Deep Learning

Lecture 8: Variational and implicit models

Chris G. Willcocks

Durham University

Lecture overview

1 Divergence measures

- Kullback–Leibler divergence
- cross entropy
- optimal transport

2 Variational models

- problems with autoencoders
- variational autoencoders
- ELBO

3 Implicit networks

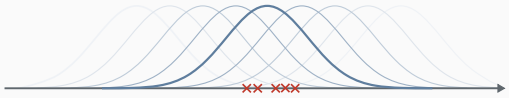
- implicit representations
- SIREN
- NeRF
- GONs

Definition: maximum likelihood estimation

Maximum likelihood estimation (MLE) is a method for estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the model the observed data is most probable

$$\begin{aligned}\theta^* &= \arg\max_\theta p_{\text{model}}(\mathbb{X};\theta) \\ &= \arg\max_\theta \prod_{i=1}^{n} p_{\text{model}}(\mathbf{x}^i;\theta) \\ &\approx \arg\max_\theta \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log p_{\text{model}}(\mathbf{x};\theta)],\end{aligned}$$

where $\mathbb{X} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^n\}$ are from $p_{\text{data}}(\mathbf{x})$

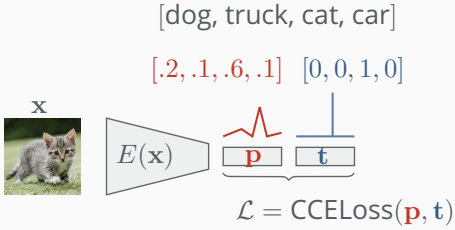


Definition: encoder

Encoders typically reduce the spatial dimensions (compress) but may increase the feature dimensions.

Link to Colab example

- For regression, we generally use an L_2 loss
- For classification, we generally use categorical cross entropy - why?
- For binary classification, you can use binary cross entropy - why?



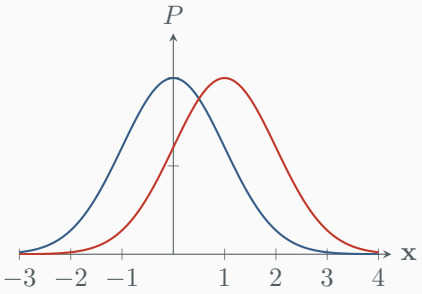
Definition: Kullback–Leibler divergence

We want to measure how different two distributions are. The Kullback-Leibler divergence (also called **relative entropy**) is one such measure that is asymmetric and non-negative:

$$D_{\mathrm{KL}}(p \,||\, q) = \int p(\mathbf{x}) \log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \mathrm{d}\mathbf{x}$$

where $D_{\mathrm{KL}}(p \,||\, p) = 0$. Practically, the KL divergence is sensitive at the tails of the distribution.

Click to try on Desmos



Definition: Cross entropy

The KL divergence can be rewritten

$$\begin{aligned} D_{\mathrm{KL}}(p||q) &= \int p(\mathbf{x}) \log\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \mathrm{d}\mathbf{x} \\ &= \int p(\mathbf{x}) \log p(\mathbf{x}) \,\mathrm{d}\mathbf{x} - \int p(\mathbf{x}) \log q(\mathbf{x}) \,\mathrm{d}\mathbf{x} \\ &= -H(p) + H(p, q) \end{aligned}$$

where $H(p, q)$ is the cross entropy and $H(p) = H(p, p)$ is the regular entropy. This is minimizing the negative log-likelihood, the same as maximizing the likelihood

$$H(p, q) = -\int p(\mathbf{x}) \log q(\mathbf{x})$$

Definition: Wasserstein metric

It's worth mentioning an elegant metric between two distributions, which is from the optimal transport problem formalised by Gaspard Monge in 1781. If interested, watch Cédric Villani on YouTube discuss this topic and see examples in [1].

In 1D, the first Wasserstein **metric** (or earth mover's distance) can be written between the two CDF's F_1 and F_2 where

$$W(F_1, F_2) = \int |F_1(x) - F_2(x)| \,\mathrm{d}x$$

This is the minimal (optimal) cost from moving dirt between the two distributions whose CDF's are F_1 and F_2 respectively.

