LEARNING ROTATION-AGNOSTIC REPRESENTATIONS VIA GROUP EQUIVARIANT VAEs

Anonymous authors
Paper under double-blind review

ABSTRACT

An emerging field in representation learning involves the study of group-equivariant neural networks, that leverage concepts from group representation theory to design neural architectures that can exploit discrete and continuous symmetries to produce more general representations. Following this direction, in this work we demonstrate how an image embedding agnostic to rotations can be naturally obtained by training a variational autoencoder equipped with a Group equivariant Convolutional Neural Network (G-CNN) encoder[1].

1 INTRODUCTION AND RELATED WORK

Group equivariant Convolutional Neural Networks (G-CNNs) are a generalization of regular CNNs (LeCun & Bengio, 1995) that can exploit symmetries by sharing more weights. They have been shown to outperform regular CNNs on rotated and intrinsically symmetric data, increasing the expressive capacity of the network (Cohen & Welling, 2016a). Recent theoretical work has expanded the capabilities of G-CNNs, which can now deal with any compact continuous group transformation, e.g. continuous image rotations (Weiler & Cesa, 2021; Cesa et al., 2022).

Most of the existing work on G-CNNs has focused on classification tasks (Weiler & Cesa, 2021). In this work, instead, we will present some interesting results obtained by implementing a $SO(2)$ G-CNN inside a VAE (Kingma & Welling, 2013), in order to disentangle the image content from its rotation. Image representations agnostic to rotations are desirable in several scenarios, such as when working with astronomical data or histopathological images (Veeling et al., 2018). Recent works have explored ways to disentangle pose features, such as in Bepler et al. (2019), where the generative part of the autoencoder is an explicit function of the spatial coordinates of the image, enabling the model to factorize and discard the underlying rotation and translation. On the other hand, as we will see shortly, implementing directly a $SO(2)$ equivariant CNN as the encoder of a β-VAE is a simple alternative that produces representations agnostic to rotations by construction.

2 METHODS

Regular group convolutions are an extension of the classical CNNs, where images and features are modeled as functions $f : \mathbb{Z}^3(\mathbb{R}^m) \longrightarrow \mathbb{R}^n$. Here $\mathbb{Z}^3$ is the space of image pixels and $\mathbb{R}^p$ is the feature space. Instead, in a G-CNN the feature maps are functions in a group set G, therefore the notion of convolution can be generalized so that the shift becomes a more general transformation in G (Cohen & Welling, 2016a). This method works very well for discrete groups, but can not be applied to continuous groups, where we need to impose the *steerability* property of the feature field (Cohen & Welling, 2016b) (Weiler & Cesa, 2021), so we end up with a G-CNN encoder.

As a further regularization to our β-VAE model, we added a hyperspherical probability distribution as the approximate posterior ($\mathcal{S}$-VAEs). In fact, vanilla VAEs with low dimensional latent spaces suffer from *origin gravity* (they present a concentrated probability mass with all the different samples near the center) (Davidson et al., 2018). Hence, we use the von Mises-Fisher distribution, to obtain a uniform spread on the sphere.

[1]All the code needed to reproduce this work is available on Kaggle: `https://www.kaggle.com/<ANONYMIZED>`

3 HYPERSPHERIC GROUP EQUIVARIANT VARIATIONAL AUTOENCODER

The main purpose of using a group equivariant encoder is to exploit the group simmetry in order to make the $\mathcal{S}$-VAE invariant to rotation. A Hyperspherical Group equivariant Variational Autoencoder ($\mathcal{S}$-GVAE) is composed of an encoder and a decoder, like any other VAE. The only difference is in the encoder part, which is now composed by a G-CNN. The objective here is always to optimize a loss function:

$$\mathcal{L}(\phi, \psi) = \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{x};\theta)}[\log p_\phi(\mathbf{x}|\mathbf{z})] - \beta \cdot KL(q_\psi(\mathbf{z}|\mathbf{x};\theta)||p(\mathbf{z}))$$

Where $\mathcal{L}$ is the function to be maximized and the right end side is called Evidence Lower Bound (ELBO). This equation is directly related to $\log \int p_\phi(\mathbf{x}, \mathbf{z}) d\mathbf{z}$ which is the log-likelihood of the data. $q_\psi(\mathbf{z}|\mathbf{x};\theta)$ is the approximate posterior which should be optimized to give the right probability distribution of the data points and it is parametrized by a neural network.
In this experiment the approximate posterior is given by the von Mises-Fisher distribution and the probability distribution $p(\mathbf{z})$ of the latent space is the hypersferical uniform distribution, so that the KL divergence in the loss function is $KL(\text{vMF}(\mu, \kappa)||U(S^{n-1}))$ where μ is the mean value, κ is the concentration and S^{n-1} is the unit sphere of dimension $n-1$ where n is the latent dimension (Davidson et al., 2018). See appendix for further details about the architecture.

3.1 EXPERIMENTAL RESULTS

During training phase we used different combinations of β and latent dimensions. Although the best validation loss is obtained for β = 0.1 and latent dimension = 10, best performace in disentanglement and recontruction was observed for β = 0.1 and latent dimension = 6, see Table 1. This is resonable since a high latent dimension decreases the information loss at cost of generalization. Furthermore, a high β gives more weight to the KL term reducing reconstruction ability.

