Comparison of Autoencoding Techniques in Few-Shot Learning for Text Classification

Bartosz Chrostowski, Bogdan Jastrzębski, Jakub Drak Sbahi

April, 2022

1 Introduction

The following is report concerns comparison of autoencoding techniques for few-shot learning, a project for NLP course conducted at the Warsaw University of Technology, MiNI department, summer-semester 2022.

2 Agenda

Deep Artificial Neural Networks (DNN) revolutionized many fields of research, Natural Language Processing (NLP) in particular. In NLP, they achieve state of the art (SOTA) results, outperforming other techniques by a large margin.

Few-shot learning is a field of machine learning research, concerning learning from a few training samples. Supervised learning requires a dataset of annotated data, which is often difficult to acquire. In many cases, however, a dataset of data without annotations is available. Performing dimensionality reduction on a large corpus of not annotated samples first, and training a supervised model with few samples on their dimensionally reduced representations (embeddings), can significantly improve final model performance.

Embeddings are widely popular in NLP, e.g., word embeddings, like GloVe[10] or Word2Vec[8], and also numerous text embedding techniques, like Sent2Vec[9], Doc2Vec[5], Doc2VecC[1], Skip-through Vectors[4], Sentence-Bert[11] and many others. Autoencoding is also a popular method, studied especially for text generation, utilizing VAE[3], β-VAE[2], InfoVAE[18], AAE[7], DAAE[13].

Autoencoding techniques have been shown to be effective for few-shot learning, like CG-BERT[16], a form of CVAE[14], and others[12][15][6][17]. These techniques are usually a form of VAE. We hypothesize, that variational autoencoders, or adversarial autoencoders, that are techniques designed for generation, are not suitable for few-shot learning. Both techniques aim to create a latent variable with a known distribution. They are specifically trained to remove the structure of the data from the reduced representations. Collapsing clusters of data into one cluster (Fig. 1) is useful for generation, but can be harmful for few-shot learning.

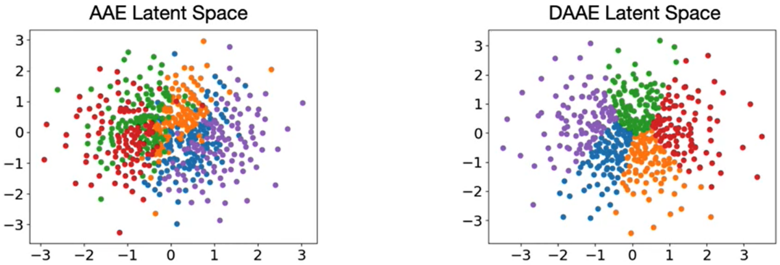


Figure 1: AAE and DAAE embeddings visualization (toy-example). For few-shot learning, ideally the clusters should be separated. It is impossible to discover clusters from this embedding, hence the representation is missing important information. Source:[13]

Our research concerns comparison of autoencoding techniques for few-shot learning in NLP. We aim to challenge the view on using generational techniques for few-shot learning. On a selected architecture, we will compare generational techniques like VAE, β-VAE, AAE, DAAE or others, against vanilla AE and Denoising Autoencoder (DAE). DAAE is a denoising adversarial autoencoder, i.e., it is an adversarial autoencoder, that is trained to restore original samples from their augmented versions. The DAAE technique provably generates well-formed embeddings, where similar observations have similar representations[13]. DAE is a simplified version of DAAE, without the adversarial loss (Fig. 2). We hypothesize, that such embedding technique will generate well-formed representations, without the loss of information about the structure of the data, which will improve performance for few-shot learning scenario. Optionally, we will refine the method with feature disentanglement techniques.

