Title
Multi-Dimensional Disentangled Representation Learning for Emotion Embedding Generation

Author
Czyzycki, Evan

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Multi-Dimensional Disentangled Representation Learning for Emotion Embedding Generation

A thesis submitted in partial satisfaction

of the requirements for the degree Master

of Science in Computer Science

by

Evan Alexander Czyzycki

2022

ABSTRACT OF THE THESIS

Multi-Dimensional Disentangled Representation

Learning for Emotion Embedding Generation

by

Evan Alexander Czyzycki

Master of Science in Computer Science

University of California, Los Angeles, 2022

Professor Majid Sarrafzadeh, Chair

In the natural language processing (NLP) research community, disentangled representation learning has become commonplace in text style transfer and sentiment analysis. Previous studies have demonstrated the utility of extracting style from text corpora in order to augment context-dependent downstream tasks such as text generation. Within sentiment analysis specifically, disentangled representation learning has been shown to produce latent representations that can be used to improve downstream classification tasks. In this study, we build upon this existing framework by (1) investigating disentangled representation learning in the multidimensional task of emotion detection, (2) testing the robustness of this methodology over varying datasets, and (3) exploring the interpretability of the produced latent representations. We discover that closely following existing disentangled representation learning methods for sentiment analysis in a multi-class setting, performance decreases significantly, and we are unable to effectively distinguish content and style in our learned latent representations. Further work is necessary to determine the effectiveness of style disentanglement for text in multi-class settings using adversarial training.

The thesis of Evan Alexander Czyzycki is approved.

Kai-Wei Chang

Yizhou Sun

Majid Sarrafzadeh, Committee Chair

University of California, Los Angeles

2022

Contents

List of Figures

ACKNOWLEDGEMENTS

The experiments described in this study were performed in collaboration with Anaelia Ovalle [1] over the course of a project under the direction of Kai-Wei Chang [2] and Majid Sarrafzadeh [3].

[1] anaelia@cs.ucla.edu
[2] kw@kwchang.net
[3] majid@cs.ucla.edu

1 Introduction

Latent representation learning is a machine learning technique where opaque random variables can be inferred from empirical data. This has been shown to be useful across many domains, particularly in those where there is significant interest in quantifying variables that are difficult, expensive, or simply impossible to measure. Traditional methods within latent representation learning include linear methods such as Principal Component Analysis (PCA), nonlinear methods such as Gaussian Mixture Modelling [14]. With the increasing prevalence of deep learning and the increasing accessibility of associated computational resources, many recent studies have found success through training neural networks on a separate task and using a subset of the resultant layers to produce latent representations. For instance, Li et al. investigated emotion recognition through latent modelling of multi-channel EEG signals in this manner [8].

In the domain of natural language processing (NLP), latent representations are frequently used via embeddings of text corpora. Generally, text embeddings aim to capture the linguistic properties of input samples (usually words) as well as the relationships between them in a low-dimensional vector space. Word2Vec [10] and GloVe [11] are common sets of word embeddings and are used in various NLP tasks. Because these embeddings are generated without additional domain-specific context, analysis atop these embeddings struggle to perform in context sensitive tasks such as emotion detection [12]. Domain-specific embeddings generated using disentangled representation learning are useful for such tasks since prior knowledge can be encoded within the embeddings during training [15]. Latent representation learning has proven useful to this end via adversarial disentangling of content and style embeddings for sentiment analysis [6]. However, sentiment analysis is traditionally framed as a classification problem along a single axis spanning negative and positive sentiment. This does not easily extend to NLP classification tasks with many classes such as emotion detection [16].

In this study, we extend previous work in disentangled representation learning for sentiment analysis by considering a multi-dimensional framework of emotion. Through disentanglement of non-parallel text using variational autoencoders and adversarial loss functions, our model learns distilled representations of multi-dimensional emotion, rather than one-dimensional sentiment. This is done by distilling both style and content vector embeddings from the autoencoder output for use in downstream tasks.

