CycBERT: Enriching Pre-Trained Transformer Encoders by Leveraging Multimodal Cyclic Translation Networks

Loten Noy (ID. 201466349), Gil-ad Katz (ID. 313591851)

Submitted as final project report for the NLP course, RUNI, 2023

1 Introduction

In today's fast-paced world, Natural Language Processing (NLP) models have become integral to our daily lives. From virtual assistants to language translation tools, these models are behind the scenes, working to make our interactions with technology smoother and more intuitive. At the core of their success lies the concept of text encoding, a fundamental process that allows machines to understand and interpret human language.

As the field of NLP continues to progress, one challenge that researchers and developers face is the diminishing returns of simply adding more parameters to improve model performance when using text-only data for training. Over time, it has become evident that solely relying on textual information may reach a point of convergence regarding the value it can provide. Consequently, there is a growing interest in exploring multimodal models to overcome this limitation and unlock new possibilities.

Multimodal models are machine learning models designed to process and integrate information from multiple modalities, such as text, images, audio, and video. These models aim to capture rich and complementary information in different modalities to enhance various tasks. For instance, emotions and sentiments are often conveyed through words, tone of voice, facial expressions, and visual cues. By leveraging multimodal embeddings, we can create a holistic representation of sentiments, considering these additional sources of information. This approach allows us to develop more nuanced and accurate sentiment analysis models that grasp the essence of emotions conveyed through various channels.

The proposed method explores how multimodal models can enrich text encoders by leveraging cross-modal learning performed by Multimodal Cyclic Translation Networks [1]. By incorporating information from diverse sources like text, images, and audio, text encoders gain a broader perspective, leading to more effective encoding of text-only information. This combination of

multimodal and text domains opens new possibilities in sentiment analysis, language understanding, and other NLP applications that benefit from non-textual communication. It uses concepts from machine translation tasks, such as seq-to-seq architecture, to enforce cross-modal learning given a multimodal training dataset and generate textual encoders that are robust to some textual discrepancies in the data and are more proficient in noticing subtle nuances in textual communication that are crucial for sentiment and emotion recognition classifiers or regression models. According to our experiments, a regression model utilizing a multimodal, continuously pre-trained BERT [2] encoder outperformed a similar regression model that relied on a BERT model trained solely on the same data without incorporating a cross-modal learning architecture. The results demonstrate that leveraging the multimodal approach led to superior sentiment regression scores, highlighting the significant benefits of cross-modal learning in enhancing the model's performance.

1.1 Related Works

Sentiment analysis in NLP aims to identify emotional tones expressed in text. Early methods used traditional machine learning. These approaches involved extracting features from the text, such as word frequencies [3], n-grams [4], or sentiment lexicons [5], and using classifiers like Support Vector Machines (SVMs) [6] or Naive Bayes to make predictions. However, deep learning revolutionized the field, starting with the introduction of word embeddings, such as Word2Vec [7] and GloVe [8], that further improved sentiment analysis performance by representing words in dense vector spaces, capturing semantic relationships between them, only to be followed by Transformer-based models like BERT [2] that enabled rich context analysis using an attention [9] mechanism that grasps delicate nuances among textual tokens by attending them dynamically.

1.1.1 Multimodal Models

Recently, multimodal sentiment analysis gained attention [10]. Fusion methods [11] and neural network models have been explored for joint representations. Generative methods like GANs [12] and conditional generative models were also used to transfer information from supporting modalities (acoustic, visual) to the main language modality. MCTN [1] (Multimodal Cyclic Translation Networks) addresses the sequential dependency of modality translations, remains effective even when modalities are missing and shows promising results in generating robust representations. The initial work involved experimenting with an ad-hoc RNN-based sequence-to-sequence architecture. This inspired adoption of a similar design for continuous pre-training of a large language encoder, such as BERT.

