Are Captions All You Need?
Investigating Image Captions for Multimodal Tasks

Stanford CS224N Custom Project

Victor de Souza Lima e Silva
Stanford University
victsou@stanford.edu

Abstract

In today's digital world, it is increasingly common for information to be multimodal: images or videos often accompany text. Sophisticated multimodal architectures such as ViLBERT and VisualBERT have achieved state-of-the-art performance in vision-and-language tasks. However, existing vision models cannot represent contextual information and semantics of images like transformer-based language models can. Fusing the semantic-rich information coming from text becomes a challenge. In this work, we study the alternative of first transforming images into text using image captioning. We then use transformer-based methods to combine the two modalities in a simple but effective way. We perform an empirical analysis on different multimodal tasks, describing the proposed method's benefits, limitations, and the situations where this simple approach can replace large and expensive handcrafted multimodal models.

1 Key Information to Include

- Mentor & External Collaborator: Leonardo Neves[1]

[1] Snap Research

2 Introduction

The power of BERT [1] and BERT-style models has played a big part in pushing the boundaries of natural language processing (NLP). State-of-the-art models for a wide range of linguistic tasks can be created after simply fine-tuning a pretrained BERT model. That level of success along with the growing importance of multimodal tasks has inspired frontier work to align and fuse vision and language models in sophisticated ways such as done by ViLBERT [2] and VisualBERT [3].

In this project, we perform empirical analysis on a different approach to represent images in multimodal tasks, originally proposed in [4]. With BERT's linguistic prowess in mind, images are brought to the language realm via image captioning, and are then concatenated to the original texts of the multimodal tasks, only to finally be given to an unimodal BERT model as text-only input.

We report results on two datasets: **CrisisMMD: Multimodal Crisis Dataset** [5] and **The Hateful Memes Dataset** [6]. We find that our simple approach performs competitively and can provide a quick and strong baseline for multimodal tasks.

3 Related Work

The extensive success of Transformers [7] and BERT [1] inspired the development of high-performance multimodal models with BERT-like architectures. VisualBERT [3] consists of a stack of Transformer layers that implicitly align elements of input text with regions in an associated

input image, using self-attention. ViLBERT [2] similarly proposes to learn joint representations of images and language, but with two separate Transformers for vision and language interacting through co-attentional layers.

[8] presents a different perspective toward to vision-and-language models, by introducing a multimodal bitransformer that jointly finetunes unimodally pretrained text and image encoders. Unimodally pretrained models are simpler and easier to adapt without requiring multimodal retraining. For example, it is straightforward to replace an image encoder with a newly developed better alternative.

Our approach is a direct extension of the work done in [4]. The authors investigate how to take advantage of information from images in Multimodal Named Entity Recognition (MNER), based on speculation that semantic-understanding models such as image captioning may provide better image representations for the task than image classification customarily does. Their analysis of different vision-and-language fusion techniques culminates in the alternative approach of using captions to represent images as text, which we replicate to other datasets and tasks in order to verify how well it can be transferred. With this strategy, our model is unimodally pretrained and also unimodally finetuned.

4 Approach

Our image preprocessing step transforms them into captions using BUTD [9]. That is done using a BUTD implementation provided by Facebook AI Research's MMF framework [10]. Ready-to-use caption generation code was generously provided by the authors of [4].

Captions for each example are then concatenated with the corresponding text. In cases where the lengths of the text and the caption sum up to more than the maximum length that our model can support, the text component is truncated.

For all text components of our work, we used the Transformers [11] library. Preprocessing was done with off-the-shelf code that lowercases text and tokenizes it using WordPiece [12]. An off-the-shelf uncased BERT [1] base model is used, conveniently pretrained on a large corpus comprising the Toronto Book Corpus and Wikipedia. Pretraining was done using a combination of masked language modeling and next sentence prediction.

4.1 CrisisMMD: Multimodal Crisis Dataset

4.1.1 Baselines

Scores from [13] were used as baselines. Their multimodal model concatenates outputs from a VGG16 network (Image modality) and a Convolutional Neural Network (Text modality) in order to build a shared representation that is then followed by a dense layer and finally a softmax. Table 1 presents their results.

Table 1: Baseline Results for CrisisMMD

Task	F1-score
Informativeness Classification	0.842
Humanitarian Classification	0.783

4.1.2 Architecture & Implementation Details

For this dataset, our BERT model has a linear layer on top of the pooled output for classification, and we fine-tune it in each experiment for the task at hand (Informativeness vs Humanitarian classification).

