Context-aware Modeling for Effective Marketing using Multimodal Knowledge-infused Learning

Trilok Padhi, Ugur Kurşuncu, Yaman Kumar, Valerie Shalin, Lane Peterson

SWAN AI Research Group, Institute for Insight

MOTIVATION

- ➔Multimodal Visual Language Models (VLM) only capture high level ***relationships across modalities*** of content leaving representation ***contextually incongruent***.
- ➔***Contextual congruence*** improves human behavioral responses, for which a similar mechanism can be designed for machine.

PROBLEM

Current multimodal VLMs often generate representations of the content with ***contextual incongruence*** and ***inaccurate*** information (i.e., hallucinations), which impacts overall performance of downstream tasks such as prediction for effective marketing.

Research Questions

- ➔**RQ1:** How can we ***improve the contextual congruence*** of the multimodal representations by incorporating external knowledge from ***commonsense knowledge graphs***?
- ➔**RQ2:** Do more contextually congruent representations ***improve the predicted success*** of ***multimodal marketing campaigns***?

OUR APPROACH

We couple explicit external knowledge in the form of ***knowledge graphs*** with ***large VLMs*** to improve the performance of a downstream task, the ***classification of marketing campaigns for effectiveness.***

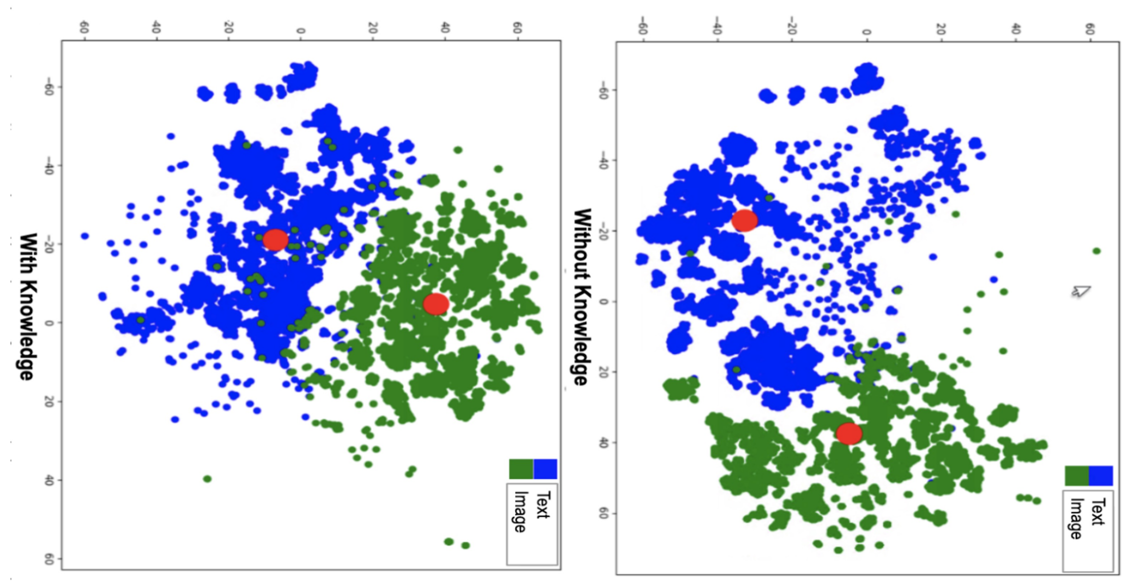
EXPLORATORY ANALYSIS

Contextual congruence refers to the ability of human intelligence to form associations in the presence of multiple cues from different modalities.

