

Denoising Autoencoders

John Thickstun

The idea of a denoising autoencoder [Vincent et al., 2010] is to recover a data point $x \sim p$ given a noisy observation, for example $\tilde{x} = x + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. These models were initially introduced to provide an objective for unsupervised pre-training of deep networks. While that training methodology has become less relevant over time, the denoising autoencoder has been adapted as a means of constructing generative models [Bengio et al., 2013] leading to recent spectacular results [Ho et al., 2020].

Denoising AEs

Formally, let $x \sim p$, $\tilde{x} \sim p(\cdot|x)$, which together define a joint distribution $p(x, \tilde{x}) = p(\tilde{x}|x)p(x)$. A denoising autoencoder is a model of the posterior distribution

$$p(x|\tilde{x}) = \frac{p(\tilde{x}|x)p(x)}{\int_{\mathcal{X}} p(\tilde{x}|y)p(y) dy}. \quad (1)$$

Specifically, we model $p_{\theta}(x|\tilde{x}) = \mathcal{N}(g_{\theta}(f_{\varphi}(\tilde{x})), \sigma^2 I)$, where $f_{\varphi} : \mathcal{X} \rightarrow \mathcal{Z}$ and $g_{\theta} : \mathcal{Z} \rightarrow \mathcal{X}$ are neural networks with parameters φ and θ respectively. Fitting this model using the MLE leads to the reconstruction objective

$$\theta^*, \varphi^* = \arg \min_{\theta, \varphi} \mathbb{E}_{(x, \tilde{x}) \sim p} \|x - g_{\theta}(f_{\varphi}(\tilde{x}))\|^2. \quad (2)$$

We can think of the composition $g_{\theta} \circ f_{\varphi}(\tilde{x})$ as projecting a corrupted data point $\tilde{x} \sim p(\cdot|x)$ back onto the support of p (the data manifold) and we can think of the latent space \mathcal{Z} as a coordinate system on the data manifold. Previewing the work of Bengio et al. [2013] and [Ho et al., 2020], we can imagine constructing a Markov chain of denoising autoencoders that guides us from samples $x_0 = \varepsilon \sim \mathcal{N}(0, I)$ through a sequence of denoising operations $x_s \sim p_{s, \theta}(\cdot|x_{s-1})$ to produce a final sample x_t distributed approximately according to p .

Denoising Score Matching

We can use a denoising autoencoder to construct an explicit score matching estimator, following Vincent [2011]. Recall that the score function estimator given by minimization of the Fisher divergence

$$\hat{\theta} = \arg \min_{\theta} \mathbb{E}_{x \sim p} \left[\frac{1}{2} \|s_{\theta}(x) - \nabla_x \log p(x)\|_2^2 \right]. \quad (3)$$

This expression requires that we evaluate gradients of the unknown density $p(x)$, which are not accessible to us. Previously, we saw an implicit score matching estimator [Hyvärinen, 2005] for estimating this quantity using samples. Denoising AEs offer another alternative.

Suppose we are willing to settle for samples of an estimate of the density $p(x)$ given by the noisy distribution

$$q(\tilde{x}) = \int_{\mathcal{X}} p(\tilde{x}|x)p(x) dx. \quad (4)$$

For example, if $p(\tilde{x}|x) = p_{\sigma}(\tilde{x}|x) = \mathcal{N}(\tilde{x}; x, \sigma^2 I)$ for small variance σ^2 , then $q(\tilde{x}) = q_{\sigma}(\tilde{x})$ is a Gaussian convolution corresponding to a mildly-smoothed version of $p(x)$, with $D(q_{\sigma} \parallel p) \rightarrow 0$ as $\sigma^2 \rightarrow 0$. Suppose we want to calculate the score matching estimator for this noisy distribution q_{σ} . The following proposition shows that this is much easier than score matching with p .

