Improving Contextual Congruence Across Modalities for Effective Multimodal Marketing using Knowledge-infused Learning

Trilok Padhi
Georgia State University
tpadhi1@student.gsu.edu

Ugur Kursuncu
Georgia State University
ugur@gsu.edu

Yaman Kumar
Adobe Research
ykumar@adobe.com

Valerie L. Shalin
Wright State University
valerie.shalin@wright.edu

Lane Peterson Fronczek
Georgia State University
lfronczek@gsu.edu

Abstract

The prevalence of smart devices with the ability to capture moments in multiple modalities has enabled users to experience multimodal information online. However, large Language (LLMs) and Vision models (LVMs) are still limited in capturing holistic meaning with cross-modal semantic relationships. In this work, we design a framework to couple explicit commonsense knowledge in the form of knowledge graphs with large VLMs to improve the performance of a downstream task, predicting the effectiveness of multi-modal marketing campaigns. We find that external knowledge improves the contextual congruence of multimodal representation, boosting predictive performance.

1 Introduction

Contemporary online multimodal platforms provide an appealing rich user experience from social media to e-commerce online shopping. Users may promote or market their ideas, products, or personal brands through a combination of various types of modalities, such as text and images, to attract consumer attention. The computational modeling of multi-modal content makes little explicit contact with human knowledge, experience, and reasoning. The smaller, earlier (multimodal) Visual Language Models (VLMs) such as MMBT (Kiela et al., 2020), ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) typically exploit independent unimodal cues (e.g., textual, visual) connecting the image and text. More recent very large VLMs, such as LLaVA (Liu et al., 2023), BLIP2 (Li et al., 2023) and GPT-4 (OpenAI, 2023), do capture *implicit* multimodal semantic relationships, but without semantic constraints, making them vulnerable to hallucinations— *superficially congruent* (Hasan, 2023; Meyer et al., 2022) and *catastrophically incorrect* (Shuster et al., 2021; Maynez et al., 2020). Moreover, these models often remain unable to identify the explicit semantic connections between modalities that influence human interpretation. (Ji et al., 2020).

We introduce here the notion of *multimodal congruence.* We measure multimodal congruence using the semantic distance between the separate image and caption (text) representations, as this distance is expected to be shorter in a congruent multimodal representation (Mandera et al., 2017; Frank and Willems, 2017; Beck and Diehl, 2011; Maki et al., 2004). We can then demonstrate how external knowledge can enhance and contextualize the representations to close the semantic gap between the modalities. In this study, we answer the following two research questions:
RQ1: Can we improve the contextual congruence of the representations of multimodal content by incorporating external knowledge while learning to unveil subtle cross-modal semantic relationships?
RQ2: Do more contextually congruent representations help the model obtain more consistent and reliable predictive performance for the success of multimodal marketing campaigns?

Our technical approach incorporates an external commonsense Knowledge Graph (KG) (e.g., ConceptNet) to learn knowledge-infused multimodal representations of the data. The novelty of our work lies in modeling a social and behavioral problem, such as effective marketing, utilizing neural and symbolic models (Sap et al., 2022). In so doing, we particularly addressed the problem of learning a contextually enhanced neural representation of multimodal content with external commonsense KG (i.e., symbolic) (Yu et al., 2022; Kursuncu et al., 2020). This helps the model to better capture the holistic meaning across modalities enhancing model performance in predicting the success of multimodal marketing campaigns.

2 Methodology

Our models incorporate external knowledge from a KG (e.g., ConceptNet) in a multimodal learning

#	Vision	Language	Knowledge	Precision	Recall	F1	AUC
1	Resnet152	BERT (Kiela et al., 2020)	-	0.86	0.77	0.81	0.86
2	ViT	BERT	-	0.88	0.84	0.86	0.86
3	ViT	RoBERTa	-	0.92	0.88	0.91	0.91
4	MDL-TIM (Cheng et al., 2019)			0.76	0.76	0.76	0.80
5	BLIP (Li et al., 2022)			0.93	0.89	0.91	0.92
6	ViT	RoBERTa	TransE	0.92	0.90	0.91	0.91
7	Resnet152	RoBERTa	TransE	0.95	0.91	0.92	0.94
8	Resnet152	BERT	TransE	0.93	0.89	0.91	0.93

Table 1: Results of the baseline models (#1-5) and the Knowledge-enhanced multimodal models, with different text and image encoders and KG embedding models.

framework, to predict the success of a marketing campaign. We formulate this as a binary classification task. We use a supervised multimodal learning framework (Kiela et al., 2020) to generate a multimodal representation. We devise a knowledge retrieval component that extracts the most relevant concepts from ConceptNet and generates their knowledge representations. For knowledge representations, we trained Knowledge Graph Embedding (KGE) models as described in the Section for Knowledge Retrieval and Representation. Then, we fuse the multimodal representation with the knowledge representation to obtain a knowledge-infused representation that is input to the classification layer.

