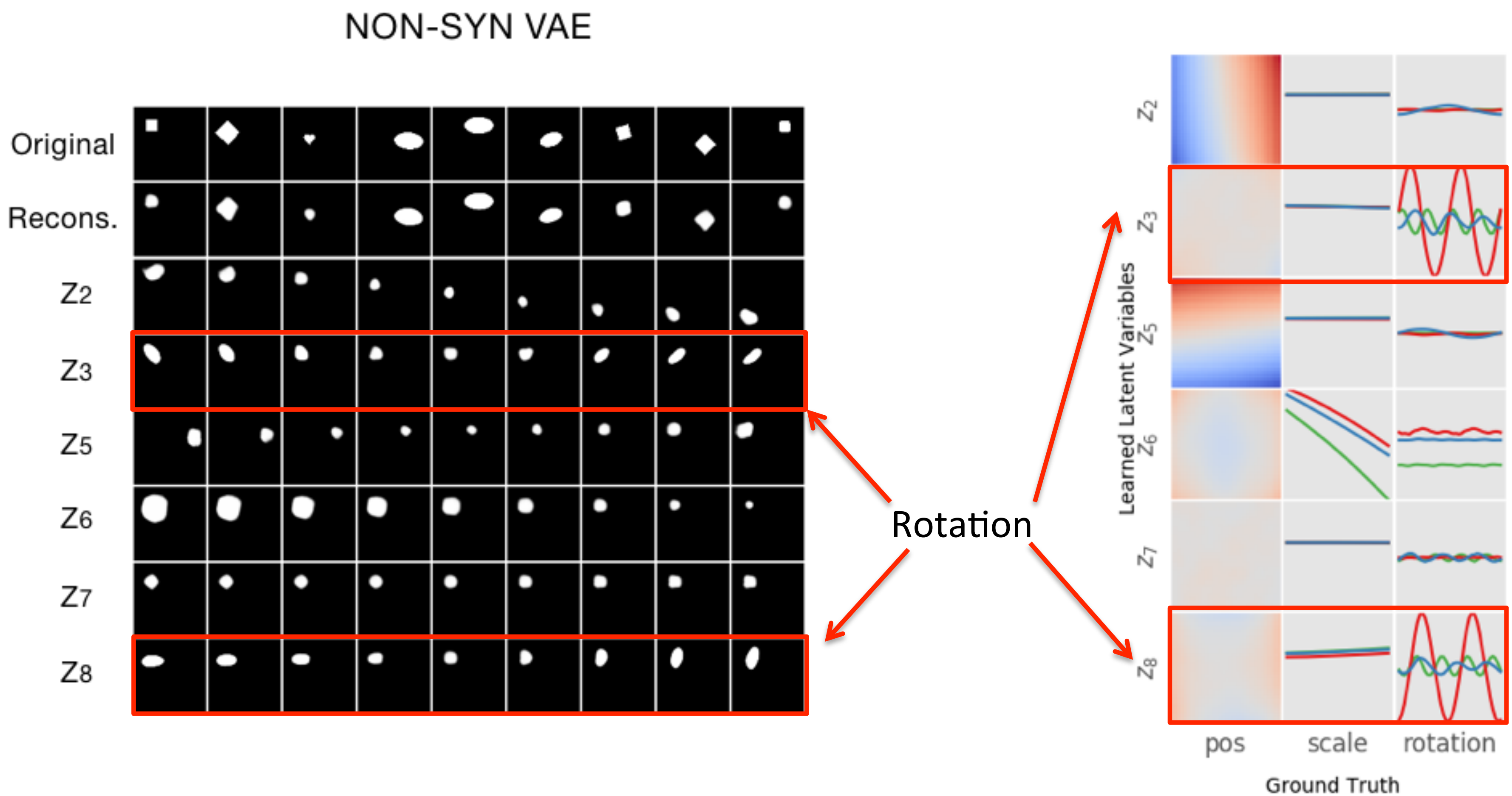




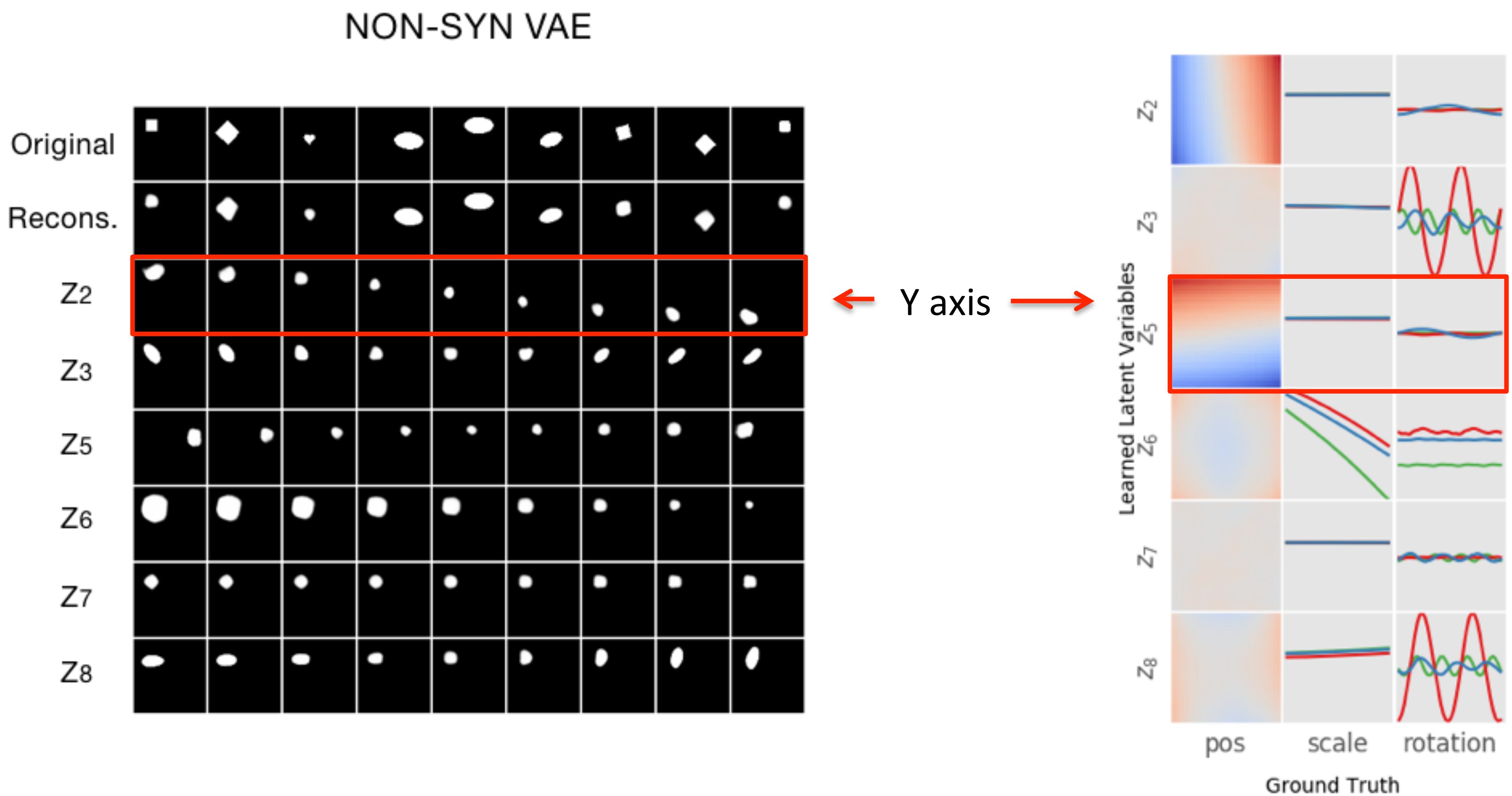
Latent traversal and Mean activation



