Disentangled Representations for Manipulation of Sentiment in Text

1 Idea

The main idea of the paper is to change the style (sentiment) of a body of text while retaining its content

2 Background

Similar to the Persona-Based Neural Conversation Model [1], this paper also utilizes an embedding for a particular sentiment that the decoder is conditioned on.

3 Method

- This system uses a CNN for text encoding
- In the case of sentiment, 2 distinct probability distributions P_{source} and P_{target} are learned. The style transfer is achieved by traversing the manifold between these 2 distributions.
- The initial training phase just uses a variational autoencoder-like setup to recreate the original sentences.
- After the initial training, the CNN is trained to classify sentiment, thus, causing the distribution of the encoded sentences to diverge based on whether the sentence is positive or negative.

- Decoding is done by conditioning the start of the sentence on a start-of-sentence token and the sentence's encoding. The sentence generation ends when an EOS (end-of-sentence) token is generated.

4 Observations

It is not clear how the representations are disentangled. It seems like the sentence encoding itself encodes information about the sentiment and hence, the representations rely on the entanglement to generate the manifold which is traversed.

References

[1] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.