Paper Review

Educating Text Autoencoders: Latent Representation Guidance via Denoising

Abraham, Frederic
i6262598

Gieshoff, Carsten
i6258391

Maastricht, February 2, 2021

Contents

The following is a review of the paper (Shen et al. 2019) published in 2020 at the International Conference On Machine Learning.

1 Summary

The paper introduces a denoisification aspect to Adversarial Autoencoder (AAE) that helps to maintain the geometry and relationships of the data space in the latent space allowing better results when using latent space arithmetic in applications like text generation. The training method applied is restricted to sequence data or more specific text data but there are similar approaches referenced for images.

First, Adversarial Autoencoders (AAE) are introduced. While such AAE are a first step towards improvement of the structure in the latent space (see gaussian distribution in left panel of figure 1), the additional denoisification introduced in this paper are a next step to further add robustness to the data representation in the latent space (see right panel of figure 1).

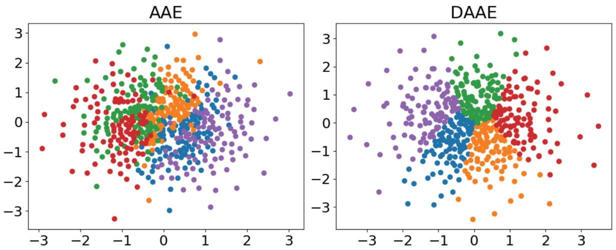

Figure 1: Latent representations learned by AAE and DAAE when mapping clustered sequences

How this additional denoising feature works is first shown on a small example before theory is developed and large scale experiemnts are conducted.
This first experiment, using synthetic data, shows quite clearly the results of this denoisification of the latent space. In the experiment they create 5 binary sequences with the length of 50 as the cluster centers. Then, noisification is performed by randomly flipping bits with a probability of 0.2, 100 permutations are created that are related to the original sequence. The architecture of the AE encodes the binary string into a two dimensional latent space so it can be plotted directly as seen in figure 1. Both models, the AAE and DAAE are able to archive a perfect reconstruction rate while

AAE looses the cluster structure in the latent space and fails to preserve the geometry of data. The left graph of figure 1 shows that with AAE sequences originating in different clusters are mapped to the same position as sequences from different clusters while DAAE is able to separate the clusters.

The authors then provide some theoretical results in a low complexity setting to back the observation from figure 1. In a couple of theorems it is established that 1) the optimal reconstruction error achievable by decoders that reconstruct from one-to-one encodings is independent of these encodings, 2) when adding intra-cluster noise the optimal reconstruction error can only be achieved when clusters from the data are retained in the latent space.
Then, the authors show that the observations from figure 1 and the simplified theoretical results generalize when applied to high complexity experiments. This includes showing that the encoder in their DAAE approach maintains cluster structure significantly better than competitive models (c.f. figure 2) and that the model outperforms other models in term of reconstruction and sentence generation quality.

Lastly, some experiments are provided that illustrate how the previous meta-results of retaining cluster structure in the latent space translate to the overall goal of achieving better results with latent space arithmetic for data generation and manipulation. The authors provide a range of qualitative experiments showing sentence tense inversion, sentiment transfer and sentence interpolation. Here, one can see that these manipulations work best for the proposed DAAE model.

2 Relevance

As pointed out in the summary, Adversarial AEs (AAEs) are a first step towards improving AEs w.r.t. the goal of useful data generation and manipulation in the latent space. While this step successfully ensures more meaningful sampling from a prior distribution in the latent space, different clusters of the data may well overlap in the latent space. This makes exploiting the arithmetic in the latent space, say by convex combination of latent vectors, difficult and results in low quality generations after decoding.
The additional introduction of denoisification/ perturbation delivers further progress in this area. With their method the authors aim to overcome the problem by separating data clusters in the latent space and thus make latent space arithmetic more usable in a intuitive fashion.