Table 1: Validation loss on various combinations of β and latent dimension after 50 epochs.

β:	0.01	0.1	1	150	500
dim. = 3	43.38k	42.56k	43.13k	43.38k	44.54k
dim. = 6	30.25k	30.94k	30.04k	32.96k	38.04k
dim. = 10	48.27k	**24.45k**	24.68k	28.5k	35.47k

After that, we retrained the latter model for 100 epochs obtaining the final validation loss of 30.29k. In Figure 1 we can see an image grid showing the disentanglement of random numbers.

Figure 1: Rotational disentanglement of random MNIST data using a $\mathcal{S}$-GVAE with β = 0.1 and latent dimension = 6.

4 CONCLUSION

In this work, we have presented $\mathcal{S}$-GVAE, a Variational Autoencoder with a hyperspherical latent space and a Group equivariant encoder. The main goal of this architecture is to achieve rotational disentanglement with minimal effort. In principle, other symmetries can be implemented by adding transformation elements to the group G. The tests performed are not definitive but provide a good basis for further experiments and improvements.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Tristan Bepler, Ellen D. Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. Explicitly disentangling image content from translation and rotation with spatial-vae. *arXiv:1909.11663v1*, 2019.

Gabriele Cesa, Leon Lang, and Maurice Weiler. A program to build e(n)-equivariant steerable cnns. *Conference paper at ICLR 2022*, 2022.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. *arXiv:1602.07576v3*, 2016a.

Taco S. Cohen and Max Welling. Steerable cnns. *arXiv:1612.08498v1*, 2016b.

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. *arXiv:1804.00891v2*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Yann LeCun and Yoshua Bengio. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 1995.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. *arXiv:1806.03962v1*, 2018.

Maurice Weiler and Gabriele Cesa. General e(2) - equivariant steerable cnns. *arXiv:1911.08251v2*, 2021.

A APPENDIX

As said before, the architecture is a classical VAE with the addition of a regularization element such as a hyperspherical latent space, and the replacement of the classical encoder with a group equivariant one.

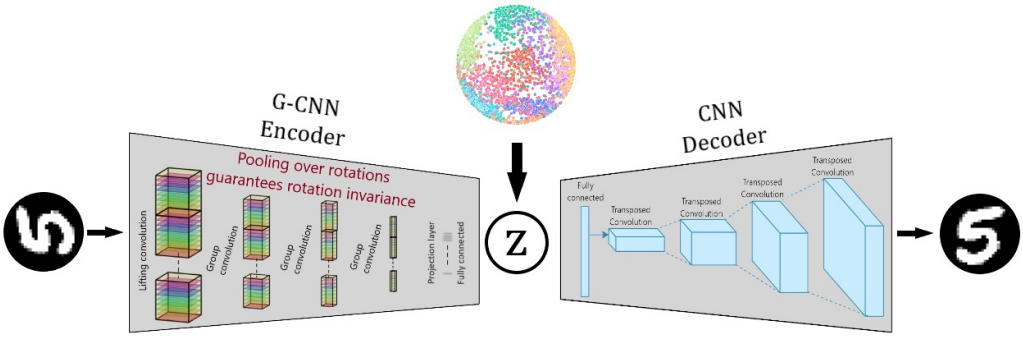

Figure 2: Sketch of the $\mathcal{S}$-GVAE: The architecture is composed by a group equivariant decoder in order to ensure rotation invariance and a hyperspherical latent space as a regularization element[3].

Now in details, the architecture used for the experiment:

Dataset:

MNIST subset of 35000 elements randomly rotated by $\theta \in [-180, 180]$ degree. The split between train and test set was stratified with test set = 1/7 of the entire dataset.

Optimizer:

Adam optimizer with initial learning rate of 1e-2, weight decay of 1e-5 and exponential scheduler with $\gamma = 0.9$.

Architecture:

Input: Batch (100 element) of 32x32x1 MNIST data (pad = 4).

Encoder: The encoder is made of an initial mask module, a FourierELU activation (frequency = 6), 3 G-steerable convolutional layers with 64, 64, 128 channels, stride = 2, 1, 2, kernel size = 4, FourierElu activation with frequency = 6, BatchNorm and Dropout = 0.1 each. A pointwise average pooling with stride = 1 was applied to the last layer. Then a G-convolution with kernel = 1 (to extract the invariant feature maps) and two last FC layers for the mean (μ) and concentration (κ) values of the von Mises-Fisher distribution with ELU activation, BatchNorm and Dropout = 0.2.

Latent: Latent dimension set to 3, 6, 10 in order to test different configurations.

Decoder: The decoder is made of a first FC linear layer with relu activation, BatchNorm and Dropout = 0.1, then 3 Transposed Convolutional layers with 128, 64, 32 channels, stride = 2, 1, 2, kernel size = 4, relu activation, BatchNorm and Dropout = 0.2 each.

[3]G-CNN encoder illustration adopted from the course "Deep Learning 2: Group equivariant deep learning" `https://uvagedl.github.io/`, University of Amsterdam.