4.2 The Hateful Memes Dataset

4.2.1 Baselines

Scores from [6] were used as baselines. Table 2 presents their results, which include versions of ViLBERT and VisualBERT that were only unimodally pretrained, and multimodally pretrained versions: **ViLBERT CC**, trained on Conceptual Captions [14], and **VisualBERT COCO**, trained on COCO [15].

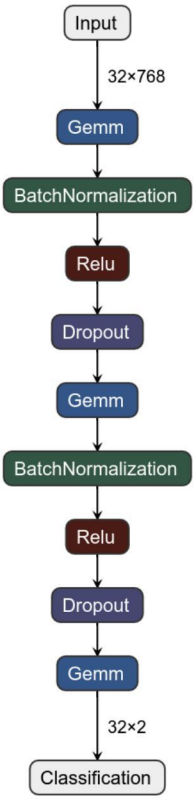
Table 2: Baseline Results for The Hateful Memes Dataset

Type	Model	Validation		Test	
		Acc.	AUROC	Acc.	AUROC
	Human	-	-	84.70	82.65
Unimodal	**Text BERT**	58.26	64.65	59.20 ± 1.00	65.08 ± 0.87
Multimodal with Unimodal Pretraining	**ViLBERT**	62.20	71.13	62.30 ± 0.46	70.45 ± 1.16
	VisualBERT	62.10	70.60	63.20 ± 1.06	**71.33 ± 1.10**
Multimodal with Multimodal Pretraining	**ViLBERT CC**	61.40	70.07	61.10 ± 1.56	70.03 ± 1.77
	VisualBERT COCO	**65.06**	**73.97**	**64.73 ± 0.50**	71.41 ± 0.46

4.2.2 Architecture & Implementation Details

For this dataset, our BERT model has a Multi-Layer Perceptron (MLP) on top of the pooled output for classification, reusing the architecture of the provided Text BERT baseline. It has two repeated blocks of (Linear, BatchNorm1D, ReLU, Dropout) layers, followed by a final linear layer with two outputs for classification, as illustrated in Figure 1. The dropout value in both layers is set to 0.5.

Figure 1: MLP Classifier Architecture



The simplicity of our approach allowed us to use off-the-shelf starter code for the Hateful Memes challenge, available in the MMF [10] library, whose configuration was already appropriate for the task at hand. Our own implementation work consisted solely on small adaptations in order to use our modified caption+text dataset.

5 Experiments

5.1 CrisisMMD: Multimodal Crisis Dataset

5.1.1 Data

The **CrisisMMD: Multimodal Crisis Dataset** [5] consists of several thousands of manually annotated tweets and images (≈1.8GB) collected during seven major natural disasters including earthquakes, hurricanes, wildfires, and floods that happened in the year 2017 across different parts of the World. We are working on two of its tasks:

1. Classify multimodal tweets as Informative or Not Informative. (**Informativeness classification**)
2. Classify multimodal tweets into one of ["Affected individuals", "Infrastructure and utility damage", "Injured or dead people", "Missing or found people", "Rescue, volunteering or donation effort", "Vehicle damage", "Other relevant information"]. (**Humanitarian classification**)

5.1.2 Evaluation Method

F1-score was used as the evaluation metric for both tasks.

5.1.3 Experiments

Because our baseline paper did not attempt to use BERT for classification, we performed two experiments: One that simply used the original text as input and ignored images completely, and one that followed our proposed approach of concatenating image captions and used that along with text.

In both experiments, our model was fine-tuned for 3 epochs, with a per-device batch size of 16. Our chosen optimizer was Adam [16], with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1\text{e}{-8}$, and with weight decay of 0.01 as introduced in [17]. Its learning rate was $5\text{e}{-5}$, and we used 500 steps of linear warmup, meaning that our learning rate began as 0 and took 500 steps to reach its target value. No layers of our model were frozen during fine-tuning.

5.1.4 Results

Table 3: CrisisMMD Results

Task	Model	F1-score
Informativeness Classification	**CNN + VGG16 (Baseline)**	0.84
	BERT Without Captions	0.86
	BERT With Captions	**0.90**
Humanitarian Classification	**CNN + VGG16 (Baseline)**	0.78
	BERT Without Captions	0.82
	BERT With Captions	**0.86**

Our results were better than expected and greatly surpassed our original baselines, largely due to BERT's proficiency in textual tasks. BERT without captions on its own outperformed the proposed multimodal approach. Our approach contributed to a F1-score increase of 0.04 in both tasks, showing that captions gave BERT useful information about their corresponding images.

5.2 The Hateful Memes Dataset

5.2.1 Data

The Hateful Memes Challenge Dataset [6] is a multimodal dataset for hateful meme detection: Given a meme image file, and a string representing the text in the meme image, classify the meme as hateful or not hateful. It consists of 4.2GB newly created examples, made with a custom-built tool.