2 Related Work

2.1 Disentangled Representation Learning

Existing disentangled representation learning methods show promising results in intuitive domain separation. This originates in computer vision with studies using adversarial encodings which frame the task as an optimal transport problem between the latent and observed data distribution[13, 3]. In the context of NLP, disentangled representations are learned over a discrete space rather than a continuous space. Previous work has shown that disentangled representations can be successfully produced in a discrete autoencoder setting using task-specific adversarial regularization [6] that is later leveraged to improve the ability to control text generation in variational autoencoders [5]. Adversarial frameworks are common in domain separation for text style transfer since they encourage the retention of relevant prior knowledge (e.g. sentiment, emotion).

2.2 Adversarial Training Frameworks for Representation Learning

Disentangled representation learning in text is often performed using simple feedforward neural network architectures or more sophisticated variational autoencoder architectures. These models are often trained using a framework which incorporates a multi-task adversarial learning objective in order to separate style from content [4, 2]. In previous studies, this adversarial framework not only

successfully separates content and style domains for style transfer, but also in doing so constructs the inductive bias necessary for truly meaningful disentanglement [9]. These frameworks have been used previously for sentiment transfer in text by linearly interpolating a style vector produced by adversarial training over a multitask objective [17].

2.3 Textual Style Transfer

Seen originally in Fu et al.[4], a cumulative multi-task adversarial infrastructure for style transfer is able to achieve state-of-the-art results in textual style transfer. This is later extended [7, 13] in the domain of sentiment transfer and generation of styled text corpora. Multi-dimensional emotion latent representation has been explored [16] in the form of a static, non-generative disentanglement method using a simple projection of word embeddings onto an emotion subspace. However, discrete multi-dimensional style transfer using variational autoencoders is relatively new and undergoing active study. In this paper, we build upon an existing adversarial sentiment analysis framework to encode latent representations of distinct emotions.

2.4 Emotion Unify Dataset

Emotion Unify Dataset is our chosen dataset for multi-dimensional emotion detection [1]. This is an aggregate dataset over the following existing datasets:

- AffectiveText
- Blogs
- CrowdFlower
- DailyDialogs
- Electoral-Tweets
- EmoBank
- EmoInt
- Emotion-Stimulus
- fb-valence-arousal
- Grounded-Emptions

- ISEAR
- SSEC
- Tales
- TEC

These datasets are aggregated over the following common label set of emotions:

- joy
- anger
- sadness
- disgust
- fear
- trust
- surprise
- love
- confusion
- anticipation
- noemo (no emotion)

3 Methods

In our analysis, we use a sequence-to-sequence variational autoencoder (VAE) to learn latent embeddings of text data with multi-class labels. This VAE uses a gated recurrent unit (GRU) recurrent neural network (RNNs) as both an encoder and a decoder. The encoder in our architecture is a bidirectional GRU RNN with a hidden dimensionality of 100, 8 hidden recurrent layers, and an output size of 136 (style vector of dimensionality 8, and a content vector of dimensionality 128; this mirrors the implementation found in John et al[7]). The decoder is a unidirectional GRU RNN with an input size of 136, a hidden dimensionality of 100, and 6 hidden recurrent layers.

The data used in our experiments contained text data associated with emotion labels from the Unify Emotion Dataset [1]. Our analysis was run on a subset of this dataset which is limited to only three class labels. Our text data was preprocessed using a tokenizer built into SpaCy[4]. Tokens

[4] https://spacy.io/

contained in the standard NLTK[5] stopword list and all instances of punctuation were filtered out. Then, tokens are transformed into word embeddings using the pre-trained distributed GloVe model trained on 6 billion tweets. The word embeddings are vectors of length 100.

In order to ensure both separation of the style and content vectors and the embedding of our prior labeling, we use a multitask loss function with several terms. Our loss function is a linear combination of the following terms (where coefficients are hyperparameters):

1. Kullback-Leibler (KL) divergence of the style embedding
2. KL divergence of the content embedding
3. Cross-entropy loss of a logistic regression classifier predicting the multi-class label using the style embedding
4. Adversarial cross-entropy loss of a logistic regression classifier predicting the word distribution of the corpus using the style embedding
5. Cross-entropy loss of a logistic regression classifier predicting the word distribution of the corpus using the content embedding
6. Adversarial cross-entropy loss of a logistic regression classifier predicting the multi-class label using the content embedding

The full loss formula is:

$$loss = a(1) + b(2) + c(3) - d(4) + e(5) - f(6) \tag{1}$$

where a, b, c, d, e, and f are hyperparameters.

