Robust Multimodal Learning

Shah Nawaz

JOHANNES KEPLER
UNIVERSITY LINZ

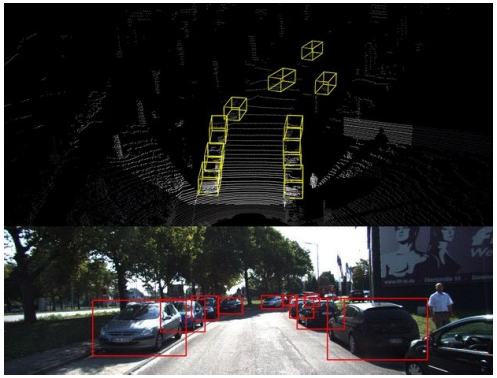
MULTIMEDIA
MINING AND
SEARCH GROUP

INSTITUTE OF
COMPUTATIONAL
PERCEPTION

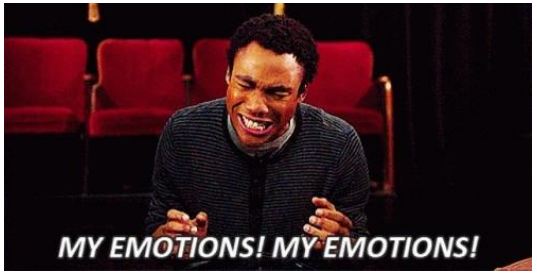
Real-world Artificial Intelligence

Digital intelligence
Multimedia Image/video description

Physical intelligence
Embodied AI, autonomous driving



Social intelligence
Affective computing
Human-AI interaction

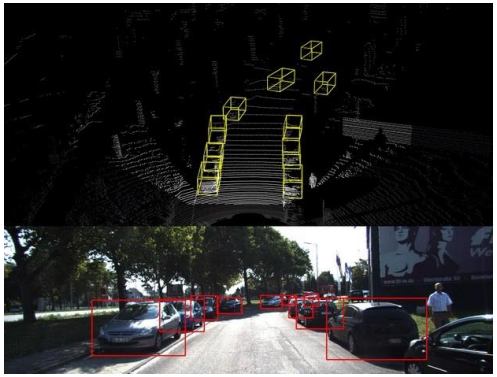


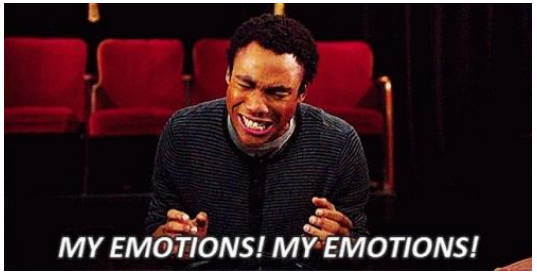
Multimodal Artificial Intelligence

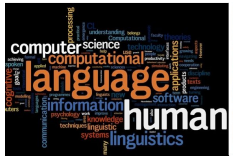
Digital intelligence
Multimedia Image/video description