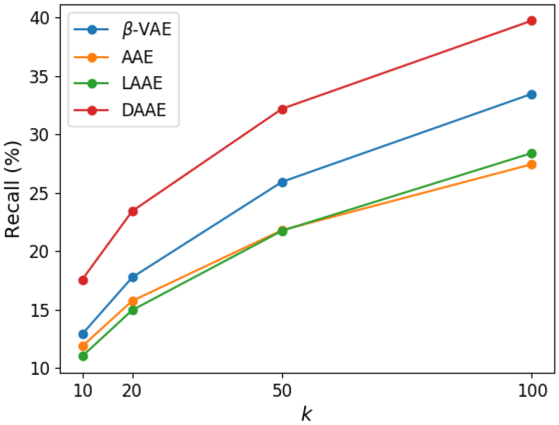

Figure 2: Recall rate of different autoencoders

3 Significance

It is quite clear that the method of denoisification AAEs (DAAEs) in the given context of text autoencoders outperforms related state of the art approaches. This is shown using quantitative experiments on meta-results and qualitative experiments on text generation and text manipulation.
With this the significance of a simple approach, that is easy to implement and that increases the reliability of text generation and reconstruction capabilities, is already given. Further more the possibility to do latent space arithmetic lays the ground brick for future work in text generation possibilities.
Also the significance is closely related to how well the noisification concept can be translated to other types of auto encoders, as the presented results and theory

do not confine to sequence based data used in NLP.

4 Novelty

The problem of generating meaning full sentences with some arithmetic is as almost as old as natural language processing itsself. The canonical example introduced by (Drozd, Gladkova, and Matsuoka 2016) is the embedding arithmetic "King" - "Man" + "Woman" ≈ "Queen" which they investigate whether analogous structure emerges in the latent space of our sentence-level models.
Thus, while the overall problem is not novel, the use of AE, especially AAE in this context has only arisen in recent years. It is worth noting that the models used in this paper (DAAEs) have been applied in other settings such as image processing previously (c.f. (Creswell and Bharath 2018)). Moreover, the general idea of using additional noise in the underlying data and thus building a more robust model has been applied in various settings.
Summarising, the authors manage to successfully apply recent advances to a new setting (NLP) aiming to solve a long existent problem.

5 Soundness and Evaluation

The paper operates on different levels of complexity; it introduces the given problem using synthetic data, conducts experiments on real datasets such as Yelp reviews [1] and provides theoretical results in a low complexity setting.
The paper is technically sound in the sense that it does provide mathematical theorems with traceable proofs. Building on these foundations the synthetic data experiments can be comprehended easily and the generalization to real datasets follows canonically. All three levels ob abstraction are concerted well with one another. Moreover, sufficient technical detail on the experiments is provided for the reader to follow all conducted procedures.

6 Clarity

In general the authors do a good job in balancing the detailed math and explaining the concept of their approach. This is done by three well constructed levels of abstraction used for introduction of the problem, theoretical foundation and further experiments. However, as this paper is an advancement to AAE the reader is in parts expected to have profound background knowledge in the related topics. A concrete example would be the metrics with which the DAAE are evaluated aren't explained because they are common in this field of research. Furthermore, is the approach to put the more elaborate proves in the appendix not a bad approach, as it does not interrupt the reading flow of the paper.

7 Comments

As indicated in several of the other sections, we experienced the structure and choice of abstraction levels in the paper really positively. This is further supported by an adequate amount and sensibly placed illustrations.
However, at some points some more detailed introduction to used methods and naming of explicit loss function might be useful to give a better intuitive understanding of why these methods/functions were used. Further, the section on Related Work could be more detailed to give the reader a broader overview on similar methods (if existent). The benchmark models used in experiments could be introduced in this context, too.

8 Questions

1. The qualitative examples provided are really impressive. Are there any bad examples that occurred in the creation of these examples and that show methodical limits to the used method, which might encourage further research?

2. Do you think there is a merit in investigating the geometry of conversations when represented as a sequence of vectors in the latent space? Are

[1] https://www.yelp.com/dataset

there e.g. recognisable geometric relationships between questions and answers? This might be helpful for e.g. chat-box AIs.

References

Creswell, Antonia, and Anil Anthony Bharath. 2018. “Denoising adversarial autoencoders.” *IEEE transactions on neural networks and learning systems* 30 (4): 968–984.

Drozd, Aleksandr, Anna Gladkova, and Satoshi Matsuoka. 2016. “Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen.” In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers,* 3519–3530.

Shen, Tianxiao, Jonas Mueller, Regina Barzilay, and Tommi Jaakkola. 2019. “Educating Text Autoencoders: Latent Representation Guidance via Denoising.” *arXiv preprint arXiv:1905.12777.*