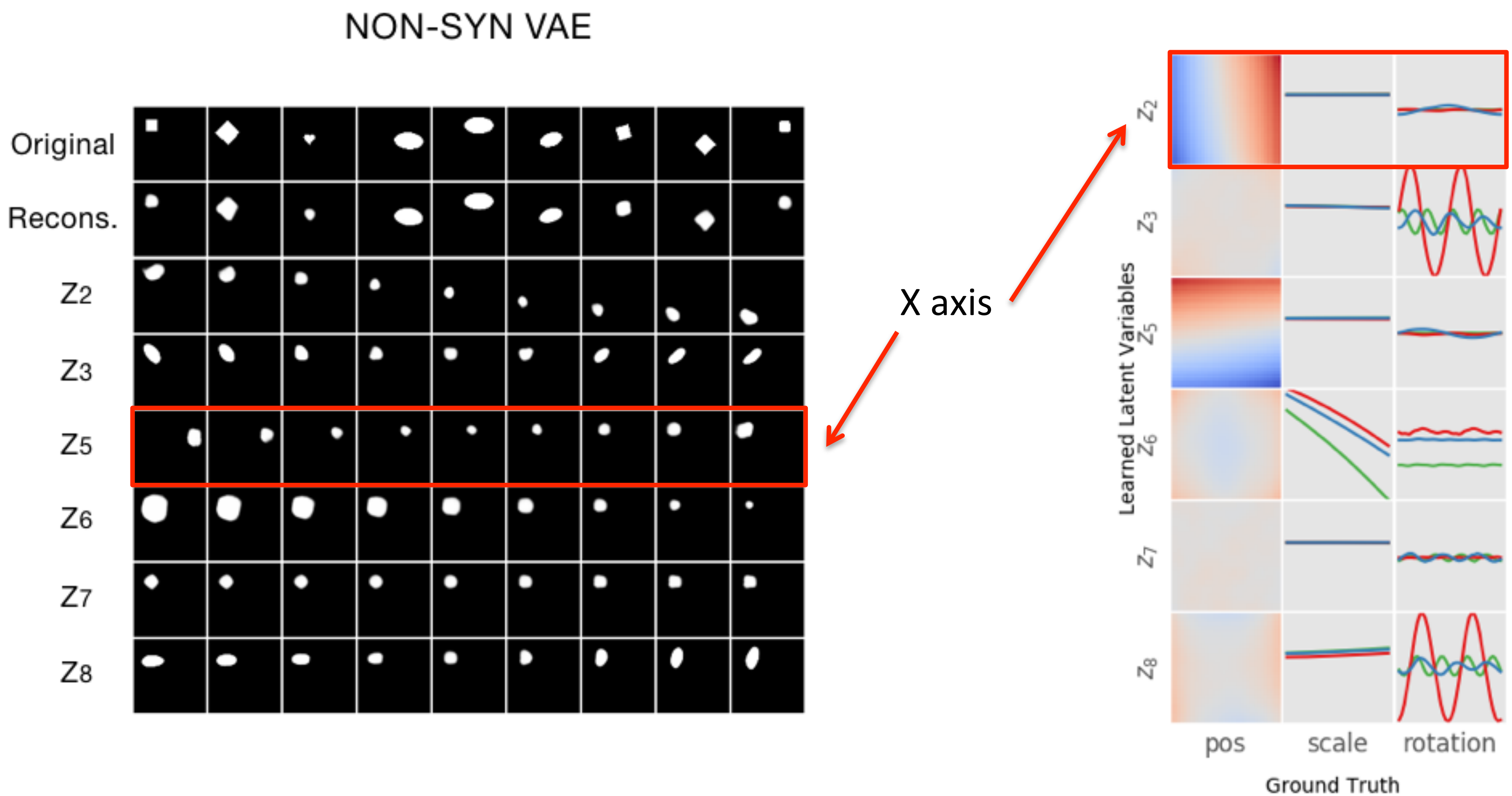
Latent traversal and Mean activation



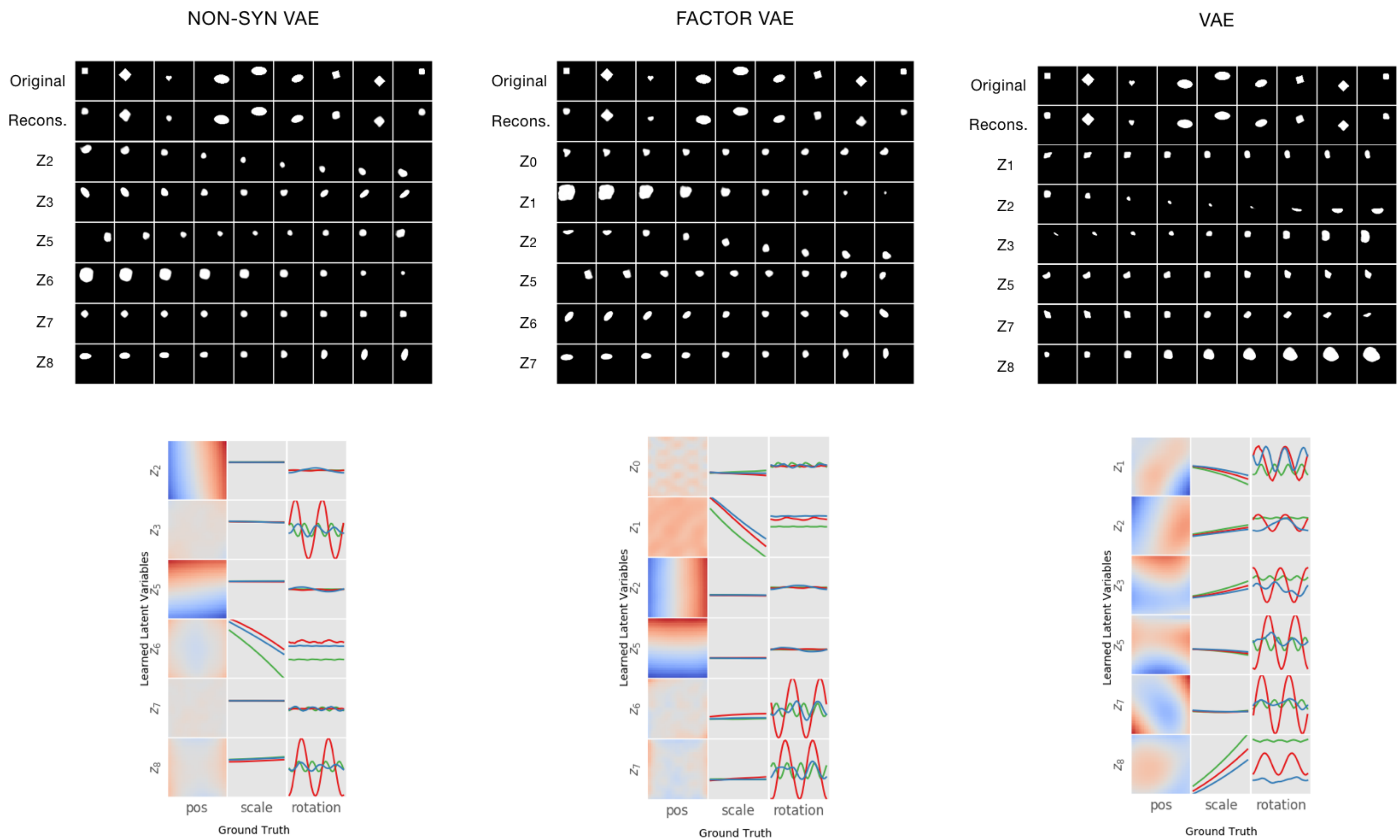