Proposition 1. (*Denoising Score Matching*) [Vincent, 2011]

$$\arg \min_{\theta} \mathbb{E}_{\tilde{x} \sim q_{\sigma}} \left[\frac{1}{2} \|s_{\theta}(\tilde{x}) - \nabla_{\tilde{x}} \log q_{\sigma}(\tilde{x})\|_2^2 \right] = \arg \min_{\theta} \mathbb{E}_{\substack{x \sim p \\ \tilde{x} \sim p_{\sigma}(\cdot|x)}} \left[\frac{1}{2} \|s_{\theta}(\tilde{x}) - \nabla_{\tilde{x}} \log p_{\sigma}(\tilde{x}|x)\|_2^2 \right]. \quad (5)$$

Proof. Expanding the quadratic and dropping the constant term, we have

$$\arg \min_{\theta} \mathbb{E}_{\tilde{x} \sim q_{\sigma}} \left[\frac{1}{2} \|s_{\theta}(\tilde{x}) - \nabla_{\tilde{x}} \log q_{\sigma}(\tilde{x})\|_2^2 \right] = \arg \min_{\theta} \mathbb{E}_{\tilde{x} \sim q_{\sigma}} \left[\frac{1}{2} \|s_{\theta}(\tilde{x})\|^2 - s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \log q_{\sigma}(\tilde{x}) \right]. \quad (6)$$

$$\mathbb{E}_{\tilde{x} \sim q_{\sigma}} [s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \log q_{\sigma}(\tilde{x})] = \int_{\mathcal{X}} s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \log q_{\sigma}(\tilde{x}) q_{\sigma}(\tilde{x}) d\tilde{x} \quad (7)$$

$$= \int_{\mathcal{X}} s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} q_{\sigma}(\tilde{x}) d\tilde{x} \quad (8)$$

$$= \int_{\mathcal{X}} s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \int_{\mathcal{X}} p(x) p_{\sigma}(\tilde{x}|x) dx d\tilde{x} \quad (9)$$

$$= \int_{\mathcal{X}} s_{\theta}(\tilde{x})^T \int_{\mathcal{X}} \nabla_{\tilde{x}} p(x) p_{\sigma}(\tilde{x}|x) dx d\tilde{x} \quad (10)$$

$$= \int_{\mathcal{X}} s_{\theta}(\tilde{x})^T \int_{\mathcal{X}} p(x) p_{\sigma}(\tilde{x}|x) \nabla_{\tilde{x}} \log p_{\sigma}(\tilde{x}|x) dx d\tilde{x} \quad (11)$$

$$= \int \int_{\mathcal{X} \times \mathcal{X}} p(x) p_{\sigma}(\tilde{x}|x) s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \log p_{\sigma}(\tilde{x}|x) d(x, \tilde{x}) \quad (12)$$

$$= \mathbb{E}_{\substack{x \sim p \\ \tilde{x} \sim p_{\sigma}(\tilde{x}|x)}} [s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \log p_{\sigma}(\tilde{x}|x)]. \quad (13)$$

From this and Eq 6, ==>

$$\arg \min_{\theta} \mathbb{E}_{\tilde{x} \sim q_{\sigma}} \left[\frac{1}{2} \|s_{\theta}(\tilde{x})\|^2 - s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \log q_{\sigma}(\tilde{x}) \right] \quad (14)$$

$$= \arg \min_{\theta} \mathbb{E}_{\substack{x \sim p \\ \tilde{x} \sim p_{\sigma}(\tilde{x}|x)}} \left[\frac{1}{2} \|s_{\theta}(\tilde{x})\|^2 - s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \log p_{\sigma}(\tilde{x}|x) \right] \quad (15)$$

$$= \arg \min_{\theta} \mathbb{E}_{\substack{x \sim p \\ \tilde{x} \sim p_{\sigma}(\tilde{x}|x)}} \left[\frac{1}{2} \|s_{\theta}(\tilde{x})\|^2 - s_{\theta}(\tilde{x})^T \nabla_{\tilde{x}} \log p_{\sigma}(\tilde{x}|x) + \frac{1}{2} \|\nabla_{\tilde{x}} \log p_{\sigma}(\tilde{x}|x)\|^2 \right] \quad (16)$$

$$= \arg \min_{\theta} \mathbb{E}_{\substack{x \sim p \\ \tilde{x} \sim p_{\sigma}(\cdot|x)}} \left[\frac{1}{2} \|s_{\theta}(\tilde{x}) - \nabla_{\tilde{x}} \log p_{\sigma}(\tilde{x}|x)\|_2^2 \right] \quad (17)$$

□

References

- Yoshua Bengio, Li Yao, Guillaume Alain, and Pascal Vincent. Generalized denoising auto-encoders as generative models. In *Advances in neural information processing systems*, pages 899–907, 2013. [\(document\)](#)
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. [\(document\)](#)
- Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 2005. [\(document\)](#)
- Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 2011. [\(document\)](#), 1
- Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 2010. [\(document\)](#)