Disentangling Variational Autoencoders for Image Classification

Chris Varano
A9 - An Amazon Company

Motivation

Goal: Improve classification performance using unlabelled data

- There is a wealth of unlabelled data; labelled data is scarce
  - **Unsupervised learning** can learn a representation of the domain
- **Disentangled representations** contain statistically independent generative factors of the data
  - Improves representation quality and prevents latent co-adaptation
- **Supervised learning** benefits from knowledge of the underlying factors

Problem Statement

Unsupervised Learning:

- **Input:** Unlabelled image data
- **Output:** Reconstructed images
- **Evaluation:** Disentanglement

Supervised Learning:

- **Input:** Labelled image data
- **Output:** Classification label of image
- **Evaluation:** Test classification error

Dataset

MNIST Handwritten Digits [1]

- 55,000 training images
- 10,000 test images
- 10 classes

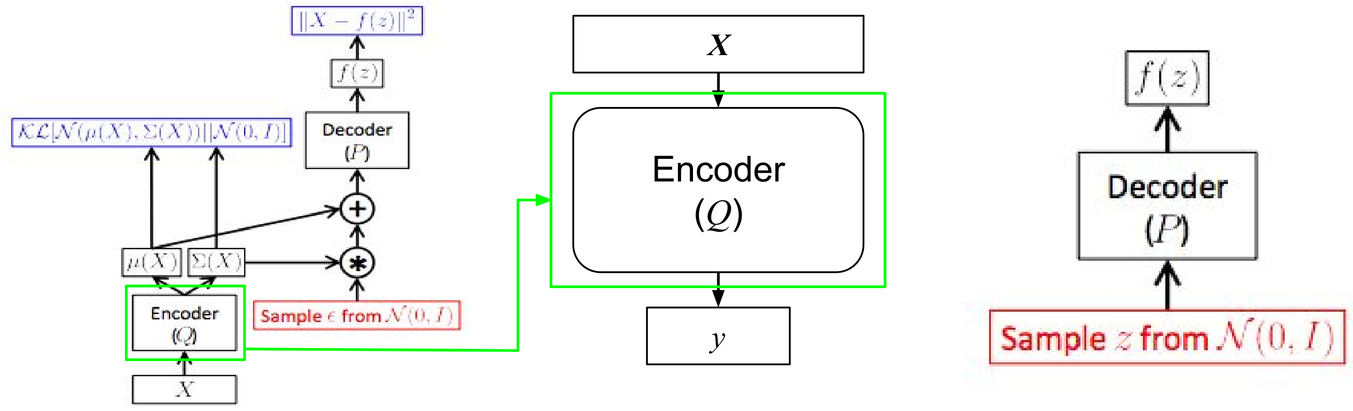
Conclusions and Future Work

- **Conclusion:** Disentangling VAE [3] improves classification performance over standard VAE and vanilla baseline when labelled data is scarce
- **Future work:** 1) Use synthetic MNIST with more continuous data (e.g. continuous rotations) so the DVAE can better learn the generative manifolds, and 2) use a semi-supervised learning objective on top of unsupervised pre-training.

References

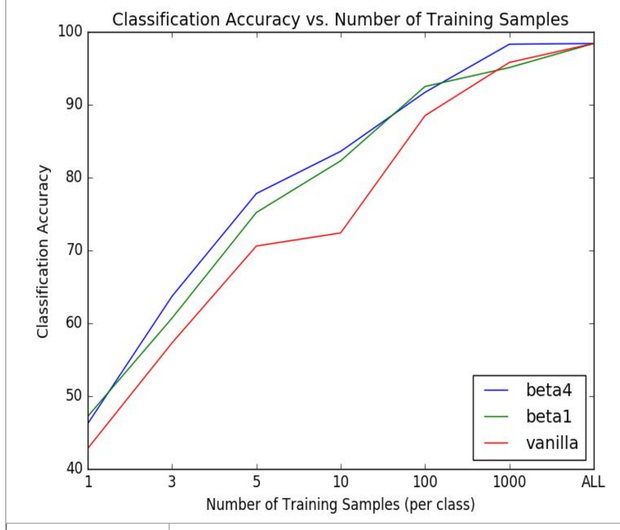
[1] Y. LeCun and C. Cortes. The mnist database of handwritten digits, 1998.
[2] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv: 1312.6114, 2013.*
[3] I. Higgins et al. Early visual concept learning with unsupervised deep learning, 2016.

Model



a) VAE architecture [2] b) Supervised architecture a) VAE sampler [2]

Quantitative Results



Images per class	Classifier		
	Vanilla	β=1 (VAE)	β=4 (DVAE)
3	57.7%	60.7%	**63.7%**
10	72.4%	82.3%	**83.6%**
ALL	98.4%	98.4%	**98.4%**

Qualitative Results

z10 - latent variable controlling rotation

z9 - latent variable controlling thickness

z7 - latent variable with no effect

CNN DVAE reconstructions of MNIST digits