[5] https://www.nltk.org/

We maximize terms (1) and (2) to most closely approximate the prior distribution with our generative model John et al [7]. We maximize terms (3) and (5) in order to ensure that style information is contained in the style embedding and content information is contained in the content vector respectively. We *minimize* terms (4) and (6) in order to ensure that style information is not contained in the content vector and that content information is not contained in the style vector. By optimizing this multi-task loss function, we generate style and content embeddings that encode style and content information respectively with minimal overlap in information.

4 Experiments & Results

4.1 Reimplementation

We were able to successfully replicate the Text Style Transfer VAE model found in John et al [7]. Having implemented our model in PyTorch, we iteratively reproduced the fundamental elements of the paper. We began by testing our implementation on the Yelp dataset used in the original study under the context of binary sentiment classification. As seen in Figures 2 and 3, we were successful in reproducing the results of the previous study. Then, we extended this framework to train on the Emotion Unify Dataset [1]. We expected to find that the model could handle this level of dimensionality and, upon clustering, observe clear segmentation into emotion classes.

4.2 Multi-class Emotion Detection Extension

As shown in Figure 3, we observe the KL-divergence decrease steadily for style and content vectors when run on the original Yelp dataset. This mimics the results contained in the original study, and this implies that we are able to successfully minimize our multipart loss function and disentangle our text.

However, upon switching datasets to the Emotion Unify Dataset, we encountered multiple

issues that resulted in poor performance. Firstly, switching to the Emotion Unify Dataset resulted in greatly increasing the cardinality of our vocabulary. Our vocabulary increased from 9,000 words when using the Yelp dataset to 60,000 words when using the Emotion Unify Dataset. This resulted in a significant increase in training time, as a larger vocabulary size combined with a larger label set size results in an exponential increase in compute time. The degradation in performance can be observed by comparing Figures 2 and 3 with Figures 4 and 5.

We also observed that the model was very sensitive to the KL weight annealment scheduling in both the Yelp and Emotion dataset. Unlike our experiment on the Yelp dataset however, the KL divergence did not decrease in our experiment on the Unify Emotion Dataset. We believe the cause of this to be the multi-class classification embedded into the loss functions that serve as the adversarial regularization of the space. Because of the greater difficulty of multi-class classification compared to binary classification, it was likely much more difficult to optimize over a loss function that includes cross-entropy loss over a multi-class classification.

As a result, we were unable to see clear separation in the embedding space pertaining to each class, indicative of more complex disentanglement than we had originally considered. This can be observed by comparing Figure 2 in John et al. [7] and Figure 1 in this study which contain t-SNE plots of the disentanglements in the original binary study and t-SNE plots of the disentanglements in our multi-dimensional study respectively. It is apparent that we fail to produce clearly separated groups over only three emotion classes.

In the original study, the authors linearly interpolate across the embedding space to achieve style transfer across their binary classes. We recognize that this method becomes increasingly complex with additional dimensions, even if we had observed clear separation in latent space. Text style transfer over multiple classes therefore requires additional future study.

5 Conclusion

5.1 Summary

Summarily, we were able to successfully reimplement existing methods in disentangled representation learning for sentiment analysis [7]. However, in extending this methodology to data containing multi-class emotion labels, we see a significant degradation in performance. We were not able to successfully produce disentangled clusters in latent space with as few as three labels. We believe that the cause of this performance degradation is the usage of a multi-class cross-entropy loss embedded in the overall loss function. Additional study is required to address this.

5.2 Future Work

As is apparent in our results, it is nontrivial to extend the methodology found in John et al. [7] to a setting with more than two class labels. Before any additional study is conducted using the Emotion Unify Dataset, we plan to reimplement the findings of John et al. [7] on this dataset. It is illogical to continue extended study on this dataset if the results of the original study cannot be reproduced on the new data. Following, we plan to investigate the effects of modifications to the reimplemented loss functions, the performance impact of separate style vectors per label in the data (e.g. 3 style embeddings for 3 labels), and the effects of modifications to the KL divergence loss scheduler since this seems to be the most sensitive aspect of our model training. Given additional computational resources and time, we also plan to perform a more exhaustive hyperparameter search.

(a) Content space (b) Style space

Figure 1: Style and content space generated over Emotion Unify Dataset.

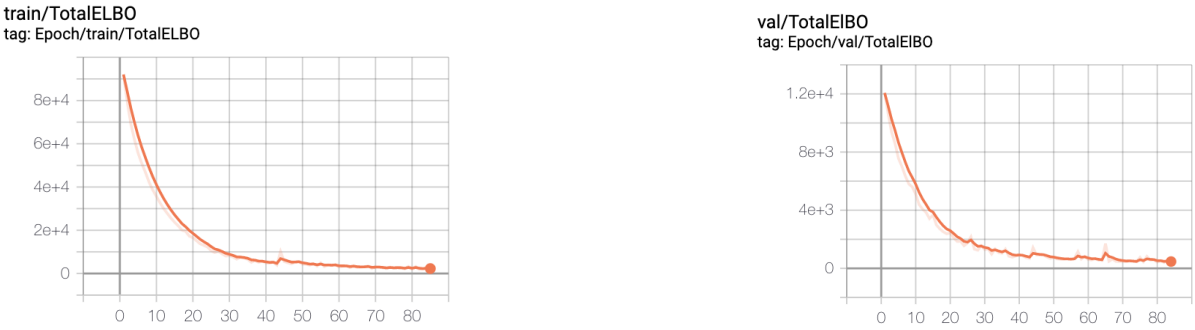

Figure 2: ELBO convergence over Yelp training data in our reimplementation.

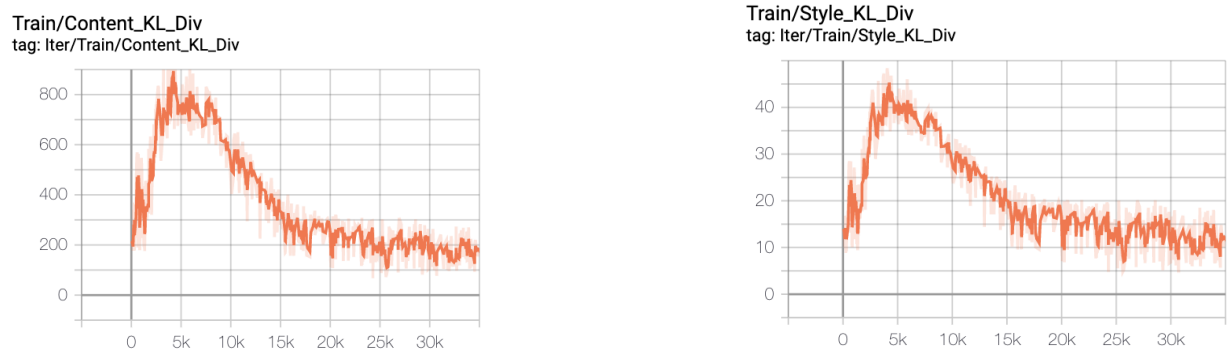

Figure 3: KL divergence convergence over Yelp training data in our reimplementation.

