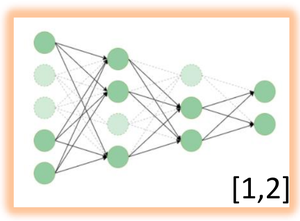
Impact of Disentanglement on Pruning Neural Networks

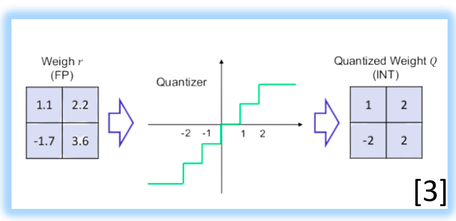
Carl Shneider*, Peyman Rostami, Anis Kacem, Nilotpal Sinha, Abdelrahman Shabayek, Djamila Aouada
Computer Vision, Imaging, and Machine Intelligence Research Group (CVI²)

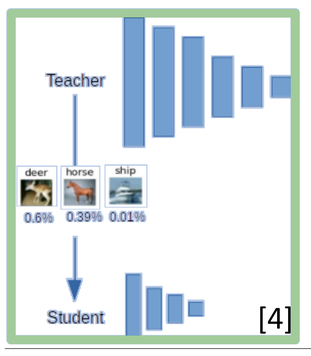
Summary: Efficient model compression techniques are required to deploy deep neural networks (DNNs) on edge devices for task specific objectives. A variational autoencoder (VAE) framework is combined with a pruning criterion to investigate the impact of having the network learn disentangled representations on the pruning process for the classification task.

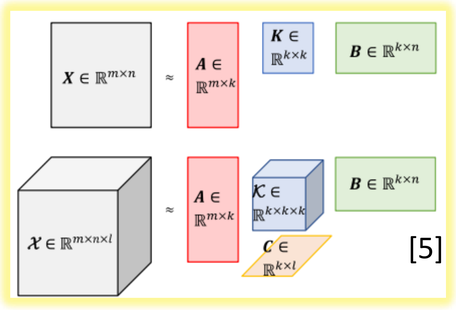
Neural Compression Methods



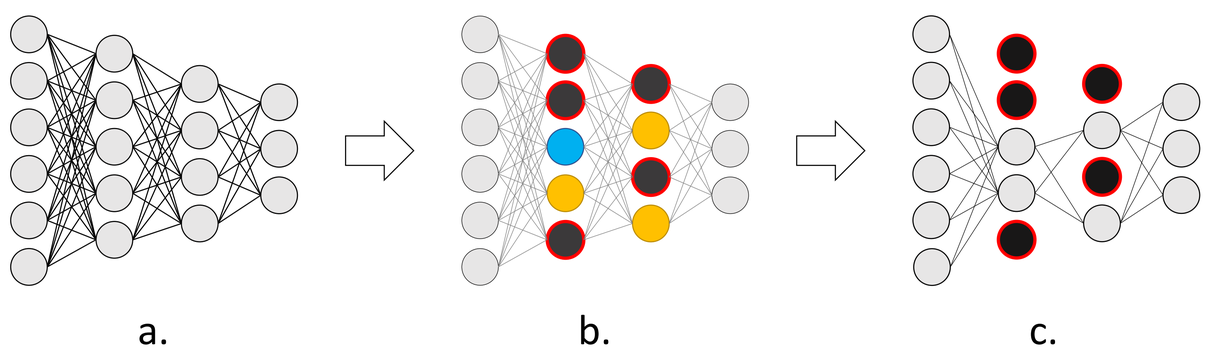
- Pruning
  - Structured [1]
  - Static/Dynamic [2]
- Weight quantization [3]
- Knowledge distillation [4]
- Low-rank factorization [5]







Disentanglement



a. Neural network contains mixed information content.
b. Ideal disentanglement of information content.
c. Network is pruned based on useful information.

