Multimodal Learning: Are Captions All You Need?

Anonymous ACL submission

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, VisualBERT, and LXMERT have achieved state-of-the-art performance in vision-and-language tasks. However, existing vision models cannot represent contextual information and semantics 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 benefits, limitations, and situations where this simple approach can replace large and expensive handcrafted multimodal models.

1 Introduction

In recent years, BERT (Devlin et al., 2019) and Transformer-based methods have revolutionized NLP and fine-tuning these pre-trained models became the standard approach for most language tasks. However, language in isolation is not enough to perform certain tasks. For example, a social media post with the text *Perfect beach weather!* might completely change meaning if accompanied by a thunderstorm image. Also, tasks like visual question answering (Antol et al., 2015) are inherently multimodal.

Prior work has approached this challenge using multimodal learning via attention (Li et al., 2019; Lu et al.; Zhang et al., 2021) or alignment techniques (Li et al., 2020). Despite their improvements, these methods require sophisticated handcrafted architectures and introduce billions of additional parameters to the model, making it hard to disentangle the effects of increased complexity

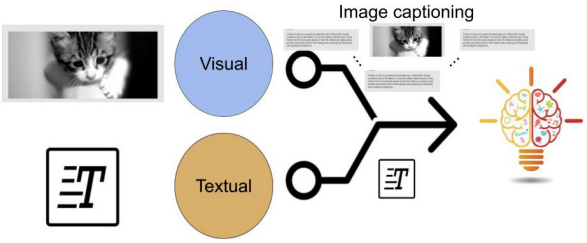

Figure 1: Multimodal learning traditionally requires a fusion stage to combine visual and textual features. This stage increases model complexity and often brings marginal gains over strong language-based baselines. Encapsulating the visual information via a caption makes the fusion step straightforward and leverages BERT's semantic knowledge.

from the benefits of combining modalities. (Chen et al., 2021) studies the impact of these methods for named entity recognition (NER) and concludes that existing fusion techniques can only bring marginal, if any, improvements to existing language-based models. Their recommendation is to use image captions (Anderson et al., 2018; Johnson et al., 2016) to represent images, making the fusion step straightforward as both representations are in the textual domain, as shown in Figure 1.

In this work, we hypothesize that their findings go beyond the NER task. We select a broader range of tasks and experiment with coarse-grained and fine-grained image caption techniques such as BUTD (Anderson et al., 2018) and DenseCap (Johnson et al., 2016). The generated captions carry semantic information that is easier to combine with textual features, as opposed to plain image features.

We perform our study on four multimodal datasets of natural disasters, social media, visual questions answering, and movie posters. We compare the image captioning method to using BERT without any image information (unimodal) and to LXMERT, a strong multimodal baseline. The method overcomes BERT(unimodal) and shows competitive results with LXMERT. We also show statistical analysis and model capacity relation with

image captions semantic knowledge, where these two components are crucial for good performance.

In summary, our contributions include:

- A study of multiple ways to fuse image and textual modalities, with a focus on incorporating image captions as a replacement for traditional image features.
- An examination of the dependency between image captioning semantic representation and model capacity.

2 Related work

2.1 Multimodal learning

Researchers are working on multimodal tasks such as image captioning (Krishna et al., 2017; Faghri et al., 2018; Aneja et al., 2019), visual question answering (Antol et al., 2015; Hudson and Manning, 2019; Goyal et al., 2017), and visual reasoning (Zellers et al., 2019).

These tasks require an understanding of textual and visual representations. Kruk et al. (2019) projects these modalities to the same space vector and learns a classification model. In addition to this intuitive approach, modern transformers models learn a joint representation of these modalities. VisualBERT (Li et al., 2019) aligns elements of text to its associated image regions using self-attention, whereas ViLBERT (Lu et al.) has separate Transformers for vision and language that only attend to each other, resulting in a much heavier and expensive model. On the other hand, LXMERT (Tan and Bansal, 2019) bridges vision and language semantics via handcrafted pre-training tasks.

