Letitia Parcalabescu

Measuring the Contributions of Vision and Text Modalities in Multimodal Transformers

Understanding natural language requires machines to grasp more than just the surface-level form of text; they must also comprehend the underlying meaning, which involves knowledge of the world humans inhabit. While text provides substantial information, humans acquire much of their knowledge through other modalities like vision and sound. For machines to develop a better understanding of the world and the natural language that describes it, they need access to multiple modalities. A great starting point is the visual modality, given its relevance in human perception and its rich contribution to our understanding of the world.

This dissertation (Parcalabescu, 2024) explores vision and language models, which are multimodal systems that take vision and text modalities to produce outputs. Specifically, it develops computational tools to assess the effectiveness of vision and language models in combining, understanding, using, and explaining information from these two modalities. We structure our investigation into three key goals: (i) measuring specific and task-overarching capabilities of vision and language models, (ii) interpreting these models to quantify how much they leverage and integrate information from both modalities, and (iii) evaluating their ability to self-consistently explain their outputs to users.

In the first part of this dissertation, we introduce VALSE, a benchmark dataset designed to assess the visio-linguistic grounding capabilities of vision and language models across specific linguistic phenomena. This benchmark challenges models to differentiate between correct image captions and so-called foils – captions that contain subtle errors targeting specific linguistic phenomena grounded in vision: existence, plurality, counting, spatial relations, actions, and entity coreference. We propose four automated strategies to construct VALSE, ensuring the development of reliable and valid foils. Our evaluation of five widely-used vision and language encoder models and three decoder models (generating text from vision and language inputs) reveals that, while these models effectively identify objects and their presence in images, they generally struggle with more complex phenomena such as actions and spatial relations. This benchmark establishes a critical, ongoing challenge for modern vision and language models, aiming to track the progress of pretrained vision and language models from a *linguistic perspective*, complementing traditional task-centred vision and language evaluations in the field.

In the second part of the dissertation, we analyse how much vision and language models integrate and use information from both modalities. To quantify this integration, we introduce a multimodality score called MM-SHAP, which we designed to complement performance metrics, such as accuracy. This score is based on Shapley values, offering a performance-agnostic method to reliably determine the extent to which a multimodal model leverages individual modalities. With MM-SHAP, we assess different model

architectures for their overall degree of multimodality and evaluate the specific contributions of each modality within individual models on specific datasets and samples. Our findings challenge the belief that unimodal collapse – where a model predominantly relies on one modality – occurs uniformly in one direction. Instead, we observe that unimodal collapse can manifest in varying degrees and in different directions. Based on these insights, we recommend MM-SHAP for interpreting multimodal models and tasks, for diagnosing and guiding progress towards true multimodal integration.

In the third part of this dissertation, we explore whether vision and language models can give self-consistent explanations for their predictions. But the utility of these explanations hinges on their faithfulness, i.e., their accuracy in reflecting the model's inner workings. Therefore, we need to test explanations for faithfulness. We clarify the status of existing faithfulness tests (developed almost solely for language-only models) in view of model explainability, characterising them as self-consistency tests instead. We compare all previous tests using the same models on the same datasets, and show that the predictions differ widely. We argue that the overall result at least questions the commonly-held view that these tests measure faithfulness, because they yield highly diverse predictions. While existing tests require input edits to test whether the model output changes, we propose CC-SHAP, an edit-free and interpretable measure, that analyses how model outputs relate to *how* the model processes the input. We compare CC-SHAP for 11 language models on 5 tasks against all other tests and show its advantages supported by individual examples for language only models. We find that chat language models show higher self-consistency than their base variants.

Finally, we extend CC-SHAP to vision and language models, and we are first to evaluate the self-consistency of vision and language models in both *post-hoc* and *chain-of-thought explanation* settings. We assess the self-consistency of 3 vision and language models on 11 datasets with CC-SHAP. We also apply the existing language-only self-consistency (faithfulness) tests in our multimodal setting. We find that vision and language models are less self-consistent than language-only models. Furthermore, the contributions of the image are significantly larger for explanation generation than for answer generation, and the difference is even more pronounced in chain-of-thought compared to the post-hoc explanation setting. This added complexity in the behaviour of vision and language models, as compared to their language-only counterparts, opens up new avenues for future research into the explainability of multimodal models.

We expect that the research contributions presented in this dissertation will continue to help measure the progress of vision and language research, and inspire future research on model benchmarking, interpretability and explainability.

References

Parcalabescu, L. (2024). *Measuring the contributions of vision and text modalities in multimodal transformers* (Doctoral dissertation). Universität Heidelberg.

Correspondence

Letitia Parcalabescu

Universität Heidelberg
Institut für Computerlinguistik
Heidelberg, Germany
parcalabescu@cl.uni-heidelberg.de

Thesis Information

Doctoral thesis defended on July 19, 2024, at the Department of Computational Linguistics at Heidelberg University. Supervised by Prof. Dr. Anette Frank. Full text available at the Heidelberger Dokumentenserver: https://archiv.ub.uni-heidelberg.de/volltextserver/35753/.