

Variational Autoencoders: From Compression to Generation

Lecture Notes Consolidated with Claude AI — CS6140 Machine Learning

Instructor’s lecture plan for the note of the same title (VAE_claude_0929.pdf). Section numbers (§) refer to that note.

This note supports the first Week 10 lecture and **HW5 Problem 3**. It assumes only Week 9 (a neuron is a logistic unit; layers have a forward and a backward pass; softmax/cross-entropy; SGD/Adam), HW3’s PCA, and HW4’s Gaussian mixture fit by EM. Nothing here needs more probability than Gaussians, Bayes’ rule and expectations. Companion materials: the Week 9 notes (**NN_BASICS**, **NN_ADVANCED**), the **GAN** note (second Week 10 lecture), and the demo notebook `vae_lecture_demo.ipynb` (every number in this note is checked there).

Lecture plan (100 minutes, mostly on the board)

Time	On the board	Worked example / demo
0–10	Three things we already have: the 8-3-8 autoencoder (HW5 P1), PCA = a linear autoencoder (HW3), a latent-variable generative model (HW4’s GMM). What an autoencoder cannot do: <i>generate</i> . Then the big picture: the VAE’s network in one drawing; leave it up all lecture. (§1, §2)	Draw the 8-3-8 net; draw “codes with gaps”; then the VAE network (§2).
10–25	A continuous latent variable: $z \sim \mathcal{N}(0, 1)$, $x = g(z) + \text{noise}$. With a <i>linear</i> g everything is computable (probabilistic PCA); with a neural g nothing is. (§3)	$\mathbf{w} = (2, 1)$, $s^2 = 1$, $\mathbf{x} = (3, 1)$: posterior mean 7/6, variance 1/6.
25–40	KL divergence from scratch; cross-entropy = entropy + KL (so MLE = min KL); KL between Gaussians in three lines. (§4)	$p = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ vs. uniform: 0.0589 / 0.0566. Table of Gaussian KLs.
40–55	The ELBO: $\log p(x) = \text{ELBO}(q) + \text{KL}(q \parallel \text{posterior})$ for <i>any</i> q ; EM is “set $q = \text{posterior}$ ”. (§5)	Two-state latent with numbers; three choices of q in a table.
55–70	From the ELBO to a network: encoder with two heads, decoder = Week 9 MLP, loss = BCE (pixel-wise logistic regression) + KL. (§6)	Draw the architecture; write the loss for one image.
70–85	Training through a random node: the reparameterization trick; computation graph. (§7)	$\mathbb{E}[z^2]$, $\mu = 1$, $\sigma = \frac{1}{2}$, $\epsilon = \pm 1$: gradients 2 and 1 exactly.
85–95	Demo (5 min): AE vs. VAE latent spaces; discuss. (§8)	Notebook §7 (8×8 digits).
95–100	EM vs. VAE table; HW5 Problem 3 map. (§9, §10)	

Board layout suggestion. Keep a narrow column on the left for the whole lecture with the three “things we already have” (AE, PCA, GMM) and add one line to each whenever the lecture connects back to it. The rest of the board is for the running example.