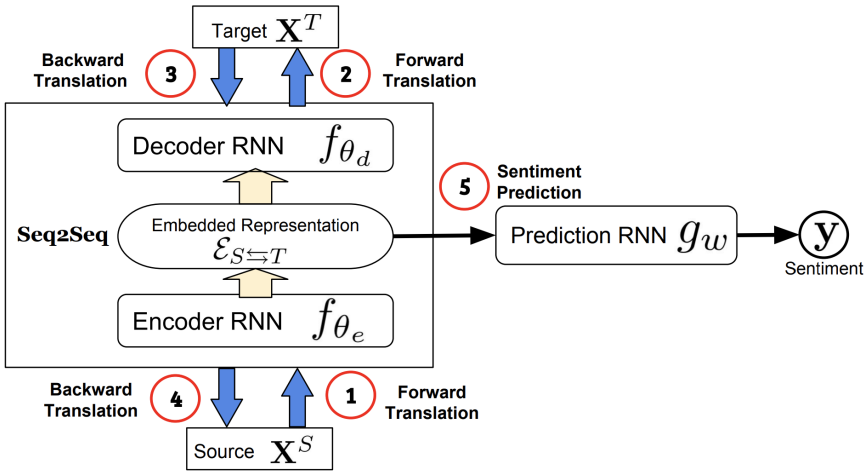

Figure 1: The MCTN Bi-modal architecture proposed in the original paper. The source modality is X^S and the target modality is X^T. The joint representation $E_{S \rightleftharpoons T}$ is obtained via a cyclic translation between X^S and X^T. Next, the joint representation $E_{S \rightleftharpoons T}$ is used for sentiment prediction. The model is trained end-to-end with a coupled translation-prediction objective. At test time, only the source modality X^S is required.

2 Solution

2.1 General approach

Our approach leverages the efficacy of BERT and similar transformers in encoding textual data for downstream tasks. By extending BERT with a corresponding visual modality during continuous pre-training, we aim to enhance sentiment analysis. This augmentation enriches the model's understanding of sentiment nuances through both linguistic and non-verbal cues. The chosen MCTN architecture aligns with BERT's common unimodal use-cases, ensuring practical compliance. Our integrated model seeks to advance sentiment analysis accuracy and adaptability while preserving BERT's strengths in textual comprehension. We conducted experiments, varying the weights assigned to different loss components. This enabled us to explore the varying influence of modalities on sentiment analysis performance. This approach deepened our understanding of how each modality contributes to sentiment analysis accuracy, revealing the intricate dynamics of multimodal interactions.

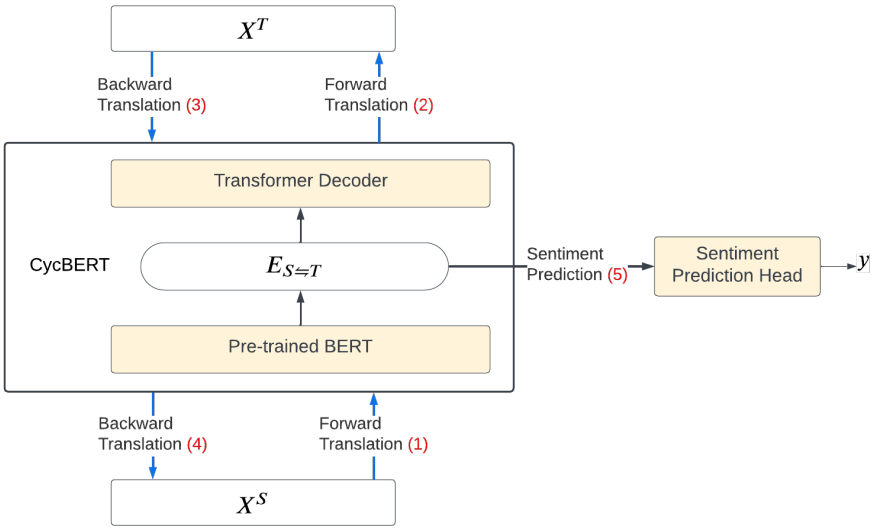

Figure 2: CycBERT Bi-modal architecture. The source modality X^S and the target modality X^T are cyclic translated to form the joint representation $E_{S \rightleftharpoons T}$. Modality encoding is done using pre-train BERT model, where a multi-layer transformer decoder model performs decoding. A linear prediction head performs sentiment classification/regression. During inference time, only the augmented BERT model and prediction head are required for text encoding and sentiment prediction

