Description of the project

Cél: Az arckép jellemzőinek a felismerése. Első lépésben címkézett adatok alapján az adatbázisban található címkékre készítsetek minél jobb modellt. További opcionális lépésként VAE (és/vagy GAN) alapú architektúrával célszerű kísérletezni, hogy a látens változók milyen arc jellemzőket tanulnak meg - akár felismerés, akár generálás idején.

Introduction

The features of images in a labelled dataset can be found by using discriminative models with convolutional neural networks, while in case of unlabelled image set several unsupervised techniques are proposed. Although with GANs ([1-]) have great advantages in generating new images and find implicit latents, we prefer working with Variational Autoencoders as (i) inference and the regularisation of the latent space can be directly controlled (ii) posterior inference (which might be intractable) can be made efficient by fitting an approximate inference (iii) β-VAE tends to consistently and robustly discover more latent factors and learn cleaner disentangled representations of them than either InfoGAN or DC-IGN [3.], (iv) furthermore, unlike InfoGAN and DC-IGN, β-VAE requires no design decisions or assumptions about the data, and is very stable to train.

Short summary of variational autoencoders

To observe the intrinsic structure of a dataset auto-encoders can be very efficient by finding a right pair of encoders and decoders keeping the maximum of information when encoding and the minimum reconstruction error when decoding. However, the lack of interpretable structures in the latent space (a.k.a lack of regularity of the latent space) and overfitting (because of the reconstruction component in the loss function) can result in severe inadequacy. I.e. neither : **continuity** (two close points in the latent space should not give two completely different contents once decoded) nor **completeness** (for a chosen distribution, a point sampled from the latent space should give "meaningful" content once decoded) can be guaranteed.

In variational autoencoder [1,2] we prevent the model to encode data far apart in the latent space and encourage as much as possible returned distributions to "overlap" with regularisation term, satisfying this way the expected continuity and completeness conditions.

Find the appropriate latents

The great novelty of Variational Autoencoders was that they combined probabilistic models, i.e. variational Bayesian approach with deep learning techniques so that to perform efficient inference and continuous latent variables even in case of intractable posterior distributions, and large datasets.

With a reparameterization of the variational lower bound, a differentiable unbiased estimator of the lower bound can be achieved: the SGVB (Stochastic Gradient Variational Bayes) estimator can be used for efficient approximate posterior inference.

However the interpretability of automatedly discovered factorized representation of the independent data generative factors needed further modification on VAEs, i.e.: the introduction of an adjustable hyperparameter β that balances latent channel capacity and independence constraints
with reconstruction accuracy. β-VAE with $\beta > 1$ outperforms VAEs when the parameter was appropriately tuned [3.]. This modification limits the capacity of latent variable, which, combined with the pressure to maximize the log likelihood of the training data, should encourage the model to learn the most efficient representation of the data as the Kubleck-Leibler divergence term of the

β-VAE objective function encourages conditional independence in the posterior: higher values of β should encourage learning a disentangled representation.

Quantitatively comparing the different unsupervised deep generative models a crucial point is the degree of disentanglement of the latent variables. In case of a disentangled representation one single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors [4.] and knowledge about one factor can generalize to novel configurations of other factors. [4.] also provides a protocol to quantitatively compare the degree of disentanglement learnt by different models.

Log likelihood of the data under the learnt model is a poor metric for evaluating disentangling in β-VAEs as latent channel capacity restriction (β > 1) can lead to poorer reconstructions due to the loss of high frequency details when passing through a constrained latent bottleneck. [3.] propose a quantitative
metric that directly measures the degree of learnt disentanglement in the latent representation. It is important to note that a representation consisting of independent latents is not necessarily disentangled: i.e. PCA or ICA do not in general align with the data generative factors and hence may lack interpretability thus a simple cross-correlation calculation between the inferred latents would not suffice as a disentanglement metric.

Instead: inference is run on a number of images that are generated by fixing the value of one data generative factor while randomly sampling all others.

A low capacity linear classifier is used to identify the fixed factor and report the accuracy value as the final disentanglement metric score (as the independence and interpretability properties hold for the inferred representations, thus there will be less variance in the inferred latents that correspond to the fixed generative factor. Smaller variance in the latents corresponding to the target factor will make the job of this classifier easier, resulting in a higher score under the metric.

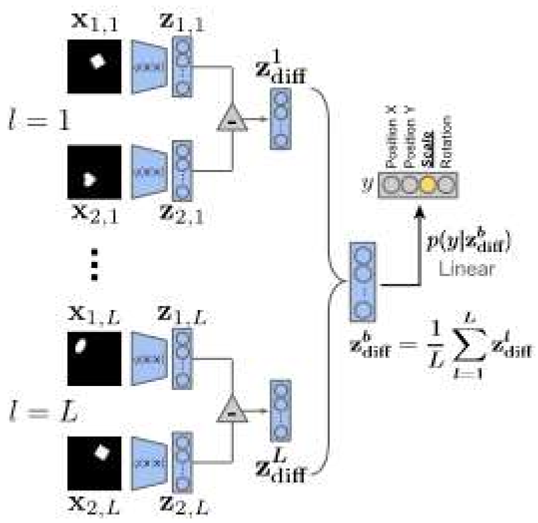

Figure 5: Schematic of the proposed disentanglement metric: over a batch of L samples, each pair of images has a fixed value for one target generative factor y (here $y = scale$) and differs on all others. A linear classifier is then trained to identify the target factor using the average pairwise difference z^b_{diff} in the latent space over L samples.

Tuning the β coefficient

β can be considered as a mixing coefficient for balancing the magnitudes
of gradients from the reconstruction and the prior-matching components of the VAE lower bound formulation. β should be normalized by latent z size m and input x size n in order to compare its

different values across different latent layer sizes and different datasets (βnorm=β*M/N). We found that larger latent z layer sizes m requires higher constraint pressures (higher β values). Furthermore, the relationship of β for a given m is characterised by an inverted U curve. When β is too low or too high the model learns an entangled latent representation due to either too much or too little capacity in the latent z bottleneck: β > 1 is necessary to achieve good disentanglement. However, if β is too high and the resulting capacity of the latent channel is lower than the number of data generative factors, then the learnt representation necessarily has to be entangled (as a low-rank projection of the true data generative factors will compress them in a non-factorial way to still capture the full data distribution well).

We proposed a principled way of choosing β for datasets with at least weak label information. If label information exists for at least a small subset of the independent data generative factors of variation, one can apply the disentanglement metric described in Sec. 3 to approximate the level of learnt disentanglement for various β choices during a hyperparameter sweep.

Hierarchical VAEs

A more sophisticated approach for image generation is the nouveau VAE (NVAE) [5], a deep hierarchical VAE using depth-wise separable convolutions and batch normalization. NVAE is equipped with a residual parameterization of normal distributions and its training is stabilized by spectral regularization. NVAE uses depthwise convolutions in its generative model with (i) a new residual parameterization of the approximate posteriors. ii) stabilized training deep VAEs with spectral regularization, iii) practical solutions are used to reduce the memory burden of VAEs. iv) deep hierarchical VAEs can obtain state-of-the-art results on several image datasets and can produce high-quality samples even when trained with the original VAE objective.
The main building block of NVAE is depthwise convolutions that rapidly increase the receptive field of the network without dramatically increasing the number of parameters. (In depth-wise convolution, we use each filter channel only at one input channel.) NVAE was the first successful application of VAEs to images as large as 256×256 pixels.

[1.] Diederik P Kingma and Max Welling.
Auto-encoding variational bayes.
In The International Conference on Learning Representations (ICLR), 2014.
https://arxiv.org/pdf/1312.6114.pdf

[2.] Diederik P Kingma and Max Welling.
An introduction to variational autoencoders
Foundations and Trends® in Machine Learning, 2019.
https://arxiv.org/pdf/1906.02691.pdf

[3.] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, Alexander Lerchner
ICLR 2017 conference submission, 2017.

Beta-vae: Learning basic visual concepts with a constrained variational framework
https://openreview.net/forum?id=Sy2fzU9gl

[4.] Y. Bengio, A. Courville, and P. Vincent.
Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013

[5.] Arash Vahdat, Jan Kautz
NVAE: A Deep Hierarchical Variational Autoencoder
https://arxiv.org/pdf/2007.03898.pdf

[6.] Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen Progressive Growing of GANs for Improved Quality, Stability, and Variation
https://arxiv.org/pdf/1710.10196v3.pdf