- Object Shape
- Object Color
- Background

Disentanglement based Neural Compression

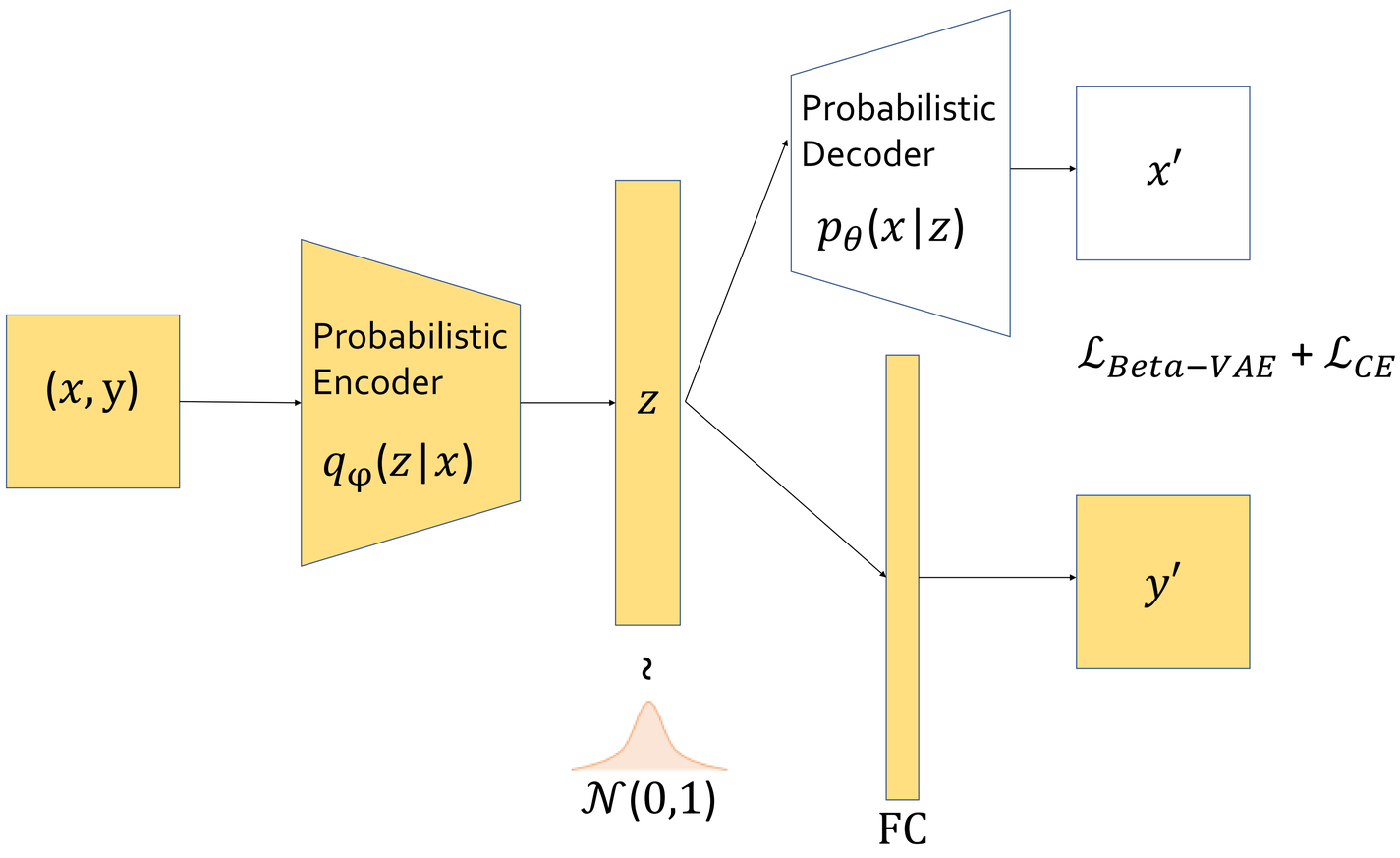


Fig. 1. The Beta-VAE model augmented by the addition of a classifier head. The combined model, Beta-VAE-Classif, is trained with all three loss terms given by the KL divergence, reconstruction loss, and classification loss. During inference, the reconstruction head *x'* is removed, leaving the shaded in blocks of the diagram.

$$\mathcal{L}_{\text{Beta-VAE}}(\theta, \phi; x, z, \beta) = -\beta\, \mathcal{D}_{KL}\left(q_\phi(z|x) \parallel p(z)\right) + \mathbb{E}_{q_\phi(Z|x)}[\log p_\theta(x|z)] \quad (1)$$

$$\mathcal{L}_{\text{CE}}(y) = -\sum_{i=1}^{N} y_i \log y_i' \quad (2)$$

$$\mathcal{L}_{\text{Beta-VAE-Classif}}(\theta, \phi; x, y, z, \beta) = \mathcal{L}_{\text{Beta-VAE}} + \mathcal{L}_{\text{CE}} \quad (3)$$

Eq. 1 denotes the Beta-VAE loss term composed of the KL divergence and reconstruction loss terms. Eq. 2 describes the cross-entropy loss of the classification head.