Latent traversal and Mean activation



Latent traversal and Mean activation



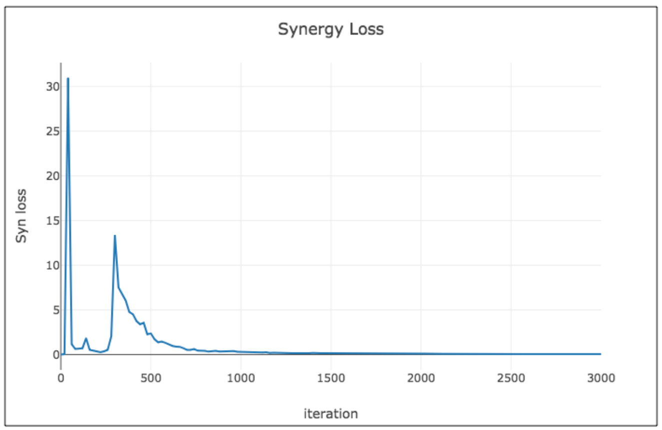
Comparison with baselines



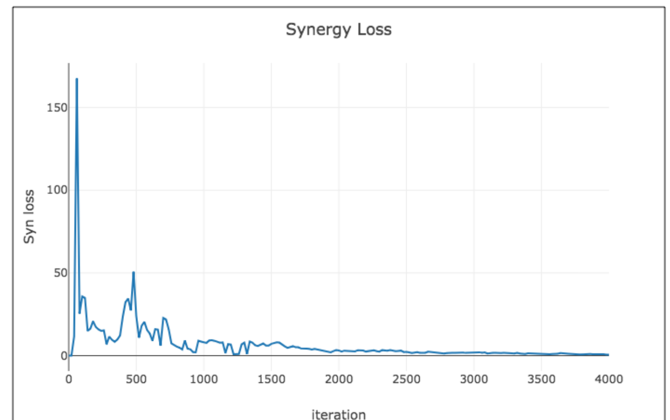
Comparison with Factor VAE

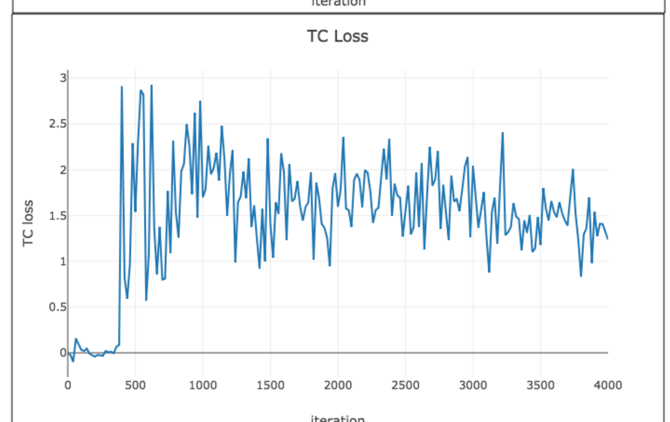
- We found that Factor VAE minimizes the synergy implicitly.