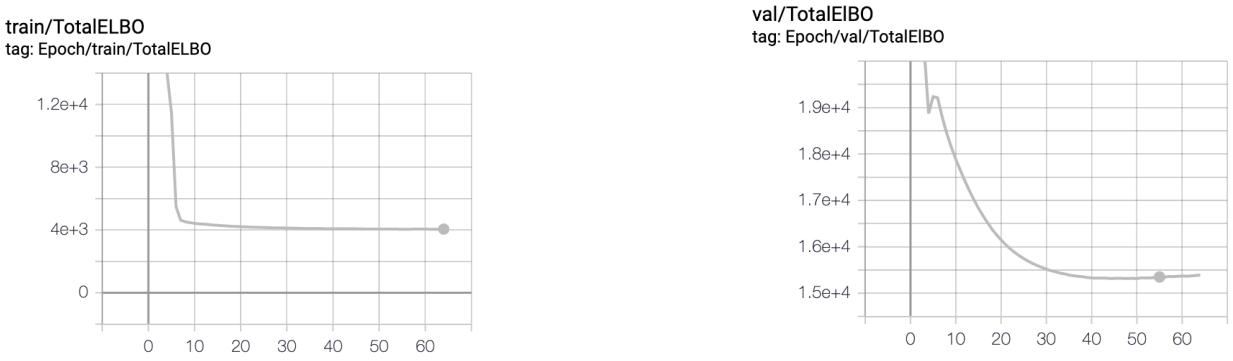

Figure 4: ELBO convergence over Unify emotion training data.

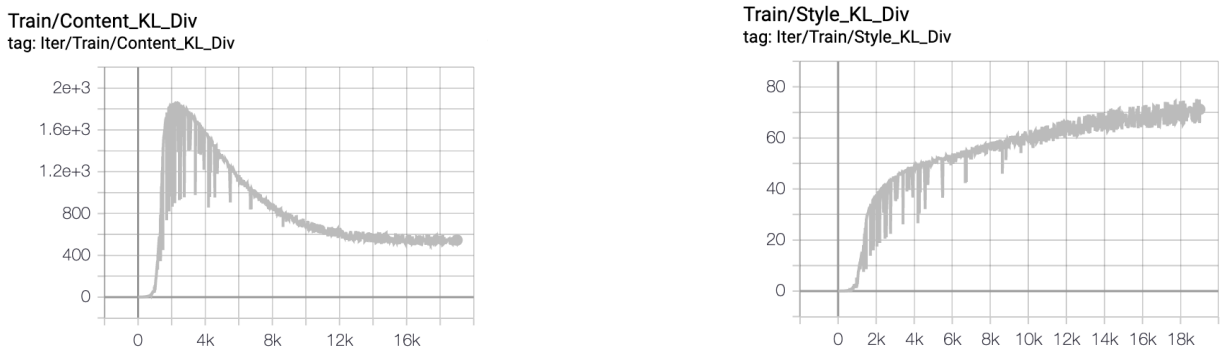

Figure 5: KL divergence convergence over Unify Emotion training data.

References

[1] L.-A.-M. Bostan and R. Klinger. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.

[2] M. Chen, Q. Tang, S. Wiseman, and K. Gimpel. A multi-task approach for disentangling syntax and semantics in sentence representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2453–2464, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016.

[4] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan. Style transfer in text: Exploration and evaluation, 2017.

[5] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. Towards a definition of disentangled representations, 2018.

[6] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1587–1596. JMLR.org, 2017.

[7] V. John, L. Mou, H. Bahuleyan, and O. Vechtomova. Disentangled representation learning for non-parallel text style transfer, 2018.

[8] X. Li, Z. Zhao, D. Song, Y. Zhang, J. Pan, L. Wu, J. Huo, C. Niu, and D. Wang. Latent factor decoding of multi-channel eeg for emotion recognition through autoencoder-like neural networks. *Frontiers in Neuroscience*, 14, 2020.

[9] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. 2018.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013, 01 2013.

[11] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[12] A. Seyeditabari and W. Zadrozny. Can word embeddings help find latent emotions in text? preliminary results. 2017.

[13] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6833–6844, Red Hook, NY, USA, 2017. Curran Associates Inc.

[14] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders, 2017.

[15] Z. Wang, G. H. T. Yeo, R. Sherwood, and D. Gifford. Disentangled representations of cellular identity. In L. J. Cowen, editor, *Research in Computational Molecular Biology*, pages 256–271, Cham, 2019. Springer International Publishing.

[16] Z. Wu and Y. Jiang. Disentangling latent emotions of word embeddings on complex emotional narratives. 2019.

[17] J. J. Zhao, Y. Kim, K. W. Zhang, A. M. Rush, and Y. LeCun. Adversarially regularized autoencoders. In *ICML*, 2018.