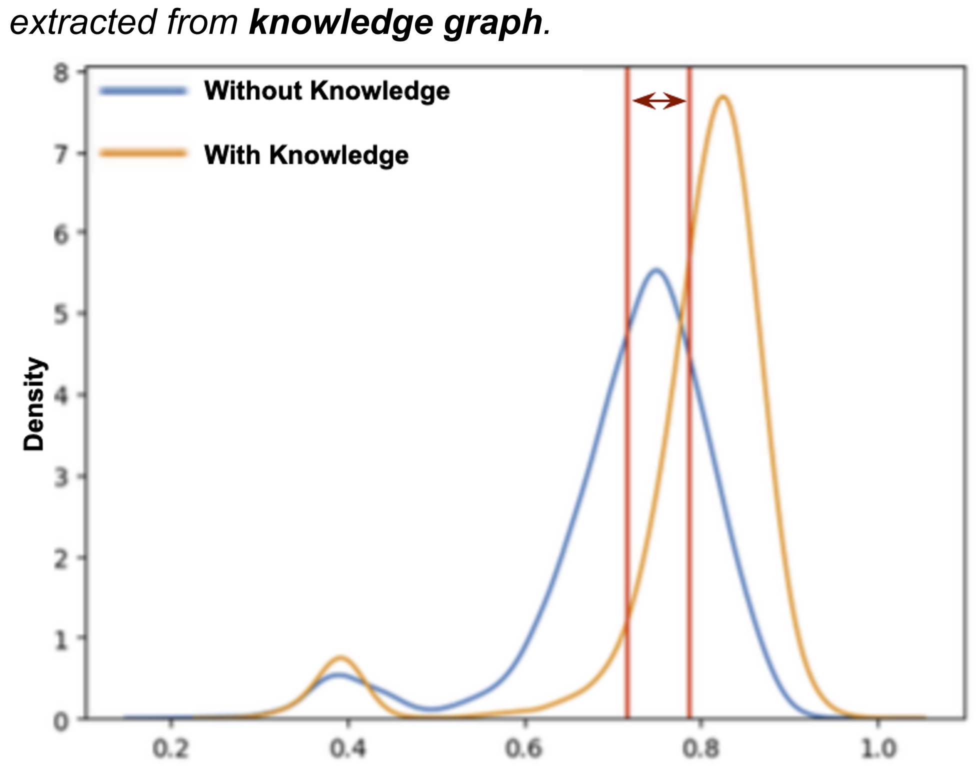
Actual human caption: "Smash the glass ceiling. Destroy the patriarchy. Save the record store."

BLIP (VLM): "Two women smiling with a hand gesture of rock and roll."

Llava (VLM): The image features two women standing next to each other, *both holding their cell phones in their hands. They seem to be taking selfies*"



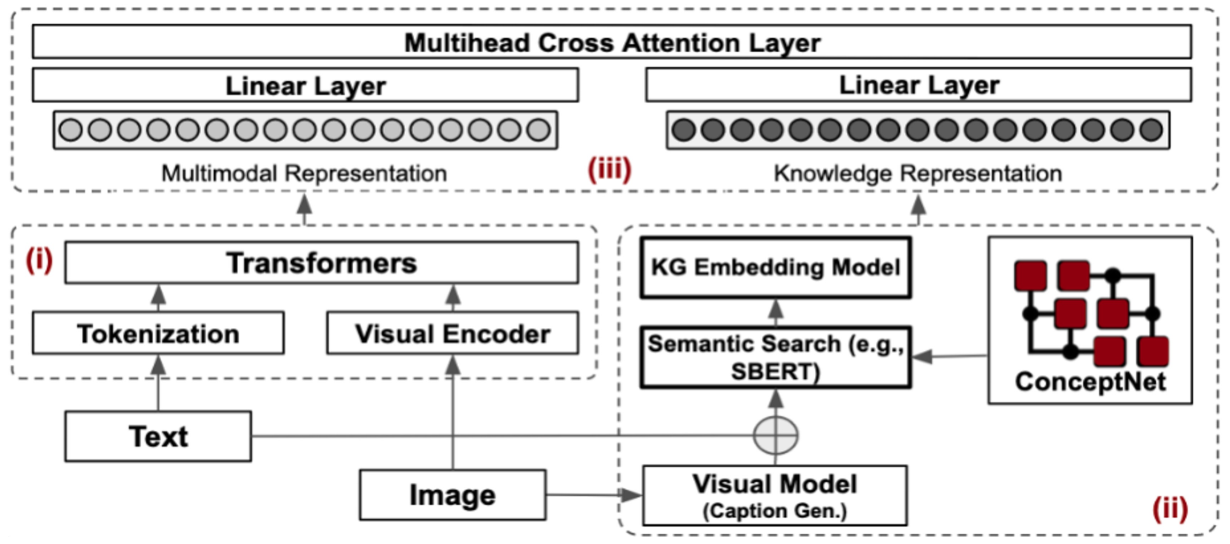
t-SNE visualization of text and image caption (BLIP) embeddings as two clusters. The ***red dots*** *represent the centroids in each cluster. The two clusters get denser and* ***the distance*** *between the* ***text*** *and* ***image*** *clusters reduces when we include* ***concepts*** *that we extracted from* ***knowledge graph****.*



