Analyzing Sentiments using Blip2 for Frame-to-Text Conversion and Deep Learning Classifiers*

Adithya Sharma C A
Computer Science and Engineering
PES University, RR campus
Bengaluru, India
pes1202100718@pesu.pes.edu

Bontha Srinivasa Reddy
Computer Science and Engineering
PES University, RR campus
Bengaluru, India
Srinivasareddy2411@gmail.com

Vunnam Sai Koushik
Computer Science and Engineering
PES University, RR campus
City, Country
pes1202101344@pesu.pes.edu

Ajey Bhat
Computer Science and Engineering
PES University, RR campus
Bengaluru, India
pes1202101344@pesu.pes.edu

Dr. Uma D
dept. Computer Science and Engineering
PES University, RR campus
City, Country
email address or ORCID

***Abstract*—This paper presents a method for sentiment analysis on GIFs and short form videos using vision transformers. With the proliferation of short visual content, automated sentiment analysis becomes crucial. Our approach combines vision transformers based pretrained models like Blip2 with sentiment analysis models like BERT, LSTM, and BiLSTM to extract spatial information and discern emotional tone. We demonstrate our method's efficacy through experiments on diverse datasets, showing promising results for sentiment analysis in short visual content.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In the era of short form communication dominated by GIFs and short videos, understanding sentiment poses a challenge. This paper presents a novel approach using vision transformers and sentiment analysis models like BERT, LSTM, and BiLSTM to effectively analyze sentiment in such content.

II. Abbreviations and Acronyms

The following are the abbreviations used in the paper: BLIP2, is a Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, BERT: Bidirectional Encoder Representations from Transformers , LSTM: Long Short-Term Memory, BiLSTM: Bidirectional Long Short-Term Memory, OpenCV: Open Source Computer Vision Library. VQA: Visual Question answering system.

III. Data

A. Tumbler GIF data

The dataset being used is carefully curated to suit our needs. We have made use of the T-GIF dataset for the sentiment analysis on it. The TGIF dataset contains over 100,000 samples of GIFs with a corresponding sentence labelling done. This allows us to use a variety of transformer based approaches for generating a sentence for a given GIF.

Fig. 1. Figure taken directly from the TGIF repository

B. Twitter data

The Sentiment140 dataset, which has been meticulously gathered and annotated for sentiment analysis purposes, was used for this project. 1,600,000 tweets were taken out of it using the Twitter API. Every tweet in the collection has a polarity label, with 0 denoting a negative sentiment and 4 a positive sentiment.

- **target**: The tweet's polarity (0 = negative, 2 = neutral, and 4 = positive).
- **ids**: The tweet's ID (for example, 2087).
- **date**: The tweet's date (e.g., Saturday, May 16, 2009, 23:58:44 UTC). **flag**
- : The tweet's query was utilized. NO_QUERY is the value that is set if no query was used.
- **user**: The user's username (robotickilldozr, for example) that tweeted the message.
- **text**: The tweet's actual text (e.g., "Lyx is cool").
  An excellent resource for developing and assessing sentiment analysis models is this dataset. Researchers and practitioners can use this dataset to create and evaluate

machine learning algorithms for tasks involving sentiment classification.

IV. Pre-trained Models

The pre-trained model called Blip2 was a vision transformer-based model developed by the research wing at sales force. It uses a pre-trained image encoder and an LLM by training a lightweight, 12-layer Transformer encoder i between them, with an appreciable performance on variou vision-language tasks. This pretrained model can be used as question answering system on the image loaded. The followin section discusses the architecture.

A. BLIP2: Architecture

- Blip2 uses an image encoder with a query transforme which also implements a query queue. It also has an LL which gives more flexibility to the image encoder mode part.
  The following image represents the entire architecture.

