Improved Variational Autoencoders for Text Modeling using Dilated Convolutions

1 Idea

The paper presents an alternative architecture to LSTM based VAEs. As shown in an earlier paper [1], LSTM-VAEs don't have a significant advantage over LSTM language models. The authors address this by using a dilated CNN decoder to vary the conditioning context of the decoder. The hypothesis is the the typical collapse of the loss function in favor of the KL-divergence term could be addressed by varying the contextual capacity of the decoder.

2 Method

- The authors use a typical LSTM based encoder model, use a dilated CNN as as the decoder of the VAE.
- The architecture of the encoder doesn't matter as long as the posterior of the latent representation resembles a Gaussian with unit variance.
- The idea of dilated CNNs was introduced with the intention of supplying varying contexts of words as features. As opposed to dense convolutions, dilated convolution skip time-steps to increase the receptive field of the operation, without increasing the computational costs. Dilations effectively introduce holes in a convolutional operation to be able to expand quickly.
- It is okay for the posterior (latent representation) to not completely mimic the Gaussian prior. This will ensure that the space of the latent probabilities offer good generative properties.

- Residual blocks are used for faster convergence and to enable building deeper architectures.

- Predictions at each step of the decoder is conditioned on the convolutional features concatenated with the latent variable z. Context, unlike in typical CNN architectures, is restricted to only words that appear in previous time-steps.

- The Gumbel softmax function is used as a continuous approximation of an otherwise discrete latent variable, in the framework for semi-supervised text classification.

- For unsupervised clustering, the authors still use a discrete label y to encode some information about an unlabeled text x, and the discrete label is then used for clustering.

- The authors use an LSTM encoder to obtain the latent representation z, followed by the dilated CNN to decode. The LSTM encoder is shared by the classifier (discriminator), since the final hidden state is fed to an MLP architecture to obtain a classification.

3 Observations

- The large CNN model (LCNN) performed marginally better than the LSTM language model as long as the encoder was pre-trained using the LSTM language model. So, this approach stills requires a pre-trained LSTM language model in-order to outperform it.

- It could be argued, as the authors do, that the dilated CNN architecture to incorporate a larger context of text helps improve language modelling and text classification performance, as evaluated by negative log-likelihood of the predicted sequences and perplexity.

References

[1] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.