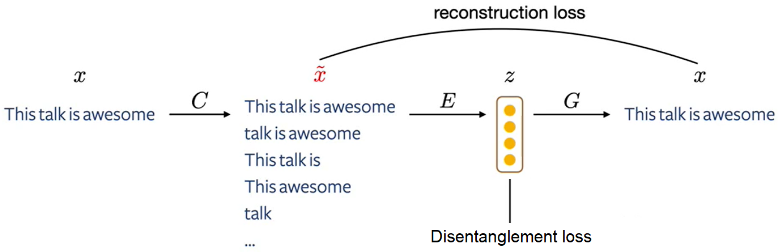


Figure 2: DAE architecture proposal. The denoising autoencoder augments observations and restores the original ones. Additionally, it could be improved using additional disentanglement loss. Architecture graph based on: [13].

3 Data

3.1 AG_news

The AG_news is a dataset of news articles. The dataset consists of four equal size classes, each of them containing 30000 samples. Articles belong to one of the four categories:

- World
- Sports
- Business
- Sci/Tech

The dataset is fully annotated. Fig. 11 shows a word cloud of the dataset. An important question we will have to answer is if the words presented here are popular in all classes or not. Some of them are stop-words, but others seem to be class related.

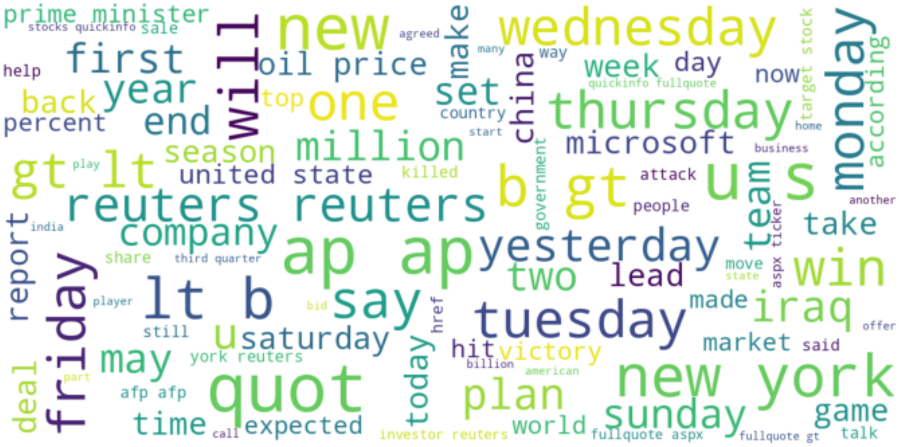


Figure 3: Word cloud for AG_news dataset. The most visible words do not belong to any of the categories, but they are news specific. The figure shows an important feature of the dataset, which is a bias towards a specific type of word choice. We can see, that popular words in articles are time-related, numbers, places, politics related concepts and stop-words.

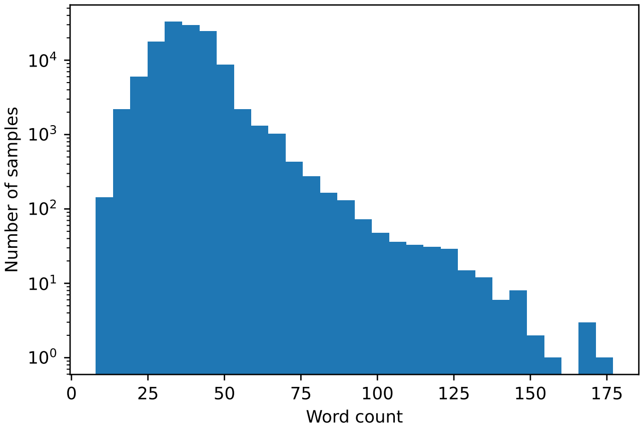
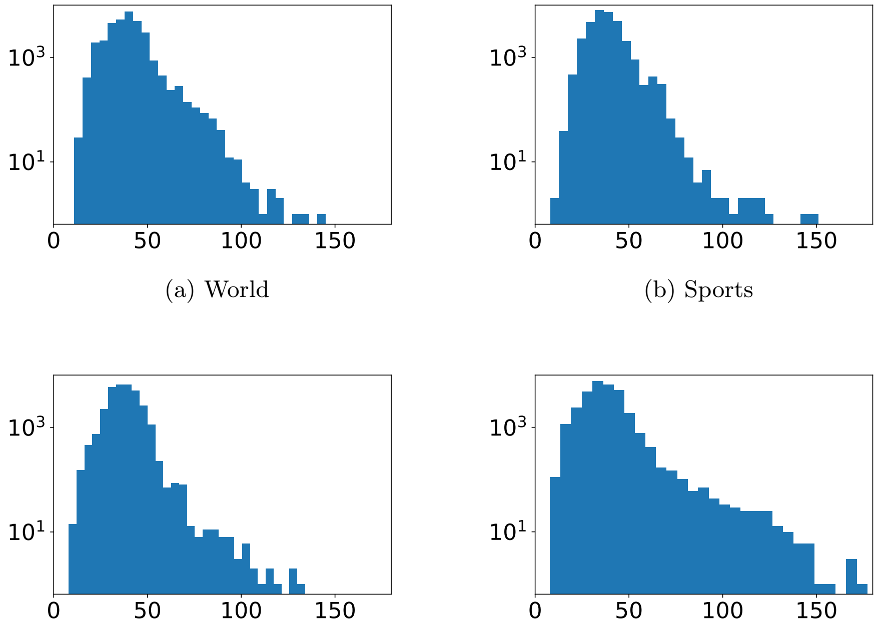