Physical intelligence
Embodied AI, autonomous driving

Social intelligence
Affective computing
Human-AI interaction

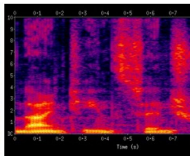




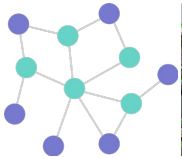


Language

Image



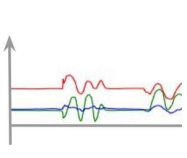
Audio



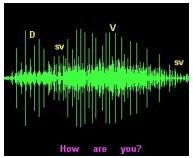
Graphs

Video

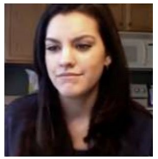
LIDAR



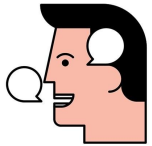
Sensors



Speech



Video

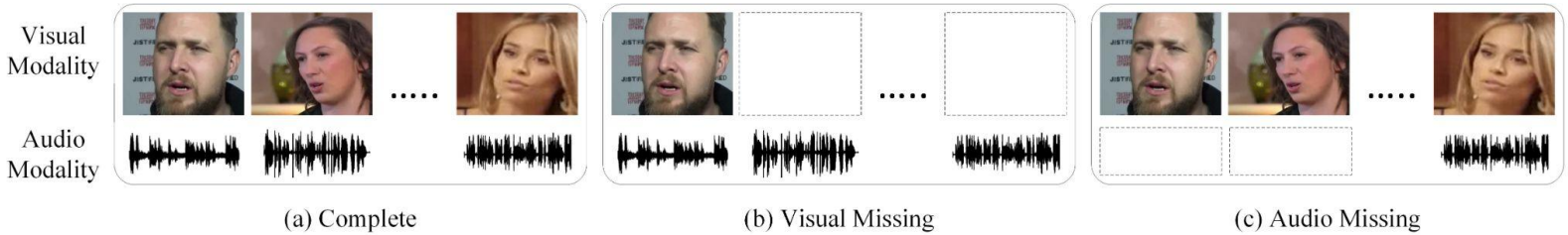


Language

JOHANNES KEPLER UNIVERSITY LINZ

Background and Motivation

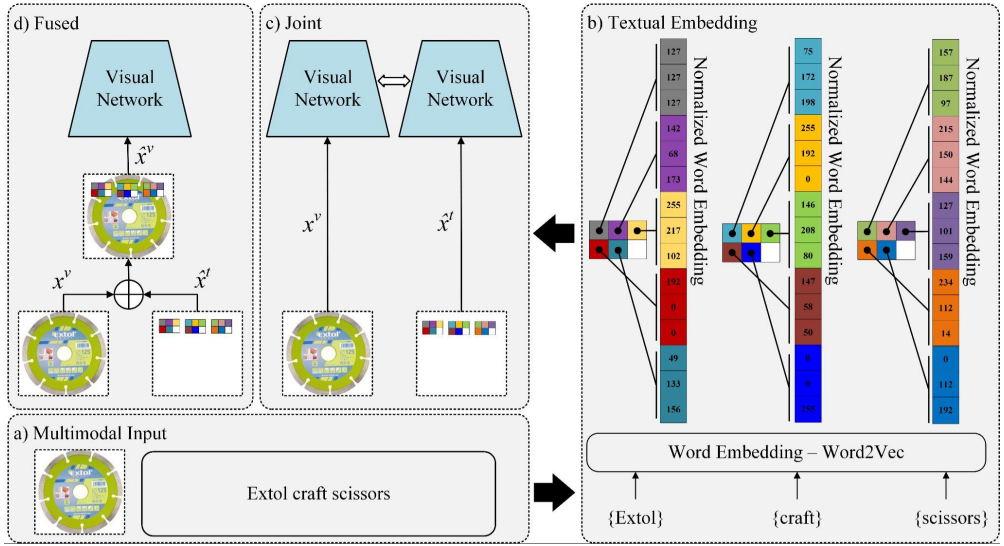
- Missing modalities scenarios



Dataset	Training		Testing		Accuracy	$\Delta \downarrow$
	Image	Text	Image	Text		
UPMC Food-101	100%	100%	100%	100%	91.9	-
	100%	100%	100%	30%	65.9	28.3%
	100%	0%	100%	0%	71.5	-

