Multi-space Variational Encoder-Decoders for Semi-supervised Labeled Sequence Transduction

1 Main Idea

The general idea [1] seems similar to style-transfer in text. Labeled sequence transduction is just a roundabout way of saying that a source text $x^{(s)}$ is to be transformed into a target text $x^{(t)}$ such that $x^{(t)}$ is conditioned on the labels $y^{(t)}$.

2 Background

- The morphological reinflection problem tries to change a sequence of characters of an inflected word. For example, convert "playing" into "played", given a set of labels $y^{(t)}$ such that $y^{(t)}_{pos} = verb$ and $y^{(t)}_{tense} = past$

3 Method

- MSVEDs (system proposed by this paper) use multiple discrete and continuous latent variables in a character based recurrent network to model the transduction of inflected words into their re-inflected forms.
- The basic architecture emulated a variational autoencoder. The encoder and decoder parameters are learned in the standard, way, by maximizing the variational lower bound on the marginal log-likelihood of the data. The neural network parametrized both the encoder and decoder weights using the VAE reparametrization trick.
- No explicit modeling seems to have been done to disentangle the latent representation z from the labels y

- Both encoder and decoder architectures are RNNs.
- The paper also describes a semi-supervised version of the architecture in which the labels y are inferred directly by a discriminative classifier on the target sequence $x^{(t)}$. This discriminative classifier is parameterized by an MLP that shares the initial layers of weights with the encoder of the network, and cuminates in an MLP and a Gumbel Softmax layer. The Gumbel softmax layer is used to obtain a continuous approximation of the latent variables, and the probability distribution varies with the temperature τ where $\tau = 0$ implies a one-hot encoded softmax, and $\tau > 0$ implies a less confident softmax output.
- The KL-term is gradually annealed from 0 to a λ_m, to prevent the latent representation collapsing into the Gaussian prior.
- Recurrent dropout is used in the form of omitting random words, and forcing the decoder to rely on the latent representation instead of just the ground-truth at each time-step.

4 Observations

- The experimental observations show that the bi-directional encoder/decoder model works better than the single directional model.
- The results also indicate a performance boost when the model is trained using unlabeled sequences, but the growth rate of performance gains grows slower as more data is added.
- The conventional method performs well when the reinflection just involves suffixing the original word, but the MSVED offers better generalized performance.

References

[1] Chunting Zhou and Graham Neubig. Multi-space variational encoder-decoders for semi-supervised labeled sequence transduction. *arXiv preprint arXiv:1704.01691*, 2017.