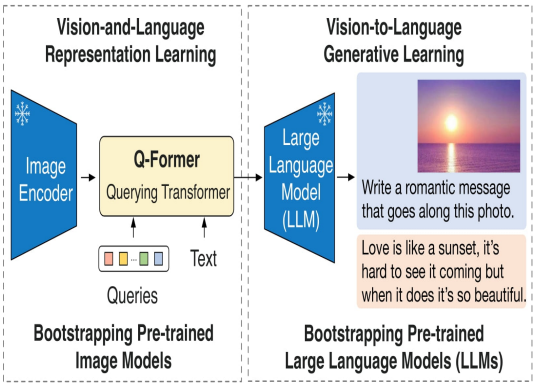

Fig. 2. Figure taken directly from the Blip2 original author

- The paper makes use of a pre-trained hugging face model.
- The Blip2 model is built on top of the pre-existing Blip model.
- The authors of Blip2 intricately describe their model as follows, "outperforms Flamingo80B by 8.7 percent on zero-shot VQAv2 with 54x fewer trainable parameters. We also demonstrate the model's emerging capabilities of zero-shot image-to-text generation that can follow natural language instructions."

B. BERT: Architecture

- BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art natural language processing (NLP) model developed by Google. It belongs to the Transformer architecture family, which utilizes self-attention mechanisms to capture relationships between words in a sentence.
- The architecture of BERT consists of multiple transformer layers, which enable bidirectional context understanding by processing the input text in both directions simultaneously.
- BERT is pre-trained on large text corpora using unsupervised learning tasks such as Masked Language Model (MLM) and Next Sentence Prediction (NSP).This pre-training process enables BERT to learn rich representations of language that can be fine-tuned for specific downstream tasks.

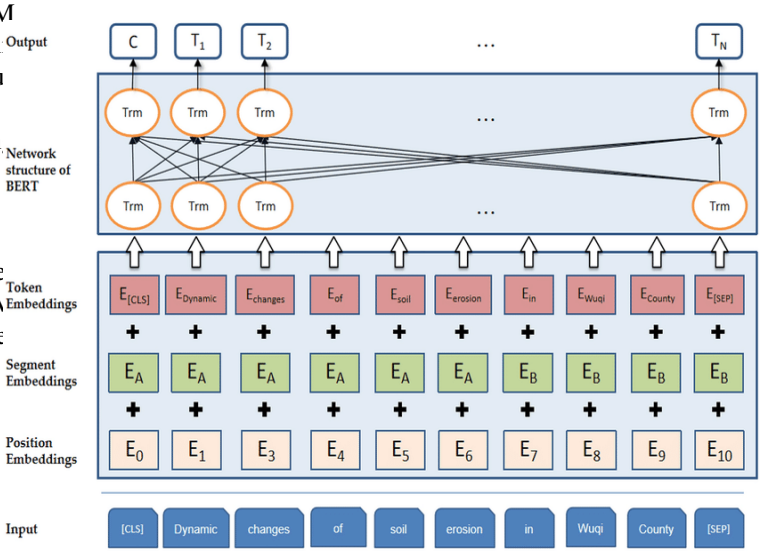

Fig. 3. The overall structure of the BERT model

- Tokenizing input text, where each word is changed into its associated token using the BERT tokenizer, is a necessary step in BERT's input representation. Next, these tokens are included into vectors with high dimensions.
- In pre-training, BERT picks up on contextual cues and word associations by learning to anticipate masked words within sentences and whether two sentences will follow one another.
- By adding task-specific layers on top of the pre-trained model and training on task-specific data, BERT can be refined for certain tasks after pre-training, such as sentiment analysis.
- Contextualized word embeddings are BERT's output, which gives tasks farther down the line useful representations of the input text.
- Because of its efficient pre-training goals and bidirectional context awareness, BERT performs exceptionally well in a variety of NLP benchmarks, including text classification, question answering, and sentiment analysis.

C. Bi-LSTM

- Bi-LSTM(Bidirectional Long Short-term Memory) is an extension of traditional LSTM architecture, which is designed to capture previous and future states of sequence.
- The architecture of BiLSTM consists of 3 layers: Embedding layer - This layer will converts input sequences into dense vector with fixed size, Bidirectional LSTM layer- This layer includes both forward and backward LSTM layers which will processes sequence in forward and backward direction, Dense layer - This is the output

layer with softmax activation function which will produce probabality distribution of 3 classes and gives output.