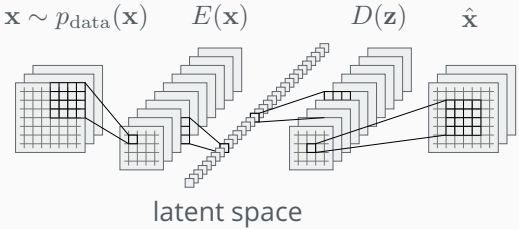
Recap: autoencoder

An autoencoder is a feedforward neural network that is trained to predict its inputs, thus learning an identity

$$\mathcal{L}_{\text{AE}} = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}}[\,\mathcal{L}(\mathbf{x}, D(E(\mathbf{x})))\,]$$

where $\mathcal{L}$ is a loss function such as mean squared error. The encoder function $E : \mathbb{R}^n \to \mathbb{R}^m$ (usually) compresses the dimensionality of the data $n \ll m$ to a latent encoding $\mathbf{z} = E(\mathbf{x})$, which is then recovered by the decoder $D : \mathbb{R}^m \to \mathbb{R}^n$, where $\hat{\mathbf{x}} = D(\mathbf{z})$. It's difficult to sample from the latent space, so its not really a generative model.

Link to Colab example



Definition: variational autoencoders

Variational autoencoders are generative models, as they impose a prior over the latent space $p(\mathbf{z})$, typically $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ which can be sampled from.

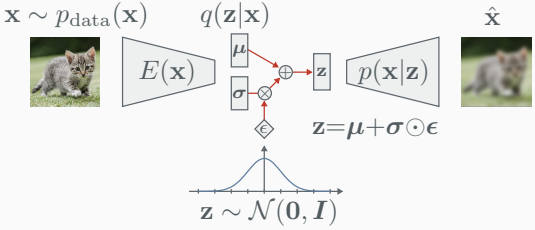
$$\mathbf{z} \sim E(\mathbf{x}) = q(\mathbf{z}|\mathbf{x}), \quad \hat{\mathbf{x}} \sim D(\mathbf{z}) = p(\mathbf{x}|\mathbf{z})$$

The VAE loss is the negated expected log-likelihood (the reconstruction error) and the prior regularization term:

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})}\right] = \mathcal{L}^{\text{pixel}}_{\text{recon}} + \mathcal{L}_{\text{prior}}$$

where

$$\mathcal{L}^{\text{pixel}}_{\text{recon}} = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})]$$
$$\mathcal{L}_{\text{prior}} = D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z}))$$



Definition: ELBO

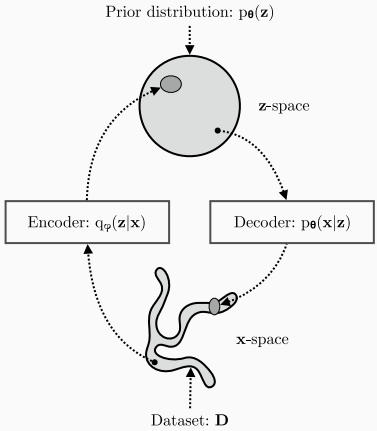
VAEs therefore have three components:

1. the decoder $p_\theta(\mathbf{x}|\mathbf{z})$
2. the *approximate posterior* (encoder) $q_\phi(\mathbf{z}|\mathbf{x})$
3. the prior distribution $p_\theta(\mathbf{z})$

They are trained with the reparameterisation trick to maximise the evidence lower bound (ELBO):

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) - D_{\mathrm{KL}}\left[q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})\right]$$

Read [2] for detail on the theory (where the figure is from) and [3] for a state-of-the-art method that stacks VAEs hierarchically (Very Deep VAEs).



Implicit networks definition

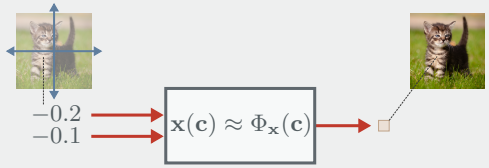
Definition: implicit representations

Consider data $\Phi\colon \mathbb{R}^m \to \mathbb{R}^n$, like a single image, as a function of coordinates $\mathbf{c} \in \mathbb{R}^m$. The aim is to learn a neural approximation of Φ that satisfies an implicit equation:

$$R(\mathbf{c}, \Phi, \nabla_\Phi, \nabla^2_\Phi, \dots) = 0, \quad \Phi\colon \mathbf{c} \mapsto \Phi(\mathbf{c}).$$

Equations with this structure arise in a myriad of fields, namely 3D modelling, image, video, and audio representation.

Example: implicit network



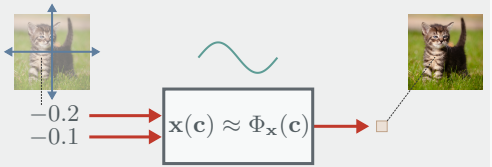
Implicit representation networks SIREN

Definition: SIREN

SInusoidal REpresentation Networks (SIREN) are a simple implicit representation network with fully connected layers, but use `sin` (with clever initialisation to scale it appropriately) as their choice of non-linearity [4].

`sin` is periodic, so it allows to capture patterns over all of the coordinate space (it's translation invariant, like convolutions).

Example: SIREN (implicit network)

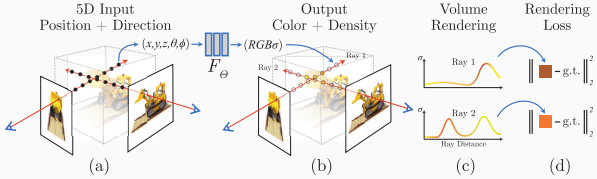


Link to project page

Definition: NeRF

Neural Radiance Fields (NeRF) are similar to SIRENs, but instead of representing an image, they represent a single 3D scene [5].

They map from pixel positions (x, y, z) and a viewing direction (θ, ϕ) to a colour and density value σ integrated via a ray on F_θ.



Link to project page

Definition: gradient origin networks

Gradient origin networks (GON) treat the derivative of the decoder as an encoder [6]. This allows us to compute the latents:

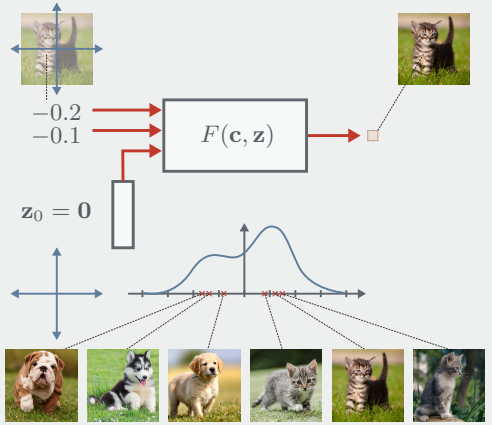
$$\mathbf{z} = -\nabla_{\mathbf{z}_0} \mathcal{L}(\mathbf{x}, F(\mathbf{z}_0))$$

which are then jointly optimised, giving the GON objective:

$$G_{\mathbf{x}} = \mathcal{L}(\mathbf{x}, F(-\nabla_{\mathbf{z}_0} \mathcal{L}(\mathbf{x}, F(\mathbf{z}_0)))).$$

Link to project page

Example: implicit GON



Take Away Points

Summary

In summary:

- Maximum likelihood, KL, classification (categorical cross entropy), and generative models are connected in the underpinning theory
- Most deep generative models approximate $\log p(\mathbf{x})$ (EBM, GAN, VAE)
- VAEs just maximise the evidence lower bound (ELBO)
- Flows are better 'in theory' as give exact likelihoods
- Some types of deep neural networks aren't technically generative models (just learning to represent one example)...
    - ...but they probably should be!
- Implicit networks can operate on continuous or irregular signals and can be sampled at any resolution

References I

[1] Gabriel Peyré, Marco Cuturi, et al. “Computational Optimal Transport: With Applications to Data Science”. In: Foundations and Trends® in Machine Learning 11.5-6 (2019), pp. 355–607.

[2] Diederik P Kingma and Max Welling. “An introduction to variational autoencoders”. In: arXiv preprint arXiv:1906.02691 (2019).

[3] Rewon Child. “Very Deep {VAE}s Generalize Autoregressive Models and Can Outperform Them on Images”. In: International Conference on Learning Representations. 2021. URL: https://openreview.net/forum?id=RLRXCV6DbEJ.

[4] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. “Implicit neural representations with periodic activation functions”. In: Advances in Neural Information Processing Systems 33 (2020).

References II

[5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: European conference on computer vision. Springer. 2020, pp. 405–421.

[6] Sam Bond-Taylor and Chris G. Willcocks. “Gradient Origin Networks”. In: International Conference on Learning Representations. 2021. URL: https://dro.dur.ac.uk/34356/1/34356.pdf.