Density Plot demonstrates the difference between the similarities (cosine) of the ***image*** *and* ***text*** *embeddings* ***with*** *and* ***without knowledge****. The inclusion of* ***knowledge*** *in the input gets text and image modalities* ***closer*** *by about* ***9.9%****.*

MODELING

- ➔An external commonsense ***Knowledge Graph (KG) (ConceptNet)*** is used to learn ***knowledge infused multimodal representations*** of data.
- ➔**Semantic search** was utilized to retrieve the most similar concepts from ConceptNet.
- ➔***Multi-Head Cross Attention Layer*** was used for the ***fusion*** of the multimodal and knowledge representations.



Our approach consists of three main components: (i) **multimodal learning***, (ii)* ***knowledge retrieval*** *and representation, and (iii)* ***Knowledge Fusion Layer****. The Retrieval component identifies the* ***most relevant concepts*** *from ConceptNet. The concepts' knowledge embeddings are generated. The knowledge fusion layer fuses the multimodal representations with knowledge embeddings.*

RESULTS

- ➔***Knowledge-infused representations*** give overall ***better performance*** compared to the baseline models.
- ➔Knowledge-infused models demonstrate **potential improvement in fairness** of the models with ***higher AUC up to 94%***.

Sl. No.	Vision	Language	Knowledge	Pr	Re	F1	AUC
1	Resenet152	Bert		0.86	0.77	0.81	0.86
2	ViT	Bert		0.88	0.84	0.84	0.86
3	ViT	RoBERTa		0.92	0.88	0.91	0.91
4	BLIP			0.93	0.89	0.91	0.92
5	**Resnet152**	**Roberta**	**TransE**	**0.95**	**0.91**	**0.92**	**0.94**
6	Resnet152	Roberta	RotatE	0.93	0.92	0.92	0.93
7	Resnet152	Roberta	DistMult	0.95	0.90	0.92	0.93

Original Label: Successful
Predicted Label w/ knowledge: Successful

Predicted Label Baseline 1: Unsuccessful
Predicted Label Baseline 5: Unsuccessful

Actual Human Caption : A project to expand our community news platform to include public education coverage in Charlottesville-Albemarle.
BLIP Caption for Image : a group of children holding a large kite
Retrieved Concepts for Actual Caption : ['hot_topic_in_public_education', 'charlottesville', 'educating_public', 'newspaper_coverage', 'newsboards']
Retrieved Concepts for Generated Caption for Image : ['kites', 'kite', 'kiters', 'kitemakers', 'kitel']

In the yellow box, the prediction of the model with ***knowledge*** *and two* ***baseline*** *models, In the green box, actual caption, BLIP caption, and retrieved concepts for each modality are shown.*

FINDINGS

- ➔Incorporating ***external knowledge*** from ***commonsense KGs*** improves ***contextual congruence*** of multimodal representations. Our approach captures the contextual connections across modalities improving the congruence.
- ➔Such ***improvement in congruence*** for multimodal representations ***improves performance*** *on downstream tasks*, ***effectiveness of multimodal marketing campaigns***.

References

1. Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." International Conference on Machine Learning. PMLR, 2022.
2. Liu, Haotian, et al. "Visual instruction tuning." arXiv preprint arXiv:2304.08485 (2023).
3. Suntis qui omnist, eaquaerum exeriaectios venihillora doloresti ape litatusapiet re sedicil mos moluptatiam fugit ides di apicium atiam et audam, sequam rem.
4. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
5. Han, Kai, et al. "A survey on vision transformer." IEEE transactions on pattern analysis and machine intelligence 45.1 (2022): 87-110.

Acknowledgements

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