In contrast to previous self-attention approaches, OSCAR (Li et al., 2020) uses object tags detected in images as anchor points for semantic alignment. VinVL (Zhang et al., 2021) extends this approach adding importance to object-centric representation using an attention mechanism.

As opposite to attention-based and alignment approaches, we encode images in a more similar representation to textual data, using image captions.

2.2 Image captioning

Image captioning is an application that combines visual and textual modalities by generating a textual description from an image (Krishna et al., 2017; Faghri et al., 2018; Aneja et al., 2019). BUTD (Anderson et al., 2018) produces captions integrating attention with feature weighting. The bottom-up attention mechanism proposes image regions, each with a corresponding feature vector, and a top-down mechanism determines feature weightings to generate a caption. DenseCap (Johnson et al., 2016) enhances BUTD predicting a set of descriptions across image regions using a localization network.

In this work, we use BUTD (Anderson et al., 2018) and DenseCap (Johnson et al., 2016) for image captioning. DenseCap generates a dense set of image descriptions to facilitate learning with their corresponding textual data.

2.3 Image encoding as captions

Chen et al. (2021) uses captions to represent images as text in Multimodal Named Entity Recognition (MNER), and argues that semantic-understanding models such as image captioning may provide better image representations.

In this work, we still represent images via captions, however, in addition to traditional image captioning, we use a set of dense captions from (Johnson et al., 2016), and extend it for multimodal classification tasks on social media, visual question answering, and movies domains.

3 Experimental setup

3.1 Image captioning encoders

We represent visual data via image captions with the following approaches.

- **BERT + capts.** Captions are generated using BUTD (Anderson et al., 2018) and concatenated to input text before being fed into the same BERT-based architectures. This baseline asses how caption type and quality influence performance. BUTD's drawback is that it creates coarse-grained semantic text, and may lose some details. Thus, we capture fine-grained and diverse captions with BERT + dense_capts.
- **BERT + den_capts.** First, we compute diverse captions (Johnson et al., 2016) from images using Visual Genome dataset (Krishna et al., 2017). We select the ten most confident captions and concatenate them with the instance correspondent text. Finally, this textual data is feed to a BERT model (Devlin et al., 2019).

3.2 Baselines

We compare our image captioning BERT models with with two different baselines:

- **BERT (unimodal)** (Devlin et al., 2019). Image components of our data are ignored, and we simply use the BERT-based architectures without any alterations to the input text. This baseline allows us to understand how much we gain by using visual information.
- **LXMERT** (Tan and Bansal, 2019). It receives an image representation and its related sentence; It produces three outputs: language representations, image representations, and cross-modality representations. We use the cross-modal representations for the classification tasks. Its performance is near to the state-of-the-art in various vision-and-language tasks.

3.3 Evaluation tasks

We use four evaluation tasks which have frequently been used for multimodal (image and text) domain: CrisisMMD (Alam et al., 2018) with informative (*CI*) and humanitarian (*CH*) classification tasks (e.g. injured or death people, rescue effort, vehicle damage) from seven major natural disasters around the world with 12'708 instances; Hateful memes (*HM*) (Kiela et al., 2020) with hateful and non hateful annotations for 10'000 instances; visual question answering (*VQ*) (Antol et al., 2015; Zhang et al., 2016) composed of 33'383 (image, question) pairs, and yes/no answers for binary classification; and movie posters (*MP*) (Cascante-Bonilla et al., 2019) composed of 5000 movies with image posters and plots categorized in action, adventure, romance, comedy, drama, fantasy, among others.

3.4 Evaluation protocol

For each dataset, we use the provided train, validation, and test splits. The only exception is for *VQ*, where we split the validation set into two halves. To increase reliability, we run five repetitions with different seeds.

We evaluate all baselines using Huging Face transformers framework (Wolf et al., 2020). We use Adam optimizer, a learning rate of 5e-5, and 30 epochs. At the end of each epoch, the network was evaluated on a validation set, and the network with a higher F1_weighed score over a validation set was selected for testing. We report F1 and accuracy metrics in our experiments.