Figure 4: Histogram of text lengths in AG_news dataset

In the figure 4 we present histogram of text lengths. The average text length for this dataset is ≈ 37.84 and median is 37.0, however we can observe that there are some outliers with over 100 words. The large length of the observation has always been a challenge in NLP. Techniques like bag-of-words perform quite poorly on long, complicated sentences. Recurrent neural networks famously forget the first part of a long sentence, which led to the construction of LSTM and GRU. However, these techniques also have the same problem, albeit to a lesser extent. Attention modules significantly improved performance on long sequences, but their execution time scales quadratically with the sentence length. Average sentence size in AG-news is reasonably long to be interesting from the few-shot learning viewpoint, and short enough to allow for a fast training.



(a) World

(b) Sports

(c) Business

(d) Sci/Tech

Figure 5: Histogram of text lengths (words) in AG_news dataset split into classes

In figures 5 and 6 we present histograms of text length in each class. It is visible that distributions are quite similar, however the Sci/Tech have longer texts than other classes. Business was the category with shortest tail in compared distributions.

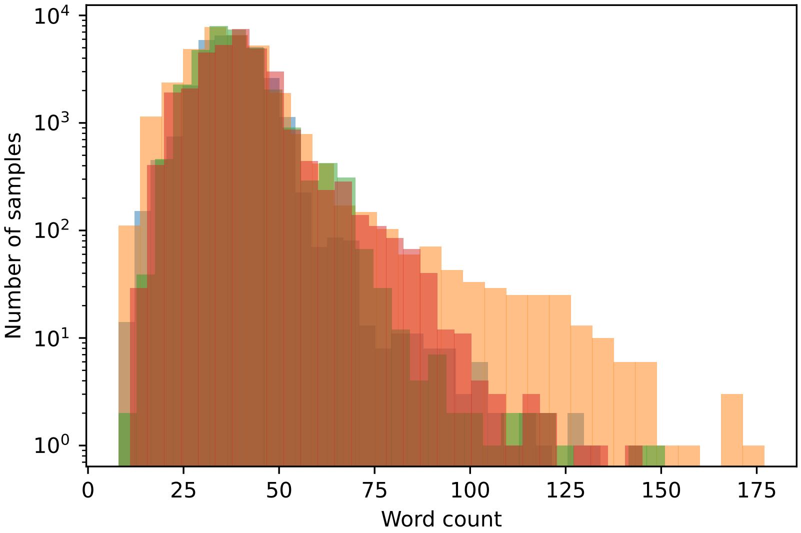


Figure 6: Histogram of text lengths in AG_news dataset split into classes overlaid

In Figure 7 there are the most common words appearing in the dataset. Most of them are just stopping words, but there is also a part of html code.