5.2.2 Evaluation Method

Area Under the Receiver Operating Characteristic curve (AUROC) is used as the main evaluation metric. Since the dataset is balanced, accuracy is also secondarily reported as an extra, easier-to-interpret signal.

5.2.3 Experiments

Our model was trained for 10000 steps, with a per-device batch size of 32. Our chosen optimizer was Adam [16], with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 1\text{e}{-8}$, and with weight decay [17] of 0.01. We used a cosine learning rate scheduler with 2000 steps of linear warmup, and our peak learning rate was $5\text{e}{-5}$. No layers were frozen, and training occurred end-to-end.

5.2.4 Results

Table 4: The Hateful Memes Dataset Results

Model	**Validation**		**Test**	
	Acc.	AUROC	Acc.	AUROC
BERT with Captions	63.60	69.28	63.00	69.51

Our results were within the expected range, with significant accuracy and AUROC gains over Text-only BERT, but with AUROC slightly below the baselines for **ViLBERT**, **VisualBERT** and **ViLBERT CC**, and Accuracy slightly above them. **VisualBERT COCO** outperforms our approach with a comfortable margin. These are not discouraging results, as our approach presents a simpler alternative to larger models that require multimodal training.

6 Analysis

6.1 CrisisMMD: Multimodal Crisis Dataset

Our model did remarkably well on both tasks of this dataset. We observed that the nature of the tweets contributed a great deal to it, as images were often related to the accompanying text, particularly for those classified as informative or into one of the humanitarian cateogires.

Error analysis was performed on 40 examples from the test set for the Informativeness Classification task, as categorized in Table 5.

Table 5: Error Analysis for Informativeness Classification Task

Error Category	**# Examples**
Bad Caption	16
Misleading Text	12
Mislabeled	4
Uncategorized	8

Our biggest source of mistakes was the **Bad Caption** category. We observed that our captions were particularly poor for images containing text. Tweets often contain screenshots and digitally created images, and their embedded textual content was simply lost in our image captioning step. We re-ran

(a) "a wall with some writing on it and a clock"

(b) "a close up of a computer keyboard"

Figure 2: Examples of bad captions generated for CrisisMMD.

predictions with better captions and found that 15 out of the 16 examples would have been classified correctly; The remaining one required a transcription of the text contained in its image.

The **Misleading Text** category had tweets that explicitly mentioned natural disasters, but contained no other informative content. Our fine-tuned model seems to easily recognize key terms, but we would likely need more task-specific training for it to be able to correctly ignore such terms in non-informative samples.

6.2 The Hateful Memes Dataset

Table 6: Error Analysis for Hateful Meme Detection

Error Category	# Examples
Bad Caption	8
Nuanced	11
Uncategorized	11

The **Bad Caption** category was a significant source of mistakes. We observed that they were simply too general for this task, and lacked important information about human subjects such as race, gender, age, and facial expressions. It also did not capture socially meaningful symbols, commonly featured in hateful memes. Without these, captions fell short in helping the model understand the meaning of each meme. Finally, captions were also of poor quality when memes consisted of two or more collaged images: Only a single part of the meme image seemed to be captioned, and its rest was ignored.

Our analysis made it clear that this is a very hard task. Many of our model's mistakes fell into the **Nuanced** category, where the meaning of the memes was simply too nuanced, not only for a model but also for untrained humans. This explains the relatively low human accuracy reported in [6], along with moderate inter-annotator agreement.

Images from our analysis could not be reproduced in this report due to their sensitive nature.

7 Conclusion

In this work, we extrapolate findings from [4] to more datasets. Our experiments present more success cases of using image captions to represent images as text, allowing unimodal BERT models to be used in multimodal tasks. We observed that this approach is promising as a simple and efficient way of incorporating visual information into models without the need for multimodal training.

A potential avenue for future exploration is improving caption quality and possibly including Optical Character Recognition (OCR). This would have been particularly useful for the tasks of the

CrisisMMD dataset, and it is favorable because it would give our model information that other vision-and-language models probably miss.

References

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

[2] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[3] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019.

[4] Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. A caption is worth a thousand images: Investigating image captions for multimodal named entity recognition, 2020.

[5] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*, June 2018.

[6] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2020.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[8] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text, 2020.

[9] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering, 2018.

[10] Amanpreet Singh, Vedanuj GoswamiG, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`, 2020.

[11] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[12] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5149–5152, 2012.

[13] Ferda Ofli, Firoj Alam, and Muhammad Imran. Analysis of social media data using multimodal deep learning for disaster response, 2020.

[14] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.