Review - Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation

Rakshan tej Baggam

January 2024

1 Introduction

- Unsupervised text attribute transfer automatically transforms a text to alter a specific attribute (e.g. sentiment) without using any parallel data, while simultaneously preserving its attribute-independent content. The dominant approaches are trying to model the content-independent attribute separately, e.g., learning different attributes' representations or using multiple attribute-specific decoders.
- However, it may lead to inflexibility from the perspective of controlling the degree of transfer or transferring over multiple aspects at the same time.
- To address the above problems, we propose a more flexible unsupervised text attribute transfer framework which replaces the process of modeling attribute with minimal editing of latent representations based on an attribute classifier.
- pecifically, we first propose a Transformer-based autoencoder to learn an entangled latent representation for a discrete text, then we transform the attribute transfer task to an optimization problem and propose the Fast-Gradient-Iterative-Modification algorithm to edit the latent representation until conforming to the target attribute.
- Extensive experimental results demonstrate that our model achieves very competitive performance on three public data sets.
- Furthermore, we also show that our model can not only control the degree of transfer freely but also allow transferring over multiple aspects at the same time

2 Text Attribute Transfer: A flexible Unsupervised Framework

- Text attribute transfer, the task of altering attributes like sentiment and style in a text, is vital for controllable natural language generation. Most existing approaches, due to the lack of parallel data, are unsupervised and struggle with flexibility.

- We propose a unique unsupervised text attribute transfer framework that overcomes limitations. Our approach uses a Transformer-based autoencoder with an entangled latent representation for both attribute and content, ensuring natural language integrity. The Fast-Gradient-Iterative-Modification (FGIM) algorithm, guided by a well-trained attribute classifier, efficiently edits the latent representation, allowing for minimal changes while conforming to the target attribute.

- Key contributions include a flexible method for controlling attribute transfer degrees and handling multiple aspects simultaneously. Our approach achieves competitive performance across three datasets, emphasizing text fluency and transfer success rate.

- Diving deeper into our approach, we're breaking away from the norm of separating attributes and content during unsupervised text attribute transfer. Instead, we focus on tweaking the combined latent representation of both attribute and content, a shift inspired by Lample et al.'s findings that this separation might not be necessary for effective text generation.

- Comparatively, our method shares some similarities with adversarial sample generation, where gradients alter continuous samples. However, we edit in the latent space, not directly on the samples, aiming for meaningful changes rather than just tricking classifiers.

- Moreover, we introduce activation maximization to text generation, a technique typically used for images but applied here to discrete texts. This involves encoding texts into a continuous latent space with an autoencoder and tweaking representations based on directions that highly activate the classifier.

- In a nutshell, our research not only changes how we approach unsupervised text attribute transfer but also brings new tricks, like activation maximization, to the table for more meaningful and controlled text generation.