Table 1. Accuracy and compression for MNIST and CIFAR10 datasets. Pure CNN models are denoted as CNN-Classif. The augmented Beta-VAE models with the classifier head are denoted as Beta-VAE-Classif. The mean and standard deviation is computed over three independent runs where available. Compression is given in terms of kilobytes for MNIST and megabytes for CIFAR10.

MNIST			
Model	Beta	Acc. (%)	Comp. (KB)
CNN-Classif.	-	98.19 ± 0.2	260.28 ± 0.1
CNN-Classif. w/ pruning	-	97.29 ± 0.4	84.25 ± 0.3
Beta-VAE Classif.	1	96.32 ± 0.5	260.89 ± 0.0
Beta-VAE Classif. w/ pruning	1	90.02 ± 7.3	84.43 ± 0.2
Beta-VAE Classif.	3	95.34 ± 1.7	260.78 ± 0.6
Beta-VAE Classif. w/ pruning	3	92.52 ± 2.5	84.81 ± 0.4
Beta-VAE Classif.	5	96.86 ± 1.0	261.37 ± 0.0
Beta-VAE Classif. w/ pruning	5	96.15 ± 0.2	85.29 ± 0.2
Beta-VAE Classif.	10	95.67 ± 0.6	260.43 ± 0.1
Beta-VAE Classif. w/ pruning	10	95.86 ± 0.8	85.67 ± 0.4
Beta-VAE Classif.	50	62.32 ± 36.2	260.45 ± 0.0
Beta-VAE Classif. w/ pruning	50	59.55 ± 34.5	83.84 ± 0.1

CIFAR10			
Model	Beta	Acc. (%)	Comp. (MB)
CNN-Classif.	-	78.14 ± 0.0	10.9 ± 0.000
CNN-Classif. w/ pruning	-	78.12 ± 0.0	3.5 ± 0.000
Beta-VAE-Classif.	1	48.97 ± 1.4	10.83 ± 0.002
Beta-VAE-Classif. w/ pruning	1	48.76 ± 5.2	3.43 ± 0.005
Beta-VAE-Classif.	3	51.08 ± 4.5	10.82 ± 0.001
Beta-VAE-Classif. w/ pruning	3	49.30 ± 5.9	3.44 ± 0.009
Beta-VAE-Classif.	5	47.17 ± 9.4	10.82 ± 0.001
Beta-VAE-Classif. w/ pruning	5	51.28 ± 10.2	3.43 ± 0.009
Beta-VAE-Classif.	10	10.00 ± 0.0	10.80 ± 0.004
Beta-VAE-Classif. w/ pruning	10	10.00 ± 0.0	3.35 ± 0.001
Beta-VAE-Classif.	50	10.00 ± 0.0	10.80 ± 0.000
Beta-VAE-Classif. w/ pruning	50	10.00 ± 0.0	3.35 ± 0.000

Conclusion & Future Work

- Preliminary results inconclusive - hint that for a certain value of β, latent space implicitly becomes sufficiently disentangled to allow for pruning to more easily discard useless information for task of classification.
- Increasing number of latent dimensions may improve results.
- More robust results expected from having ground truth disentanglement labels together with an appropriately selected metric to directly measure degree of disentanglement for task of classification.

References
[1] L. Deng, G. Li, S. Han, L. Shi and Y. Xie, "Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey," in Proceedings of the IEEE, 2020.
[2] Li, Changlin, et al., "Dynamic slimmable network," CVPR, 2021.
[3] Gholami, Amir, et al., "A survey of quantization methods for efficient neural network inference," 2021.
[4] Gou, Jianping, et al., "Knowledge distillation: A survey." International Journal of Computer Vision, 2021.
[5] L. Deng et al., "Model Compression and Hardware Acceleration for Neural Networks: A Comprehensive Survey," in Proceedings of the IEEE, 2020.

CONTACT US

Acknowledgement
This work was funded by the Luxembourg National Research Fund (FNR) under the project reference C21/IS/15965298/ELITE.

Fonds National de la Recherche Luxembourg

https://cvi2.uni.lu/

*carl.shneider@uni.lu