NON-SYN VAE

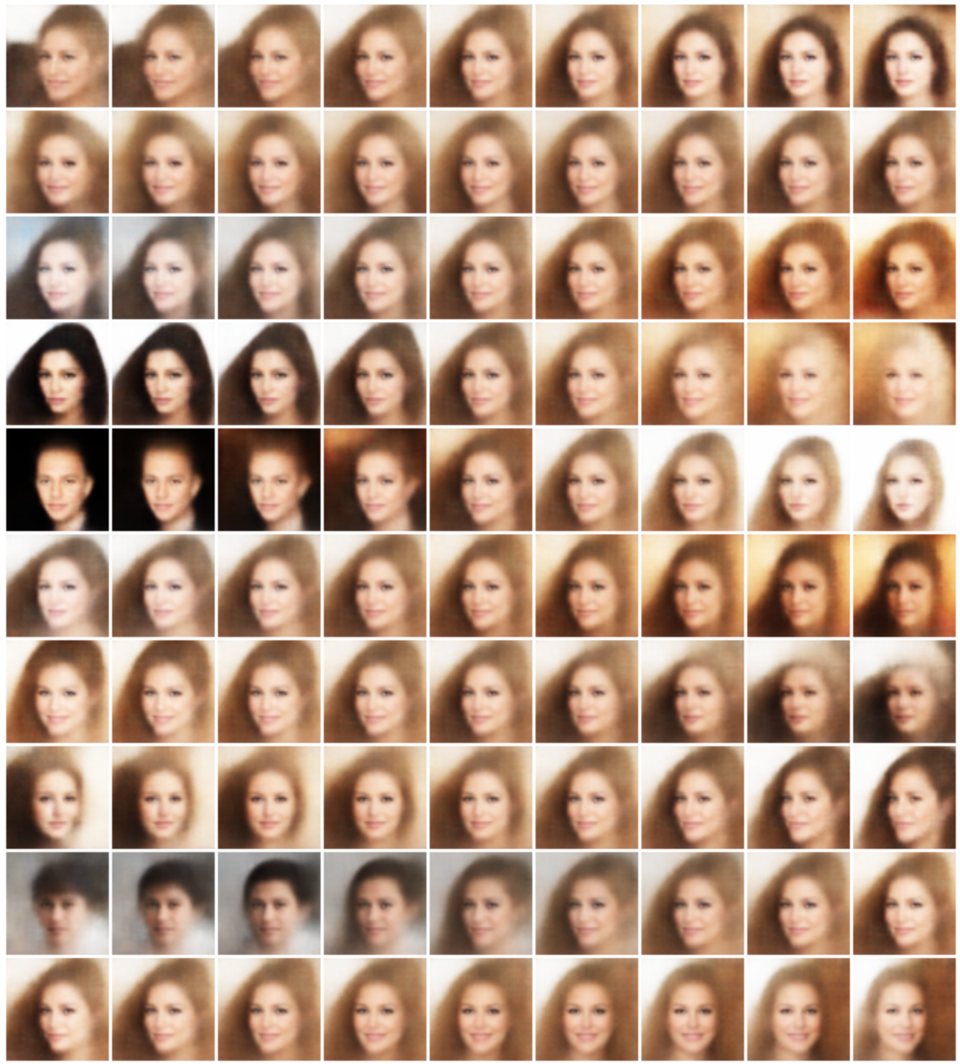


FACTOR VAE





CelebA – Traverse latents



Chairs – traverse latents

Conclusions

- Learning disentangled representations in an unsupervised setting could be useful to build the path towards the creation of truly intelligent machines, since our models will be able to understand the structure of the world.
- Fields such as neuroscience or information theory provide a useful insight that could inspire the next state of the art models.
- Future work needs to find a way to learn representations from different visual domains, such as the work from Achile et al., 2018 (presented at this conference)

References

- Y. Bengio, A. C. Courville, and P. Vincent. Representation learning: A review and new perspectives.
- N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost. Shared information – new insights and problems in decomposing information in complex systems. CoRR, abs/1210.5902, 2012.
- T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. CoRR, abs/1802.04942, 2018
- I. Gat and N. Tishby. Synergy and redundancy among brain cells of behaving monkeys. In Advances in Neural Information Processing Systems 11, [NIPS Conference, Denver, Colorado, USA, November 30 - December 5, 1998], pages 111–117, 1998.
- V. Griffith and C. Koch. Quantifying synergistic mutual information.
- D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-inspired artificial intelligence. Neuron, 95(2):245 – 258, 2017. ISSN 0896-6273.
- I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner. Early visual concept learning with unsupervised deep learning.
- I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework.
- M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In Workshop in Advances in Approximate Bayesian Inference, NIPS, 2016.
- H. Kim and A. Mnih. Disentangling by factorising. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, pages 2654–2663, 2018

Thank you!

Gonzalo Barrientos

Gonzalo.ayquipa.16@ucl.ac.uk