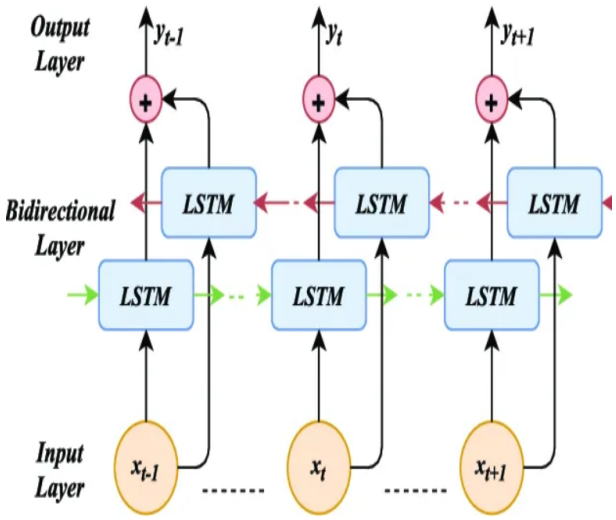

Fig. 4. Architecture of BiLSTM

- The BiLSTM generate contextual embedding for each tokenns in sentence which is encoded with future and past contexts information.This helps in capture tokens relation with complete sentence.Overall,using bidirectional processing and capturing comprihencide contextual information BiLSTM can provide better accurecy than unidirectional LSTM in sentiment analysis.

V. Methodology

We propose to make use of the Blip2 pretrained model, which is available on hugging face for captioning the frame wise split of the images. We make use of a custom frame split function to convert the GIF into user defined number of frames. These frames are then captioned using the Blip2 model and a narration of the frames is stored individually per frame. We then run these generated sentences through a sentiment anlyser as discussed above to find the final sentiment score. There are various sentiment analysers that we are using, like BERT, BiLSTM, LSTM etc. These can also be used in combination to get a final sentiment of the GIF.

VI. Results

Using the BERT architecture, the sentiment analysis model performed exceptionally well, achieving an accuracy of 84.58% on the test dataset. For both negative and positive sentiment classes, precision, recall, and F1-scores were excellent: precision was 0.84 and 0.85, recall was 0.86 and 0.83, and F1-score was 0.85 and 0.84, respectively. The weighted-average and macro-average F1-scores were both close to 0.85, indicating a balanced performance across sentiment classes. In contrast, alternative models such as LSTM and BiLSTM achieved lower accuracies, approximately 77.8% and 78%, respectively, on the same test dataset. This result underscores the superior performance of the BERT model in sentiment analysis tasks, showcasing its ability to capture subtle sentiment nuances effectively.

A. Conclusion

The study presented a comprehensive approach for sentiment analysis on GIF content, leveraging cutting-edge deep learning models such as BERT. Through an ensemble technique combining BERT for sentiment analysis and BLIP2 for image captioning, the sentiment expressed in GIFs was effectively captured.

The results underscore the superior accuracy and robustness of the BERT model compared to traditional sequential models like LSTM and BiLSTM in sentiment analysis tasks. Additionally, the combination of BLIP2 and BERT models enhanced overall performance by leveraging both textual and visual information from GIF content.

This research highlights the importance of utilizing state-of-the-art deep learning models and ensemble techniques for complex multimedia analysis tasks like sentiment analysis on GIFs. The proposed model holds promise for applications in various domains requiring multimedia content sentiment analysis, including content recommendation systems, social media monitoring, and customer feedback analysis. Further exploration could focus on integrating additional modalities or optimizing the ensemble approach for specific domains to enhance sentiment analysis accuracy.

References

[1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).Batra, Himanshu, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal.

[2] "Bert-based sentiment analysis: A software engineering perspective." In Database and Expert Systems Applications: 32nd International Conference, DEXA 2021, Virtual Event, September 27–30, 2021, Proceedings, Part I 32, pp. 138-148. Springer International Publishing, 2021.

[3] S. Ramakrishnan and L. D. Dhinesh Babu, "Enhancing Twitter Sentiment Analysis using Attention-based BiLSTM and BERT Embedding," 2023 9th International Conference on Smart Computing and Communications (ICSCC), Kochi, Kerala, India, 2023, pp. 36-40, doi: 10.1109/ICSCC59169.2023.10335010.

[4] BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi

[5] Shirzad, Amirhossein Zare, Hadi Teimouri, Mehdi. (2020). Deep Learning approach for text, image, and GIF multimodal sentiment analysis. 419-424. 10.1109/ICCKE50421.2020.9303676.