Carnegie Mellon University
Computer Science Department

Speaking Skills Talk

Unsupervised Text Style Transfer using Language Models as Discriminators

Zichao Yang

Tuesday, January 22, 2019
10:00 am
GHC 6501

Binary classifiers are often employed as discriminators in GAN-based unsupervised style transfer systems to ensure that transferred sentences are similar to sentences in the target domain. One difficulty with this approach is that the error signal provided by the discriminator can be unstable and is sometimes insufficient to train the generator to produce fluent language. In this paper, we propose a new technique that uses a target domain language model as the discriminator, providing richer and more stable token-level feedback during the learning process. We train the generator to minimize the negative log likelihood (NLL) of generated sentences, evaluated by the language model. By using a continuous approximation of discrete sampling under the generator, our model can be trained using back-propagation in an end-to-end fashion. Moreover, our empirical results show that when using a language model as a structured discriminator, it is possible to forgoe adversarial steps during training, making the process more stable. We compare our model with previous work using convolutional neural networks (CNNs) as discriminators and show that our approach leads to improved performance on three tasks: word substitution decipherment, sentiment modification, and related language translation.

Presented in Partial Fulfillment of the CSD Speaking Skills Requirement.