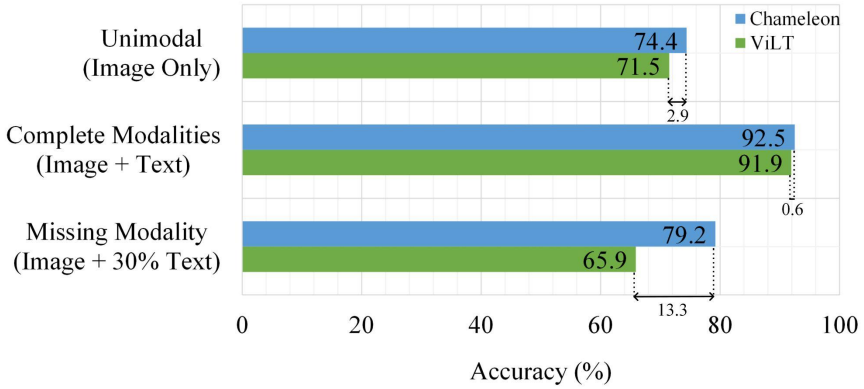
Ma, M., Ren, J., Zhao, L., Testuggine, D., & Peng, X. (2022). Are multimodal transformers robust to missing modality?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18177-18186).

Common Input format (Pixel Level) - Architecture



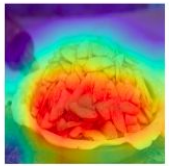
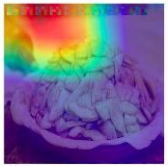
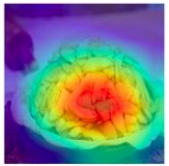
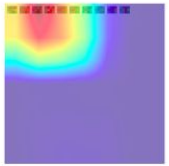
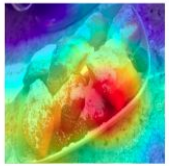
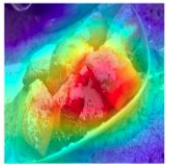
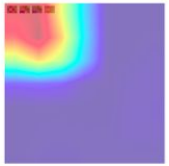
Liaqat, M. I., **Nawaz, S.**, Zaheer, M. Z., Saeed, M. S., Sajjad, H., De Schepper, T., ... & Schedl, M. H. K. M. (2024). Chameleon: Images Are What You Need For Multimodal Learning Robust To Missing Modalities. arXiv preprint arXiv:2407.16243.

Common Input format (Pixel Level) - Results



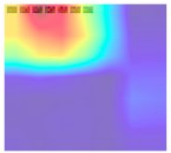
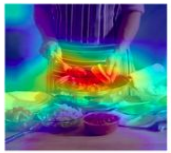
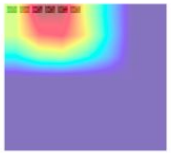
Liaqat, M. I., **Nawaz, S.**, Zaheer, M. Z., Saeed, M. S., Sajjad, H., De Schepper, T., ... & Schedl, M. H. K. M. (2024). Chameleon: Images Are What You Need For Multimodal Learning Robust To Missing Modalities. arXiv preprint arXiv:2407.16243.

Common Input format (Pixel Level) - Internal Working

		Chameleon			
	a. Input Image	b. Unimodal	c. Multimodal	d. Missing Text	e. Missing Image
		Successful Cases			
Apple Pie		 Apple Pie (0.9984)	 Apple Pie (0.9999)	 Apple Pie (0.7052)	 Apple Pie (0.9998)
Beignets		 Beignets 0.9989	 Beignets 0.9999	 Beignets 0.9999	 Beignets 0.9998