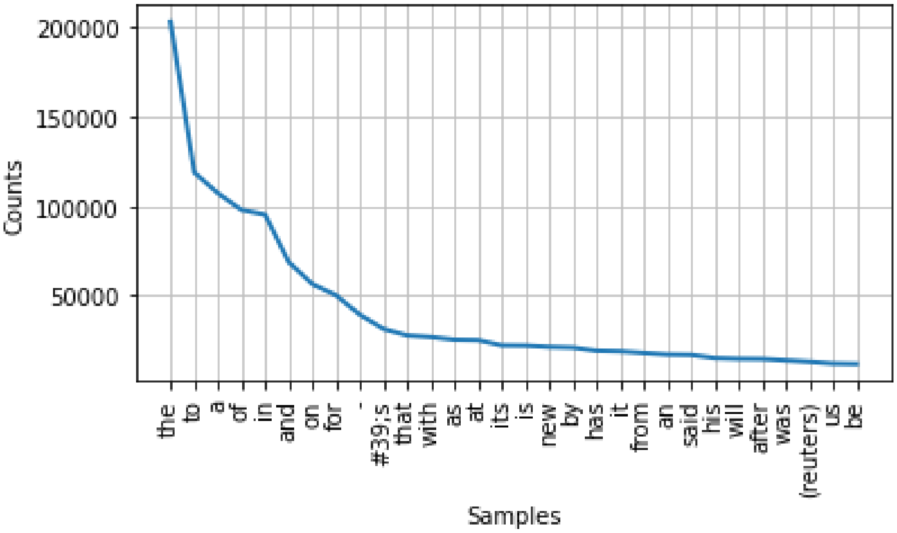


Figure 7: The most common words for AG_news dataset

Much more html code is visible in the Figure 8. It shows 10 most common trigrams. All of them are parts of html or some link to another page or subpage.

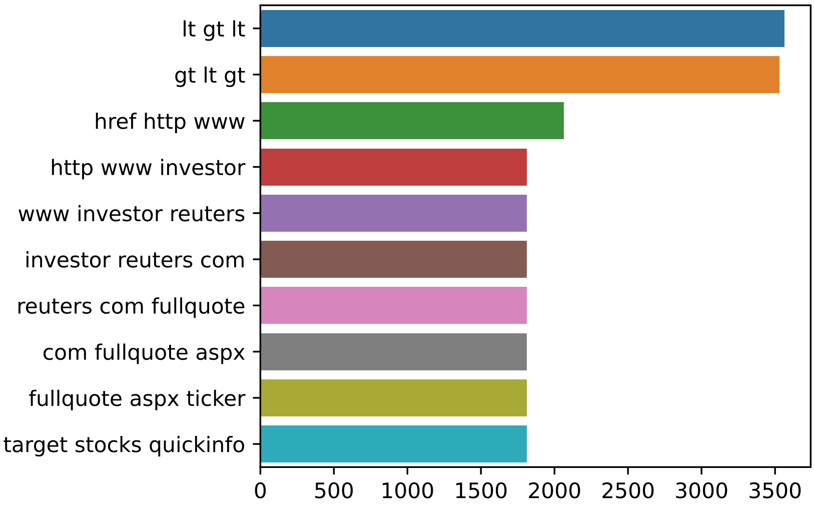
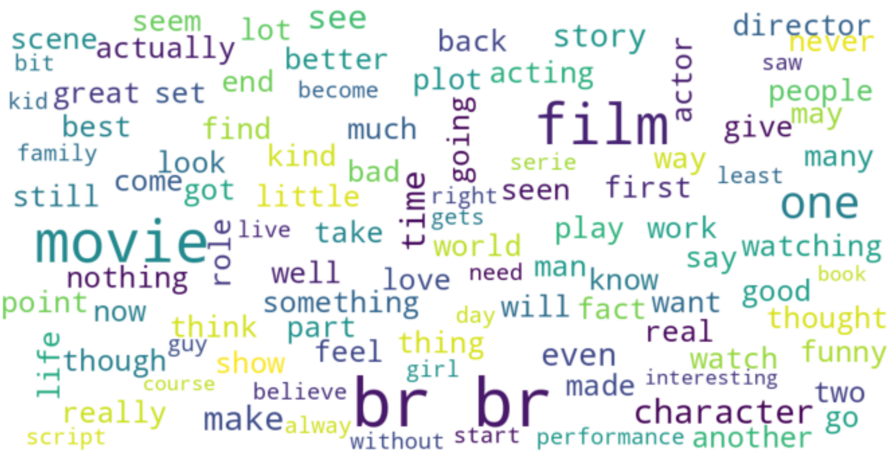