2.1 Technical Approach

Our technical approach incorporates an external commonsense Knowledge Graph (KG) (e.g., ConceptNet) to learn knowledge-infused multimodal representations of the data. We first take pairs of images and text from our dataset and generate embeddings using text and image encoders (i.e., via LLMs, LVMs). To query the KG, we generate captions for images using BLIP (Li et al., 2022). Then, we retrieve the most similar concepts from ConceptNet based on semantic search. We train knowledge graph embedding (KGE) models to generate representations of the resulting concepts. Initially, we learn multimodal data representations using bidirectional transformers. These representations are then fused with the KGEs through a linear and a multi-head cross-attention layer.

3 Results & Discussion

The results of the success prediction of online crowdfunding campaign success is presented in Table 1. Our models with knowledge-infused representations (#6-8) performed better overall than the baseline models (#1-5). The best performance was achieved with Resnet152 and RoBERTa with the TransE KG embeddings with 0.95, 0.91, and 0.92, precision, recall, and F1-score, respectively. The lowest performance came from the MDL-TIM model (#4) while the two baseline models, ViT/RoBERTa (#3) and BLIP-only (#5), were much better. However, knowledge-infused models consistently performed better. Regarding AUC, our best knowledge-infused model (#7) outperforms other models with 94%, indicating a higher true positive rate and a lower false positive rate. The outperformance of our model in AUC also signals potential improvement in the fairness of the model, warranting future work on the fairness assessment of knowledge-infused models. While we have performed other experiments, we left them out due to space limitation.

4 Conclusion

Online marketing involves both image and text, and hence provides an ideal domain for computing the relationship between these two modalities. The domain also offers a success metric for the adequacy of a semantic interpretation; good interpretations are those that predict which image text pairs will result in successful campaigns. We developed a computational approach to capture the environmental context and congruence in multimodal online marketing content. External knowledge from KGs (e.g., ConceptNet) enhanced the multimodal representation that narrowed down the contextual distance between the two modalities of a multimodal content. External knowledge from KGs improved the success prediction of online crowdfunding campaigns. The approach facilitates the prediction of successful marketing campaigns, including the crowd funding application studied here and more generally, the effectiveness of persuasive economic and public service campaigns.

Ethics Statement

While our work is in the application domain of marketing, which is benign, we recognize understanding (and thereby predicting) multimodal persuasiveness has extensive applications. These are both benign, as in purchasing products or services, or malicious where prediction is a paramount tool for ensuring safety. While we demonstrate such enhancement for effective multimodal marketing, a similar approach will be necessary for other context-sensitive social and behavioral problems including toxicity and extremism, among others. Further, there are still unknown impli- cations of these AI-enabled technologies for individuals and communities, specifically, underrepresented groups in society. This work specifically provides potential avenues to address social biases, inadvertently encoded in large foundation models (e.g., LLMs, LVMs, VLMs) using a framework that incorporates external knowledge in learning, for future work.

Acknowledgements

We acknowledge support from Adobe Inc. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Adobe.

References

Fabian Beck and Stephan Diehl. 2011. On the congruence of modularity and code coupling. In *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering*, pages 354–364.

Chaoran Cheng, Fei Tan, Xiurui Hou, and Zhi Wei. 2019. Success prediction on crowdfunding with multimodal deep learning. In *IJCAI*, pages 2158–2164.

Stefan L Frank and Roel M Willems. 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. *Language, Cognition and Neuroscience*, 32(9):1192–1203.

Ali Hasan. 2023. Why you are (probably) anthropomorphizing ai (short version).

Zizheng Ji, Lin Dai, Jin Pang, and Tingting Shen. 2020. Leveraging concept-enhanced pre-training model and masked-entity language model for named entity disambiguation. *IEEE Access*, 8:100469–100484.

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

Ugur Kursuncu, Manas Gaur, and Amit Sheth. 2020. Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning. *Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice (AAAI-MAKE.*

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597.*

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485.*

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

William S Maki, Lauren N McKinley, and Amber G Thompson. 2004. Semantic distance norms computed from an electronic dictionary (wordnet). *Behavior Research Methods, Instruments, & Computers*, 36:421–431.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92:57–78.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization.

Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E Losada. 2022. Do we still need human assessors? Prompt-Based GPT-3 user simulation in conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces*, number Article 8 in CUI '22, pages 1–6, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. Gpt-4 technical report.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312.*

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567.*

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Dongran Yu, Bo Yang, Qianhao Wei, Anchen Li, and Shirui Pan. 2022. A probabilistic graphical model based on neural-symbolic reasoning for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618.