4 Experimental results

4.1 Comparison to baselines

Tables 1 and 2 show average accuracy and weighted F-measure per dataset, respectively. In both tables, we observe that *BERT + den_capts* and *LXMERT* methods perform similarly and are the best methods among all baselines. Our *BERT + den_capts* is better for *CH* and *HM* datasets, while *LXMERT* has better performance for the remaining datasets. Also, we add a multi-label dataset in Table 3, where *BERT + den_capts* is the best performer. We believe this top performance is due to the complex dataset, where poster images do not provide too much information about movie categories, while a textual description may provide semantic details. We observe that *LXMERT* is the worst approach, while other competitors with some semantic knowledge are accurate.

Accuracy	CI	CH	VQ	HM	Avg
BERT (unimodal)	0.859	0.814	0.513	0.580	0.691
BERT + capts.	0.879	0.836	0.513	0.642	0.717
BERT + den_capts	0.887$^{\sim}$	**0.854**$^{\triangle}$	0.513$^{\triangledown}$	**0.654**$^{\sim}$	**0.727**
LXMERT	**0.892**$^{\sim}$	0.844$^{\triangledown}$	**0.518**$^{\triangle}$	0.645$^{\sim}$	0.725

Table 1: Comparative results for BERT using accuracy on Crisis Informative (CI), Crisis Humanitarian (CH), Visual Question Answering (VQ), and Hateful Memes (HM) datasets.

F1-weighted	CI	CH	VQ	HM	Avg
BERT (unimodal)	0.856	0.814	0.348	0.555	0.643
BERT + capts.	0.877	0.835	0.348	0.621	0.670
BERT + den_capts	0.885$^{\sim}$	**0.854**$^{\triangle}$	0.348$^{\triangledown}$	**0.637**$^{\sim}$	0.681
LXMERT	**0.892**$^{\sim}$	0.844$^{\triangledown}$	**0.384**$^{\triangle}$	0.636$^{\sim}$	**0.689**

Table 2: Comparative results for BERT using F1-weighted.

MP	F1 macro	F1 weighted
BERT (unimodal)	0.338	0.511
BERT + capts.	0.348	0.527
BERT + den_capts.	**0.360**$^{\triangle}$	**0.542**$^{\triangle}$
LXMERT	0.295$^{\triangledown}$	0.513$^{\triangledown}$

Table 3: Comparative results for BERT using F1 metrics for multi label movie posters dataset.

It is important to highlight that textual captions boost initial performance as shown in Tables 1, 2, and 3; where *BERT (unimodal)* and *BERT + den_capts.* improve performance over *BERT - capts.* We believe the success of *BERT + den_capts.* is due to encapsulation of visual information via

a caption semantic representation and its diverse fine-grained captions. Captions capture semantics in a more structured and meaningful representation that images alone, and also it is easy to combine with textual representation for multi modal tasks.

4.2 In-depth analysis

From our previous section, we observe that *BERT + den_capts.* and *LXMERT* show similar performance and we dig on this result via a statistical t-test on the last two rows on Tables 1, 2 and 3. We observe that there is no statistical difference ($p_{value} > 0.05$) on *CI* and *HM* tasks with symbol $\sim$. And there is statistical difference ($p_{value} < 0.05$) for the remaining three tasks. *CH* and *MP* tasks are in favor of *BERT + den_capts.* ($\triangle$), while vqa is for *LXMERT* ($\triangle$). Hence, we can sum up that *BERT + den_capts.* outperforms *LXMERT*.

4.3 Model capacity

Our previous results support that semantic representation of images improves multimodal learning, however, it is still unclear if model capacity is also a key component.

We replicate our previous experiments with distillBERT on Tables 4, 5, and 6. We observe that in average *distillBERT + capts.* is the best performer. From accuracy table, *distillBERT + capts.* outperforms other baselines for two datasets, while *distillBERT + den_capts.* outperforms for *CH* and *MP* data sets.

