CS287 Final Project Abstract

Francisco Rivera[*] Jiafeng (Kevin) Chen[†] Yufeng Ling[‡]

March 27, 2019

1 Area

Variational autoencoders (VAE) are a powerful method for learning latent representations of high-dimensional objects; these representations preserve the underlying structure of the objects, and have useful features such as interpolation and natural clustering. In vector-quantized VAE (VQ-VAE), we discretize the representation space, using the vector quantization algorithm. Discretizing the representation space effectively compresses inputs, and in some cases, may lead to interpretable discrete latent codes. We are interested in using recent advances in vector-quantized variational autoencoding (VQ-VAE) to improve machine translation, particularly in low-resource settings.

VQ-VAE has been used (Kaiser *et al.*, 2018) to speed up machine translation. VQ-VAE achieves this by compressing sentences into shorter discrete representations, which are then translated using a encoder–decoder framework (e.g. transformer). The speedup is provided since the variational encoding and decoding parts are parallelizable. We anticipate that VQ-VAE may also be helpful in low-resource settings, since variational autoencoding is unsupervised. In particular, recent advances in VQ-VAE has shown that it can style transfer audio into being spoken by different speakers, in a fully unsupervised manner (van den Oord *et al.*, 2017). Speaker style transfer is a task that bears some similarities to translation, and we hope to leverage these recent advances for low-resource machine translation.

[*] frivera@college.harvard.edu
[†] jiafengchen@college.harvard.edu
[‡] yufengling@college.harvard.edu

2 Papers

- Our main paper is van den Oord, Vinyals *et al.* (2017), which introduces VQ-VAE. VQ-VAE differs from traditional VAEs in two crucial aspects: (1) VQ-VAE is discrete, and (2) the prior is learned.
- Kaiser, Roy, Vaswani, Pamar, Bengio, Uszkoreit and Shazeer (2018)
- Guzmán *et al.* (2019)

3 Time

We have already signed up for Monday April 22.

4 Baseline

We will be using the same Nepali-English and Sinhala-English datasets as in Guzmán *et al.* (2019). Those languages are only spoken by less than 20 million people in Nepal and Sri Lanka respectively. They are categorized by small amount out-of-domain parallel data and large amount of monolingual data.

With very limited supervised dataset ($\sim$ 500K sentences), the paper achieved BLEU scores of less than 8 for Nepali and Sinhala to English. With a combination of supervised translation data and monolingual data on the target side, BLEU scores increased to 15.1 for both. Finally, unsupervised learning generated BLEU score very close to 0. We intend to use those as baseline and improve upon them.

References

Guzmán, F., Chen, P.-J., Ott, M., Pino, J., Lample, G., Koehn, P., Chaudhary, V. and Ranzato, M. (2019). Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv preprint arXiv:1902.01382*.

Kaiser, L., Roy, A., Vaswani, A., Pamar, N., Bengio, S., Uszkoreit, J. and Shazeer, N. (2018). Fast decoding in sequence models using discrete latent variables. *arXiv preprint arXiv:1803.03382*.

van den Oord, A., Vinyals, O. *et al.* (2017). Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6306–6315.