Effective Estimation of Deep Generative Language Models

Tom Pelsmaeker and Wilker Aziz

ILLC, University of Amsterdam

Deep Latent Variable Models

Latent variable models with neural likelihood (or *sampling distribution*)

$$p(x|\theta) = \int p(z)p(x|z,\theta)\, dz$$

Motivation:

- Statistical (expressiveness): marginalisation breaks independence assumptions
- Practical (usefulness): generative model may exploit neighbourhood in latent space to explain structural similarity in data space.

(Kingma and Welling, 2014; Rezende et al., 2014)

Text Generation with Strong Generators

In text generation tasks, models that make **no independence assumptions**

- e.g., those parameterised by RNNs or Transformers

can model the data *arbitrarily well* without the need for marginalisation.

Example: auto-regressive product of Categorical distributions

$$X_i \mid \theta, x_{<i} \sim \text{Cat}(\text{NN}(x_{<i}; \theta))$$

We call these ***strong generators***.

(Bowman et al., 2016)

Posterior Collapse

Strong generator can model X_i independently of Z given $X_{<i}$

$$p(x_i|z,x_{<i},\theta) = p(x_i|x_{<i},\theta)$$

This means X is **independent** of Z in the joint distribution

$$p(x,z|\theta) = p(z)p(x|\theta)$$

Thus the **true posterior** is **independent of the data**

$$p(z|x,\theta) = p(x,z|\theta) / p(x|\theta) = p(z)$$

Visualise Data Space via a Walk in Latent Space

A collapsed models is ***not*** a poor generator, but its latent space ***is*** useless.

The two sides hadn't met since Oct. 18.
I don't know how much money will be involved.
The specific reason for gold is too painful.
The New Jersey Stock Exchange Composite Index gained 1 to 16.
And some of these concerns aren't known.
Prices of high-yield corporate securities ended unchanged.

Collapsed model

The inquiry soon focused on the judge.
The judge declined to comment on the floor.
The judge was dismissed as part of the settlement.
The judge was sentenced to death in prison.
The announcement was filed against the SEC.
The offer was misstated in late September.
The offer was filed against bankruptcy court in New York.
The letter was dated Oct. 6.

Non-collapsed model

Collapsed Variational Auto-Encoders (VAEs)

VAEs maximise the *evidence lowerbound* (ELBO) and thus minimise

$$KL(q(z|x,\lambda) \,||\, p(z|x,\theta))$$

where $p(z|x,\theta) = p(z)$ this leads to Z being independent of X in q, i.e.

$$q(z|x,\lambda) = p(z)$$

(Chen et al., 2017; Alemi et al., 2018)

How do we criticise VAEs?

Quantitatively

- Importance-sampling estimates of held-out log-likelihood
- Distortion: a notion of *reconstruction error*
  - Average held-out $E_q[- \log p(x|z)]$ gives us an estimate
- Rate: the *maximum mutual information* between X and Z possible
  - Average held-out $KL(q(z|x,\lambda) \,||\, p(z))$ gives us an estimate

Qualitatively: data generated from a

- prior sample shows the decoder is well trained
- posterior sample shows the posterior is not independent of the data.

Contributions

1. We review a number of strategies and *test* them in language modelling
    a. Word dropout
    b. Annealing
    c. β-VAE
    d. LaggingVAE
    e. Free-bits (FB) and Soft-FB (SFB)
    f. InfoVAE
    g. LagrangeVAE
2. We also make technical contributions to promote higher rates
    a. Directly by targeting a specific positive rate via constrained optimisation (**MDR** in the paper)
    b. Indirectly by creating a mismatch between the prior and the approximate posterior families (**strong priors** in the paper)

Experimental Setup

- English corpora: Penn Treebank, Yahoo, Yelp
- Bayesian optimisation: systematic hyperparameter optimisation
- All likelihoods are parameterised by GRUs: z is used to initialise the GRU
- All posterior approximations are diagonal Gaussian
- Model criticism
  - Intrinsic indicators of quality
  - Diagnostics based on samples, greedy samples, homotopies, and retrieval

Summary of Findings

Most techniques work

- except word dropout and annealing

But getting them to work is not equally easy

- e.g., LangrangeVAE works quite well, but its many hyperparameters are all very important and easy guesses are not obvious.

They all aim at higher rates, but FB and MDR do so by specifying a target rate directly

- the rate is a single number with a clear interpretation
- it does not seem to require a very fine-grained range of values

Final recommendations

- Target a particular rate when training your VAEs
- Monitor degree of determinism and copy in generations
- Always use importance sampling to estimate log-likelihood (and perplexity)

  This is crucial point has received very little attention! For a thorough discussion, check On Importance Sampling-Based Evaluation of Latent Language Models (Logan IV et al., at this conference)

Future work (or *free research idea*): a sandwich of mutual information

- Rate (an upperbound) is used in MDR. Distortion (relates to a lowerbound) is used in LagrangeVAE.

Thank you for Watching!

Code: https://github.com/tom-pelsmaeker/deep-generative-lm