2.2 Design

2.2.1 Data Modeling

Similar to the MCTN approach, our data modeling strategy involved aligning visual modality features with OpenFace [13] attributes extracted from video segments to correspond with the textual tokens provided by the BERT uncased tokenizer. This alignment resulted in a dataset structure comprised of triplets: (X_l, X_v, y). X_l and X_v consist of sequences, each sharing a similar k length, denoted $(X_l^1, ...X_l^k)$.

Notably, each X_v^i represents the average of all visual features associated with the appearance of the WordPiece language token X_l^i. This methodology establishes a coherent link between textual and visual cues.

2.2.2 CycBERT Architecture

The newly designed architecture closely resembles the MCTN model. However, we've replaced the sequence-to-sequence encoder with a pre-trained BERT

model (uncased), while the decoder component remains an untrained transformer decoder model. Given the inherent structure of transformers operating in predefined embedding spaces, we incorporated an extra linear layer. This layer transforms the target modality's initial vector representations into dimensions aligned with BERT.

2.2.3 Cyclic Translation Loss

As demonstrated in the original MCTN paper, we incorporated a cyclic translation loss into the training process to enforce the creation of robust representations that capture the information from both source and target modalities. The training loss is a coupled Translation-Prediction objective [1] where the first two losses a the forward translation loss L_t defined as

$$L_t = \mathbf{E}[\ell_{X^T}(\hat{X}^T, X^T)]$$

and the cycle consistency loss L_c defined as

$$L_c = \mathbf{E}[\ell_{X^S}(\hat{X}^S, X^S)]$$

where ℓ_{X^T} and ℓ_{X^S} represent the respective loss functions. Finally, the prediction loss L_p is defined as

$$L_p = \mathbf{E}[\ell_y(\hat{y}, y)]$$

with a loss function ℓ_y defined over the labels. The overall objective used for training was therefore L defined as

$$L = \lambda_t L_t + \lambda_c L_c + L_p$$

where λ_t, λ_c are weighting hyperparameters.

Using discrete sentiment score datasets (non-binary), our approach involved applying distinct loss functions to each loss component. To ensure robustness against outliers, we employed the Smooth L1 loss for translation losses ℓ_{X^T} and ℓ_{X^S}. Simultaneously, we utilized Mean Squared Error (MSE) as ℓ_y loss for the sentiment scores, enhancing the model's sensitivity to inaccurate predictions. This strategic selection of loss functions aimed to balance the model's ability to handle extreme cases while maintaining accuracy across the sentiment spectrum.

3 Experimental results

3.1 Data

We used The CMU-MOSI [14] dataset for training, which is a multimodal dataset for sentiment analysis and emotion recognition. It comprises a diverse collection of video clips from YouTube, covering a wide range of topics, contexts, and speakers. The dataset includes spoken language, transcribed text, and corresponding visual and acoustic features. This is also the dataset experimented with in the original MCTN paper. We used a movie review dataset containing more than 80K textual movie reviews and their sentiment score for unimodal evaluation.

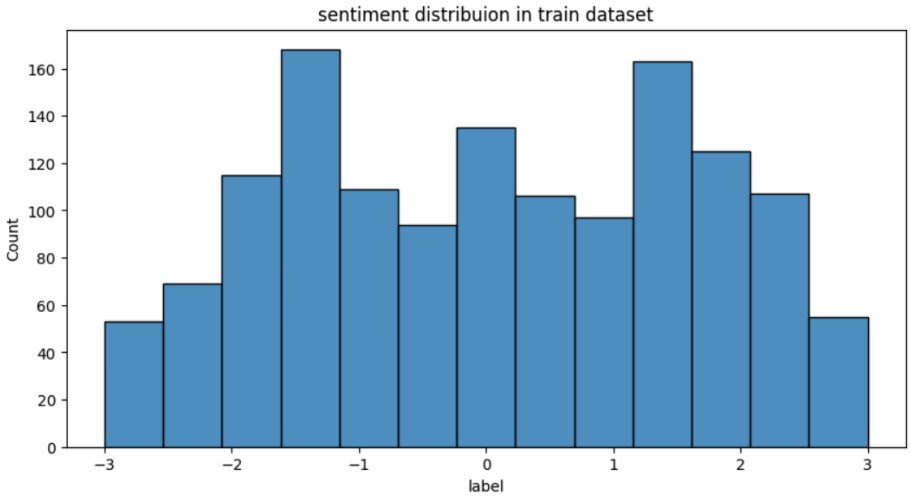

Figure 3: CMU-MOSI training data sentiment distribution. The data contained a total of 1396 labeled data points used for training throughout the described experiments

3.2 Training Setup

We implemented the CycBERT architecture. Pre-trained on the 'bert-base-uncased' variant, the BERT model served as the foundational sequence-to-sequence encoder. Based on an untrained transformer architecture, the decoder was designed with two layers for contextual decoding with 8 attention heads each. The prediction head was a dropout regularization (0.1) followed by a linear layer producing a single output. As mentioned, we used Smooth L1 loss for transnational loss measurements and back-propagation and MSE loss for sentiment prediction corrections.

Regarding training parameters, we employed the Adam optimizer with a learning rate of $1e-5$ to optimize the model's parameters. The lambda values (λ_t and λ_c) were dynamically adjusted per run to weigh the influence of the translation and cycle consistency losses during training. The training process spanned multiple epochs, and the progression of training and validation losses was monitored to assess the model's convergence.

To manage the entire training procedure, we utilized the Trainer class. This class streamlined various tasks like data loading, optimization, and evaluation. Once training concluded, the model's performance was evaluated using a dedicated test dataset to gauge its proficiency in real-world sentiment analysis scenarios.

3.3 Evaluation

The evaluation of our model was carried out in two distinct stages to assess its performance comprehensively. In the first step, a preliminary evaluation was conducted using a test dataset derived from a subsection (0.2) of the original training data as part of the training process. This step enabled us to understand the model's initial predictive capabilities and robustness.

The second step involved a more extensive evaluation by employing a separate unimodal movie reviews dataset. This dataset exclusively contained textual reviews or excerpts relating to movies, each accompanied by an associated sentiment score. This approach allowed us to isolate the model's performance within a unimodal context, focusing solely on textual data. By applying the model to this specific dataset, we gained valuable insights into its ability to comprehend sentiment patterns in a text-centric scenario, thereby providing a deeper understanding of its proficiency in sentiment analysis tasks.

3.4 Experiments

Our experimental approach encompassed two crucial phases. Initially, we conducted a hyper-parameter tuning process to determine the optimal lambda values. By systematically varying these lambda values, we identified the configuration that yielded the lowest global validation loss, refining the model's performance.

Subsequently, we engaged in training two distinct models for performance comparison. The first model, CycBERT, was subjected to the selected lambda values from the tuning phase. This tailored model's performance was then evaluated over the unimodal movie reviews dataset. In contrast, the second model represented our baseline approach, where lambda values were set to zero. This baseline model was, in essence, a BERT model that underwent continuous pre-training exclusively over the CMU-MOSI dataset. Notably, the baseline model did not incorporate translation loss considerations during pre-training.

The comparative analysis of CycBERT against the baseline BERT model enabled us to assess the effectiveness of our approach. By evaluating both models over the unimodal movie reviews dataset, we aimed to unravel the impact of our proposed modifications, particularly the integration of translation loss, on the sentiment analysis task's performance.

Model	λ_t	λ_c	$\downarrow$ MSE	$\uparrow$ R^2
CycBERT	0.5	0.5	**1.051824**	**0.414871**
Baseline	0	0	1.139170	0.366281

Table 1: In direct comparison, our CycBERT model using a balanced loss function outperformed the baseline model (BERT) over the test movie reviews dataset. Notably, non-trivial CycBERT achieved better scores in terms of both Mean Squared Error (MSE) and R^2 metrics. This outcome highlights the potential of our approach in enhancing sentiment analysis accuracy.

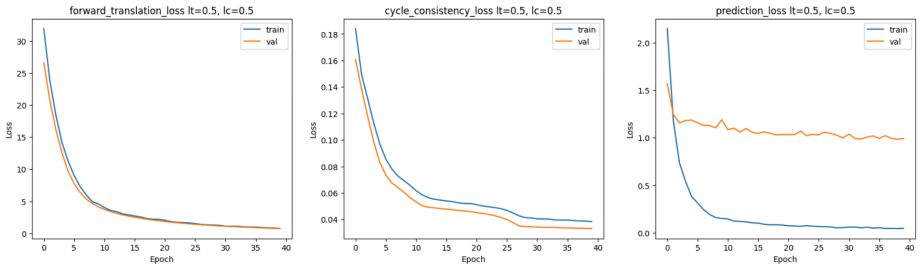

Figure 4: CycBERT training losses plot of the optimal hyper-parameters setup $\lambda_t = 0.5$ and $\lambda_c = 0.5$. The validation loss is evidently improved across epochs, and so is sentiment prediction (although in a shallower manner

4 Discussion

Our study offers an interesting stride forward in tackling the challenge of enriching large text encoders with non-language modalities. By introducing the MCTN architecture and incorporating cyclic translation loss, we have demonstrated the efficacy of combining textual and visual cues for sentiment analysis. The cyclic translation loss emerges as an effective mechanism for enhancing sentiment understanding by bridging the gap between diverse modalities, enabling the model to capture nuanced and balanced expressions of sentiment.

As we look to the future, there are several exciting routes for exploration. Expanding the scope of modalities to include acoustic cues can deepen the model's understanding, particularly in scenarios where tone and intonation hold significance. Amplifying the training process by leveraging more powerful computational resources can potentially unlock further performance improvements. Integrating additional multimodal datasets into the evaluation process can provide a more comprehensive understanding of the model's adaptability. Additionally, broadening the analysis to encompass tasks beyond sentiment analysis, such as stress detection and other mental illness-related symptoms, holds the potential to illuminate the broader spectrum of applications that can benefit from the incorporation of non-verbal communication cues. Collectively, these directions underscore the ongoing evolution of multimodal techniques and their pivotal role in enhancing the depth and accuracy of large text encoders.

5 Code

Full experimentation code and results can be found at:
https://colab.research.google.com/drive/1CoxBF3mnT0qW-7M430yZGNy4wgtWhEmr?usp=sharing

References

[1] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabas Poczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[3] Bijoyan Das and Sarit Chakraborty. An improved text sentiment classification model using tf-idf and next word negation. 2018.

[4] Rungroj Maipradit, Hideaki Hata, and Kenichi Matsumoto. Sentiment classification using n-gram idf and automated machine learning. 2019.

[5] Chetan Kaushik and Atul Mishra. A scalable, lexicon based technique for sentiment analysis. 2014.

[6] Anuj sharma and Shubhamoy Dey. Performance investigation of feature selection methods. 2013.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013.

[8] Robin Brochier, Adrien Guille, and Julien Velcin. Global vectors for node representations. In *The World Wide Web Conference*. ACM, may 2019.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[10] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. 2017.

[11] Seungwhan Moon, Suyoun Kim, and Haohan Wang. Multimodal transfer deep learning with applications in audio-visual recognition, 2016.

[12] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks, 2017.

[13] Qiao Han, Jun Zhao, and Kwok-Yan Lam. Facial landmark predictions with applications to metaverse, 2022.

[14] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016.