Generating Sentences from a Continuous Space

1 Idea

The paper presents an alternative strategy to language modeling using RNNs. The authors attempt to use this to impute missing words in a sentence, to interpolate between the latent representations of 2 sentences, and also generate sentences by sampling from the space of the latent representation's prior probability.

2 Background

- The authors differentiate between types of sentence embeddings. Sequence Autoencoders have RNNs as encoders and decoders are are just used to regenerate the original text. Skip-Thought models are similar but the target sentence is different from the original sentence. Paragraph Vector models simply try to predict the words that are present in a given sentence.
- Variational autoencoders impose a prior distribution on the latent representation, but a standard autoencoder does not.
- The latent representation is usually parameterized by a diagonal Gaussian distribution.

3 Method

- The prior distribution of the latent representation acts as a regularizer.

- Unclear what a 'global latent representation' is. Intuitively, each sentence would have its own representation.
- The authors suggest KL-term annealing, which involves having a cost function like

  $$L(\theta; x) = \alpha(-KL(q_\theta(z|x)||p(z))) + E_{q_\theta(z|x)}[\log p_\theta(x|z)]$$

  where the value of α is raised from 0 to 1 during the course of the training. This can be thought of as a steady progression from a standard autoencoder to a VAE.
- Word-level dropout used to force the decoder to rely primarily on the latent space for the sentence generation.
- A beam search strategy is used by the decoder with beam-size 15.
- Training sequences are read from right to left, to shorten the word dependencies.
- An alternative to qualitative evaluation is presented by the usage of an adversarial classifier, which is partially similar to what a GAN does. Non-differentiability of the discrete RNN decoder network disallows usage of the adversarial criterion during training.
- Model is compared against RNNLMs.

4 Observations

- The paper doesn't talk about the presence of dead-zones in the latent space. This should be more of a problem due to the discrete nature of word representation. It also doesn't explain why the reconstruction error doesn't dominate, and only the KL-divergence error dominates. Perhaps several experiments with a fixed KL-divergence factor would be better to arrive at a correct loss-function balance.
- It is difficult to train models for which the KL-divergence term dominates.
- A word-dropout of about 75% seems to work best from qualitative assessments. (These could very well have been cherry-picked)

- RNNLMs seem to favour very generic sentences, and are less likely to be diverse.