Figure 8: The most common words for AG_news dataset

3.2 IMDB



The IMDB dataset consists of 25000 labeled, and more unlabeled samples. Each record is a movie review. There are only two categories of reviews - positive and negative.

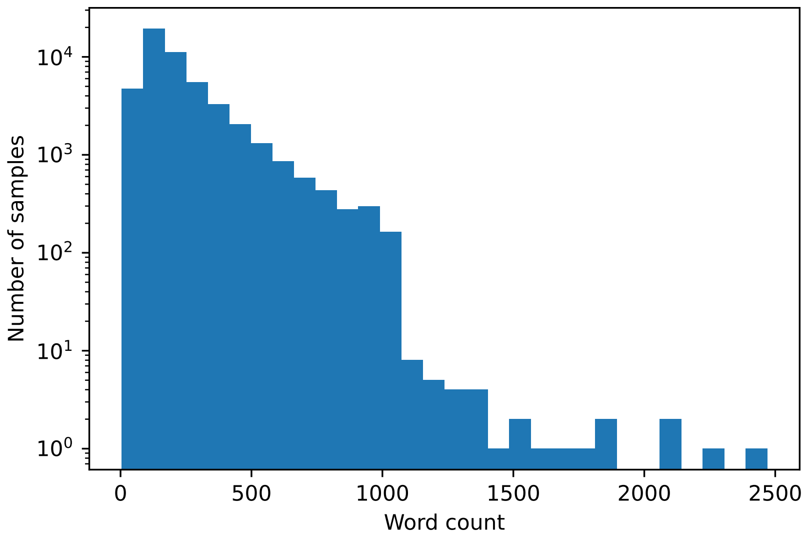
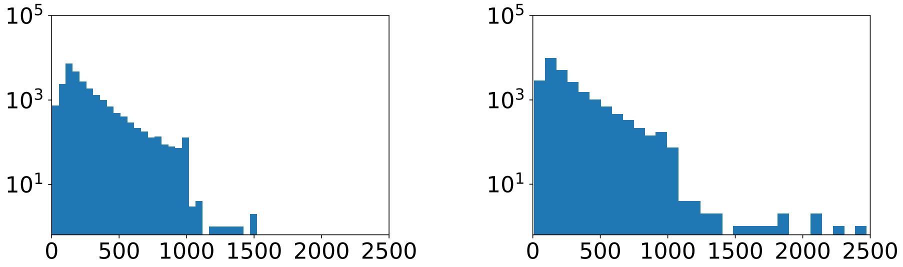


Figure 9: Histogram of text lengths in IMBD dataset



(a) Negative

(b) Positive

Figure 10: Histogram of text lengths (words) in IMBD dataset split into classes

Overall as can be seen in 9 text lengths in IMBD are much longer when compared to AG news dataset. The average text length in this dataset is ≈ 231.15 with median 173.0 and standard deviation ≈ 171.4. In case of IMBD dataset the distribution text lengths in classes distribution is almost the same for both classes. Positive class contains more outliers when compared to Negative class, however when it comes to the overall distribution of texts lengths both classes are similar.