It is interesting to note that *distillBERT + den_capts.* outperforms for complex tasks such as movie poster classification, where poster images alone can be intriguing and have an advertising intent, than providing context about the movie category. In this case, captions facilitate suitable semantic knowledge for machine learning models (See Tables 3 and 6).

Accuracy	CI	CH	VQ	HM	Avg
dBERT (unimodal)	0.804	0.668	0.513	0.550	0.634
dBERT + capts.	**0.827**	0.725	0.513	**0.540**	**0.651**
dBERT + den_capts.	0.802	**0.769**	0.513	0.510	0.648

Table 4: Comparative results for distill-BERT using Accuracy on Crisis Informative (CI), Crisis Humanitarian (CH), Visual Question Answering (VQ), and Hateful Memes (HM) datasets.

Overall, we still observe improvement adding captions over *distillBERT (unimodal)*, however, *distillBERT + den_capts.* is not the best performer for all datasets as in the BERT model. We believe an explanation is that distillBERT does not have a good capacity to manage a diverse set of captions as opposed to BERT, which is a much deeper and complex model.

F1-weighted	CI	CH	VQ	HM	Avg
dBERT (unimodal)	0.801	0.703	0.348	**0.535**	0.597
dBERT + capts.	**0.825**	0.754	0.348	0.518	**0.611**
dBERT + den_capts.	0.775	**0.780**	0.348	0.345	0.562

Table 5: Comparative results for distill-BERT using F1-weighted.

MP	F1 macro	F1 weighted
dBERT (unimodal)	0.154	0.336
dBERT + capts.	0.152	0.340
dBERT + den_capts.	**0.170**	**0.383**

Table 6: Comparative results for distill-BERT using F1 metrics for multi label movie posters dataset.

5 Conclusion

In this work, we study the use of image captions in multimodal systems as a replacement for traditional image features. We show that this simple but effective approach can yield comparable, when not superior, results to strong multimodal baselines for most tasks. By leveraging the existing semantic knowledge from text-based models, fusing the original text to the image captions becomes trivial, reducing model complexity at the fusion stage. We also show the impact of model capacity and that the results still hold even when models are smaller. Our findings benefit practitioners, as incorporating captions as image representations might be more efficient than handcrafting complex multimodal alternatives, and help determine future directions for research in image captioning and multimodal fusion.

An interesting future direction is incorporating external knowledge for multimodal learning for subjective tasks such as ads understanding (Hussain et al., 2017; Ye and Kovashka, 2018) and personalization (Murrugarra-Llerena and Kovashka, 2019; Veit et al., 2018). These tasks are more complex and may benefit from external knowledge to disambiguate and better understand the context.

References

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natu-

ral disasters. In *International AAAI Conference on Web and Social Media (ICWSM)*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. 2019. Sequential latent spaces for modeling the intention during diverse image captioning. In *International Conference on Computer Vision (ICCV)*. IEEE.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*. IEEE.

Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. 2019. Moviescope: Large-scale analysis of movies using multiple modalities. *ArXiv*, abs/1908.03180.

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Can images help recognize entities? a study of the role of images for multimodal ner. In *Workshop on Noisy User-generated Text (W-NUT) at Empirical Methods in Natural Language Processing (EMNLP)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota. ACL.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference (BMVC)*. BMVA Press.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. ACL.

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

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *European Conference of Computer Vision (ECCV)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Nils Murrugarra-Llerena and Adriana Kovashka. 2019. Cross-modality personalization for retrieval. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Andreas Veit, Maximilian Nickel, Serge Belongie, and Laurens van der Maaten. 2018. Separating self-expression and visual content in hashtag supervision. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

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. 2020. Transformers: State-of-the-art natural language processing. In *Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*. Association for Computational Linguistics.

Keren Ye and Adriana Kovashka. 2018. Advise: Symbolism and external knowledge for decoding advertisements. In *European Conference on Computer Vision (ECCV)*. Springer.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and Yang: Balancing and answering binary visual questions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE.