Investigating Multimodal Contributions in Vision and Language Models: An Observation for Applying MM-SHAP Analysis

Yu-Chuan Cheng
Scientific Computing, Heidelberg University
yu-chuan.cheng@stud.uni-heidelberg.de

Abstract

This paper presents an inquiry into the performance-agnostic metric, MM-SHAP, designed for evaluating multimodal contributions within vision and language models and associated tasks. Building upon the insights gleaned from (Parcalabescu and Frank, 2023), the study endeavors to both replicate and expand upon its analysis by employing MM-SHAP within the context of the latest iteration of Large Language Models (LLMs). Specifically, our investigation aims to probe the MM-SHAP scores of the same model architecture but with an augmented patch size or different training dataset. Our hypothesis suggests that the MM-score will demonstrate a consistent proportionality across varying patch sizes. Additionally, this research seeks to widen its purview by applying MM-SHAP analysis to the LLaVa model. Through this exploration, we seek to uncover the model's understanding of contextual cues and its reliance on either visual or textual modalities. Finally, while this paper does not conclude the LLaVa work, it lays some effort for the last section, titled Future Work.

1 Introduction

1.1 Motivation

Initially, my interest in the intersection of computer vision and text was piqued by the Sora model, which translates text into videos. In today's landscape, large language models often integrate various modalities. Following discussions in seminars, particularly the insight gleaned from Julius Steen that "There is no golden metric, there is only a metric highly related to the purpose," I found myself drawn to the MM-SHAP evaluation method. This approach introduces a fresh metric that could provide preliminary insights transcending mere interactions between visual and textual models. Its potential extends to evaluating a broader range of model types in the future. Furthermore, the Clip model highlighted in literature showcases the symbiosis between visual and language models. Consequently, I am compelled to delve deeper into its application specifics and the influence of model size. This endeavor aligns with the overarching objective of this paper: to investigate the potential of MM-SHAP analysis for probing multimodal contributions in vision and language models.

1.2 MM-SHAP

This paper (Parcalabescu and Frank, 2023) proposes MM-SHAP, a novel unbiased metric for measuring the degree of contribution from different modalities (vision and language) in multimodal models. It highlights the limitations of existing accuracy-based evaluation methods, which fail to accurately assess a model's reliance on each modality. MM-SHAP is grounded in the theoretical foundation of Shapley values(a game-theoretic concept), a principled approach for fairly distributing a model's prediction contributions.

Additionally, the paper introduces how MM-SHAP can be used to evaluate and interpret modality contributions at the sample level, dataset level, model level , and measuring fine-tuning effects. It explains the computation of MM-SHAP, which involves approximating Shapley values through Monte Carlo sampling of token coalitions, and aggregating token-level contributions to obtain modality-wise scores.

Experiments are conducted with various vision-language models, including LXMERT, CLIP, and ALBEF variants, on tasks such as image-sentence alignment, visual question answering, and the VALSE benchmark. The results demonstrate the effectiveness of MM-SHAP in quantifying the extent to which models rely on different modalities.

One of the limitations mentioned is the computational cost associated with approximating Shapley values, which grows exponentially with the number of input tokens. The authors suggest increasing the number of sampled coalitions for more precise measurements at the expense of higher environmental impact. Additionally, the work focused on evaluating a limited number of English vision-language models in a zero-shot setting, leaving room for future exploration of autoregressive models, models with additional modalities, and tracking MM-SHAP scores during pretraining and fine-tuning.

1.2.1 Properties of MM-SHAP

- **Performance-Agnostic:** Unlike accuracy-based metrics, MM-SHAP does not rely on the model's task performance or prediction correctness. It solely measures the relative contribution of each modality (e.g. vision, text) to the model's output, regardless of whether the prediction is right or wrong.

- **Interpretability:** MM-SHAP provides insights into how much each input token (e.g. word, image patch) contributes towards the model's prediction, allowing fine-grained interpretation at the token level. (From (Parcalabescu and Frank, 2023) abstract, section 3.5 Fig. 1 and Section 4.4 App D.2 and Fig. 10)

- **Multi-Level Analysis:**MM-SHAP enables analysis at the sample level (individual predictions), dataset level (by aggregating over samples), and model level (comparing different models).

- **Additivity:** The additive property of Shapley values allows MM-SHAP scores to be averaged over instances and datasets, providing a holistic measure of a model's modality reliance. (From (Parcalabescu and Frank, 2023) Section 3.5)

- **Stability:** The paper shows that MM-SHAP remains relatively stable even when data distributions or model accuracies vary, allowing reliable measurement of multimodal contribution. (From (Parcalabescu and Frank, 2023) Section 4.4)

In essence, MM-SHAP inherits the desirable properties of Shapley values while extending them to quantify the degree of multimodal integration in a theoretically grounded and performance-agnostic manner, enabling comprehensive analysis and interpretation of multimodal models.

1.3 Contrastive Language-Image Pre-training (CLIP) model (Radford et al., 2021)

In this analysis, we will focus specifically on characterizing the key capabilities and limitations of CLIP, rather than providing an overly broad treatment of its training data, objectives, or other background details. By directly enumerating the salient characteristics and shortcomings observed, we can establish a factual basis to adequately ground the subsequent discussion and conclusions. Contextual factors outside the scope of CLIP's core representational properties and performance will not be examined in depth, with the intent of maintaining a precise and concentrated analysis.

1.3.1 Properties of CLIP

- **Natural language supervision:** It is a scalable method for learning visual representations directly from natural language supervision, leveraging large datasets of (image, text) pairs collected from the internet.

- **Contrastive pre-training:** It is trained on a simple contrastive objective - predicting which text caption goes with which image from a batch. This contrastive pre-training is more efficient than methods that try to predict the exact words describing an image.

- **Zero-shot transfer:** After pre-training, CLIP can perform zero-shot transfer to downstream tasks by using the names or descriptions of the target dataset classes to generate a classifier in the learned joint embedding space. I have several examples here: (Patashnik et al., 2021), (Frans et al.), (Gu et al., 2022).

- **Strong zero-shot performance:** CLIP demonstrates strong zero-shot performance on a wide variety of existing computer vision datasets, spanning object classification, OCR, action recognition, geo-localization and more. On many datasets, its zero-shot performance rivals or exceeds fully supervised baselines.

- **Robust representations:** CLIP's representations learned from natural language supervision are more robust and generalize better to distribution shifts compared to standard supervised ImageNet models.

- **Predictable scaling:** CLIP scales smoothly and predictably - increasing its size and compute used during pre-training leads to better zero-shot performance following predictable scaling laws.

In summary, CLIP is an efficient way to learn high-performing, robust and flexible visual representations from the abundant supervision contained in (image, text) pairs on the internet.

1.3.2 Limitations of CLIP

- **Subpar transfer performance:** On datasets with training splits, the zero-shot performance of CLIP is on average competitive with a simple supervised baseline of a linear classifier on top of ResNet-50 features, but this baseline is now below the overall state-of-the-art on most datasets. Significant work is still needed to improve CLIP's task learning and transfer capabilities.

- **Weak on fine-grained/systematic tasks:** CLIP's zero-shot performance is quite weak on several kinds of tasks like fine-grained classification (e.g. differentiating car models, flower species), counting objects, and novel tasks unlikely to be in the pretraining data.

- **Poor out-of-distribution generalization:** While CLIP generalizes well to many natural image distributions, it still generalizes poorly to truly out-of-distribution data, as demonstrated by its relatively poor performance on MNIST handwritten digits.

- **Lack of open-ended generation:** CLIP is limited to the concepts present in its training data for zero-shot classification, lacking the flexibility of approaches like image captioning to generate truly novel outputs.

- **Data inefficiency:** CLIP does not address the poor data efficiency of deep learning models, instead compensating by using a huge scale of supervision.

In summary, the main limitations revolve around CLIP's still limited capabilities on many tasks compared to supervised approaches, poor out-of-distribution generalization, inflexibility of text prompts and lack of data efficiency challenges.

2 Experiment

The experimentation scope was delimited by the finite computing resources at our disposal. Consequently, emphasis was directed towards the examination of the CLIP model, renowned for its efficacy in transfer learning paradigms. Within the Computational Linguistics Department at Heidelberg University, our computational arsenal comprised four Nvidia GeForce GTX 1080 Ti graphics cards, totaling 10GB of GPU RAM. Adhering to the foundational architecture of the CLIP model as outlined in (Parcalabescu and Frank, 2023), three discrete experiments were roughly conducted. These experimental configurations deviated from the canonical CLIP implementation by introducing alterations in model specifications, image inputs, and patch size parameters. Despite the inherent computational limitations, these preliminary experiments were undertaken with the overarching goal of scrutinizing the potential ramifications of modifying fundamental aspects of the CLIP model on its performance metrics and behavioral attributes.

The task I set is image sentence alignment and the dataset I use is Visual7W (Zhu et al., 2016) for one of the metrics within the paper, so called "existence"

The experimental framework was initially predicated upon the assessment of general machine knowledge performance. Traditionally, an anticipation exists that diminishing patch size coupled with augmented module dimensions should yield superior outcomes. This conjecture stems from the prevalent adoption of the top-1 performance metric, universally acknowledged as a benchmark for accuracy evaluation. It is commonly presumed that finer granularity in patching and increased complexity in module size contribute positively to the overall efficacy of the model, as corroborated by widespread acceptance of this metric as a standard for gauging accuracy.

3 Evaluation

It is unsurprising that with increasing model size, the evaluation time exhibits a non-proportional, near-exponential growth pattern. This observation underscores the significant impact of both larger dataset image sizes and smaller patch sizes on elongating evaluation times.

Of particular interest is the finding that larger models do not uniformly improve the model's Acc_r. Notably, both large-path14-336 and large-patch14 models demonstrate lower Acc_r values compared to base-patch32. However, large-path14-336 showcases markedly higher $T - SHAP_c$ values relative to other modules, suggesting a greater reliance on textual rather than visual models for this task. Ideally, a balanced contribution from both types of models would be expected in scenarios where a large model incorporates both.

"In addition, $T - SHAP_f$ demonstrates superior performance as expected in larger models, but under the training of 'laion,' its performance is mediocre at best, and even extremely poor in terms of Acc_r.

Finally, discussing possible reasons, there are two. Firstly, the clip training data consists of only 400 million Image-Text pairs from the internet. However, I reasonably speculate that OpenAI must have invested considerable time in data cleaning and labeling. This is because the open-source model 'laion' utilizes data directly scraped from the web, without filtering inappropriate content. Despite laion having a dataset of 2.3 billion English data, the training results still vary significantly. Secondly, it's possible that there is a higher degree of overlap or similarity between the dataset used in the task and the dataset originally used to train Clip. Therefore, the original Clip model may perform well."

4 Conclusion

In this paper, we explored the application of the recently proposed MM-SHAP metric to investigate multimodal contributions in vision and language models. Motivated by the potential of MM-SHAP to provide performance-agnostic insights into how models leverage different modalities, we conducted experiments focusing on the CLIP model architecture.

Our analysis involved modifying key parameters such as patch size, model size, and training dataset to observe their impact on MM-SHAP scores as well as traditional evaluation metrics. Through these experiments, we gained valuable insights into the behavior and characteristics of CLIP under varying configurations.

While our findings shed light on some intriguing patterns, such as the influence of data quality and the potential trade-offs between model size and multimodal balance, several questions remain unanswered. Further research is warranted to establish more definitive relationships between architectural choices, training procedures, and the resulting multimodal dynamics as measured by MM-SHAP.

Additionally, the computational challenges associated with approximating Shapley values, especially for larger models and finer-grained inputs, demand the exploration of more efficient approximation techniques. Overcoming these computational bottlenecks could pave the way for more extensive and high-fidelity MM-SHAP analyses across a broader range of multimodal architectures and tasks.

Ultimately, our work serves as an initial step towards leveraging the potential of MM-SHAP for a deeper understanding of multimodal models. By continuing to refine and apply this performance-agnostic metric, we can gain valuable insights into the intricate interplay between different modalities, fostering the development of more robust and interpretable multimodal systems.

5 Future Work

Initially, the experimental design aimed to comprehensively evaluate the latest LLaVA model. However, the observed results from the execution appeared significantly incongruous. Therefore, I present here a list of parameters that I believe require attention when transitioning to non-CLIP models: image_size, vocab_size, patch_size, and possibly other unnoticed issues. For those interested in tracking this issue further, please refer to the GitHub issue [1] link provided. Additionally, as suggested by the original author, exploring the evaluation of more than two types of models using the same MM-SHAP methodology is also feasible.

6 Use of Chatgpt

I begin by articulating the contents of my article in English, encompassing motivation, model features, experimental design, and conclusions. Subsequently, I utilize ChatGPT to refine and complete the entirety of this article.

References

Kevin Frans, Lisa Soros, and Olaf Witkowski. CLIPDraw: Exploring text-to-drawing synthesis through language-image encoders. In *Advances in Neural Information Processing Systems*, volume 35, pages 5207–5218. Curran Associates, Inc.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2022. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

[1] https://github.com/Heidelberg-NLP/MM-SHAP/issues/7

Table 1: In the initial stages of experiment design, it was hypothesized that larger models would yield superior pairwise accuracy across all dimensions. Factors under consideration encompassed the application of text SHAP analysis on caption pairs comprising both text and image. Furthermore, the dataset includes foil pairs of text and image, where images are mismatched with incorrect descriptions. The term 'Acc_r' signifies pairwise accuracy. The ultimate evaluation metric also incorporates the time required for MM-SHAP analysis.

CLIP-vit-...	Acc_r	T-SHAP_c	T-SHAP_f	Patch	Model	Image	Eval_time
base-patch32	77.03	47.05	28.68	32	base	224	12.5 h
B-32-laion2B-s34B-b79K	54.06	53.38	25.23	32	base	224	14.0 h
large-patch14	69.50	48.00	29.28	14	large	224	132.0 h
large-patch14-336	71.68	63.15	23.02	14	large	336	253.0 h

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.