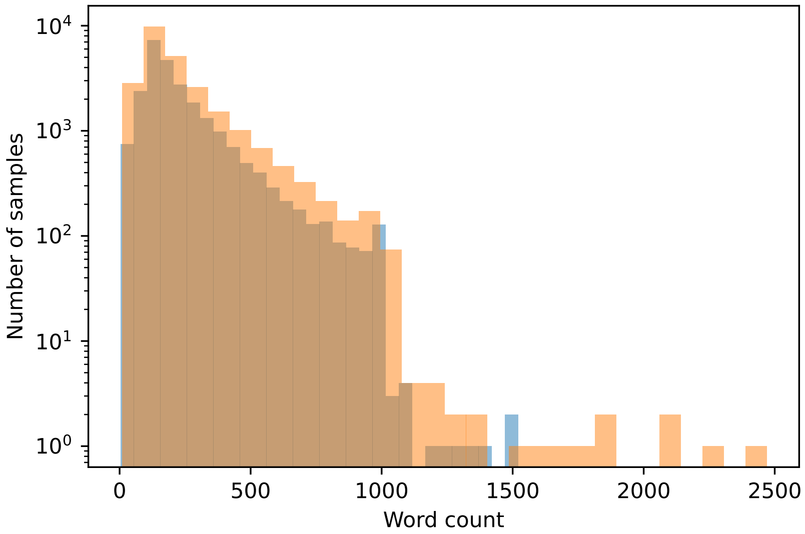


Figure 11: Histogram of text lengths in AG_news dataset split into classes overlaid

In Figure 12 there are the most common words appearing in the dataset.

Most of them are just stopping words, but there is also a part of html code for new line.

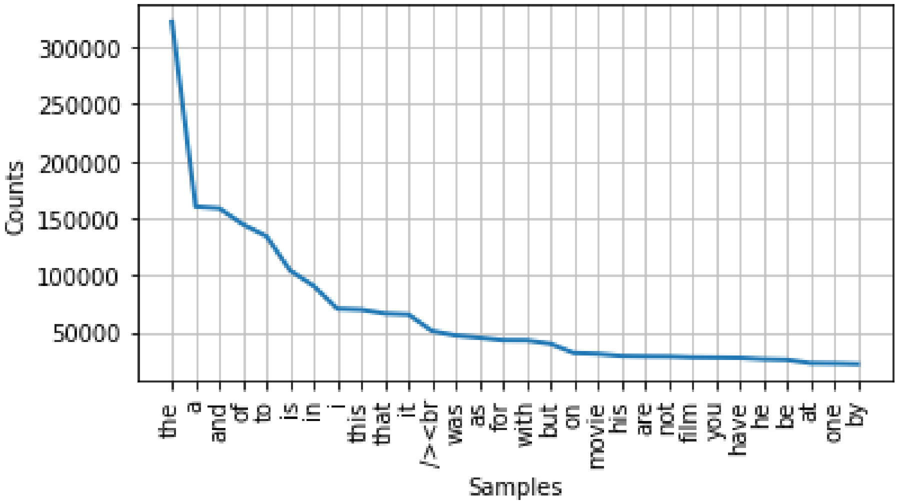


Figure 12: The most common words for IMDB dataset

In Figure 13 there are 10 most common trigrams. For IMDB dataset they are just simple phrases. Many of them are related to videos, or are used very often when describing a movie.

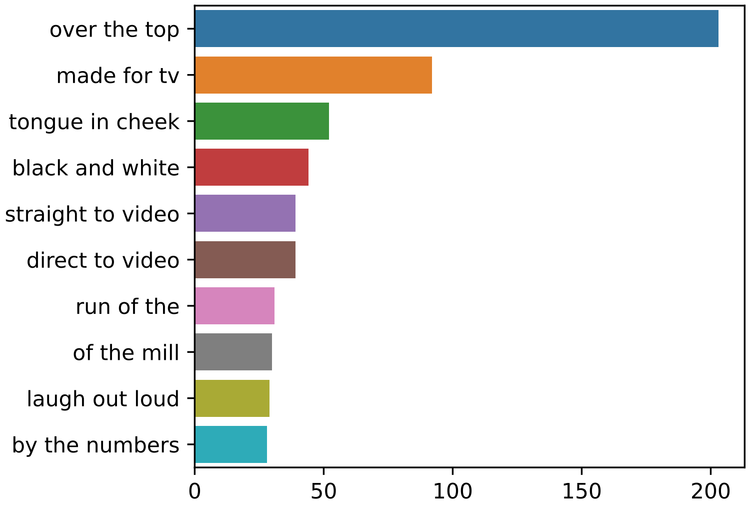


Figure 13: The most common trigrams for IMDB dataset

References

[1] Minmin Chen. "Efficient Vector Representation for Documents through Corruption". In: *CoRR* abs/1707.02377 (2017). arXiv: `1707.02377`. URL: `http://arxiv.org/abs/1707.02377`.

[2] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *ICLR*. 2017.

[3] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2014. arXiv: `1312.6114 [stat.ML]`.

[4] Ryan Kiros et al. "Skip-Thought Vectors". In: *CoRR* abs/1506.06726 (2015). arXiv: `1506.06726`. URL: `http://arxiv.org/abs/1506.06726`.

[5] Quoc V. Le and Tomás Mikolov. "Distributed Representations of Sentences and Documents". In: *CoRR* abs/1405.4053 (2014). arXiv: `1405.4053`. URL: `http://arxiv.org/abs/1405.4053`.

[6] Peirong Ma and Xiao Hu. "A variational autoencoder with deep embedding model for generalized zero-shot learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 11733–11740.

[7] Alireza Makhzani et al. "Adversarial Autoencoders". In: *CoRR* abs/1511.05644 (2015). arXiv: `1511.05644`. URL: `http://arxiv.org/abs/1511.05644`.

[8] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[9] Mahdi Naser Moghadasi and Yu Zhuang. “Sent2Vec: A New Sentence Embedding Representation With Sentimental Semantic”. In: *2020 IEEE International Conference on Big Data (Big Data)*. 2020, pp. 4672–4680. DOI: `10.1109/BigData50022.2020.9378337`.

[10] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: `10.3115/v1/D14-1162`. URL: `https://aclanthology.org/D14-1162`.

[11] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084* (2019).

[12] Edgar Schönfeld et al. “Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders”. In: *CoRR* abs/1812.01784 (2018). arXiv: `1812.01784`. URL: `http://arxiv.org/abs/1812.01784`.

[13] Tianxiao Shen et al. “Latent Space Secrets of Denoising Text-Autoencoders”. In: *CoRR* abs/1905.12777 (2019). arXiv: `1905.12777`. URL: `http://arxiv.org/abs/1905.12777`.

[14] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. “Learning Structured Output Representation using Deep Conditional Generative Models”. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015. URL: `https://proceedings.neurips.cc/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf`.

[15] Tailin Wu et al. “Meta-learning autoencoders for few-shot prediction”. In: *CoRR* abs/1807.09912 (2018). arXiv: `1807.09912`. URL: `http://arxiv.org/abs/1807.09912`.

[16] Congying Xia et al. “CG-BERT: Conditional Text Generation with BERT for Generalized Few-shot Intent Detection”. In: *CoRR* abs/2004.01881 (2020). arXiv: `2004.01881`. URL: `https://arxiv.org/abs/2004.01881`.

[17] Congying Xia et al. “Low-shot Learning in Natural Language Processing”. In: *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*. 2020, pp. 185–189. DOI: `10.1109/CogMI50398.2020.00031`.

[18] Shengjia Zhao, Jiaming Song, and Stefano Ermon. “InfoVAE: Information Maximizing Variational Autoencoders”. In: *CoRR* abs/1706.02262 (2017). arXiv: `1706.02262`. URL: `http://arxiv.org/abs/1706.02262`.