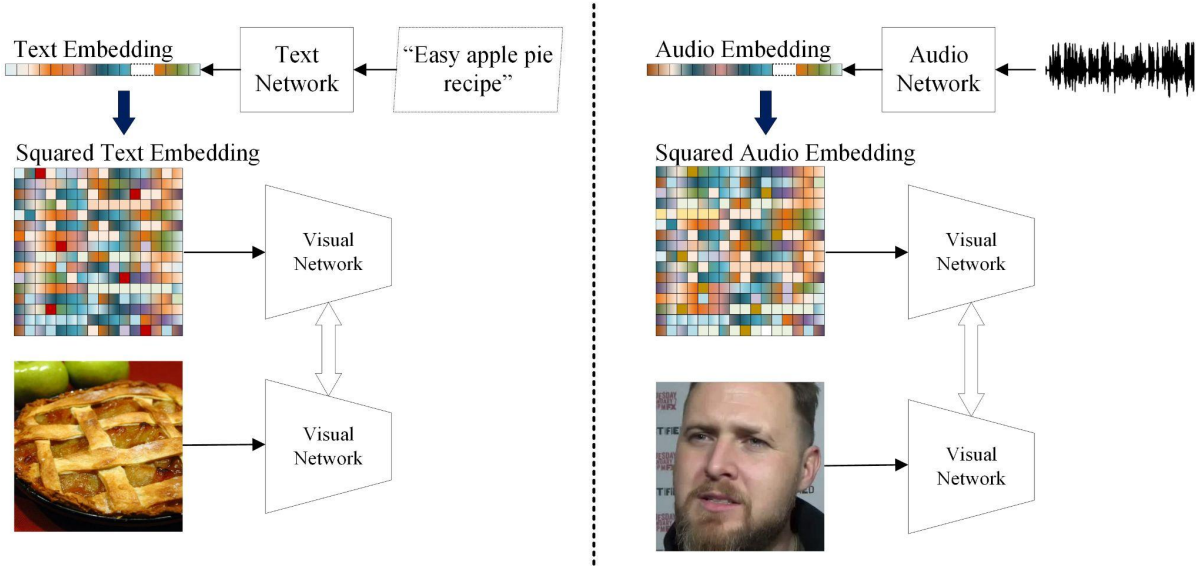
Liaqat, M. I., **Nawaz, S.**, Zaheer, M. Z., Saeed, M. S., Sajjad, H., De Schepper, T., ... & Schedl, M. H. K. M. (2024). Chameleon: Images Are What You Need For Multimodal Learning Robust To Missing Modalities. arXiv preprint arXiv:2407.16243.

Common Input format (Pixel Level) - Internal Working

		Chameleon		
	a. Input Image	b. Unimodal	c. Multimodal	d. Missing Text	e. Missing Image
Failure Cases					
Sashimi		 Sashimi (0.6064)	 Sushi (0.9371)	 Sashimi (0.9960)	 Sushi (0.9956)
Tacos		 Spaghetti Bolognese (0.3015)	 Tacos (0.9991)	 Spaghetti Bolognese (0.2771)	 Tacos (0.9997)

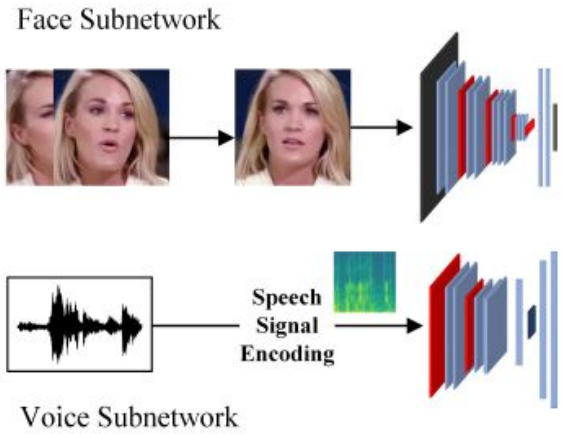
Liaqat, M. I., **Nawaz, S.**, Zaheer, M. Z., Saeed, M. S., Sajjad, H., De Schepper, T., ... & Schedl, M. H. K. M. (2024). Chameleon: Images Are What You Need For Multimodal Learning Robust To Missing Modalities. arXiv preprint arXiv:2407.16243.

Common Input format (Pixel Level) - Architecture



Liaqat, M. I., **Nawaz, S.**, Zaheer, M. Z., Saeed, M. S., Sajjad, H., De Schepper, T., ... & Schedl, M. H. K. M. (2024). Chameleon: Images Are What You Need For Multimodal Learning Robust To Missing Modalities. arXiv preprint arXiv:2407.16243.

Common Input format (Embedding Level) - Motivation



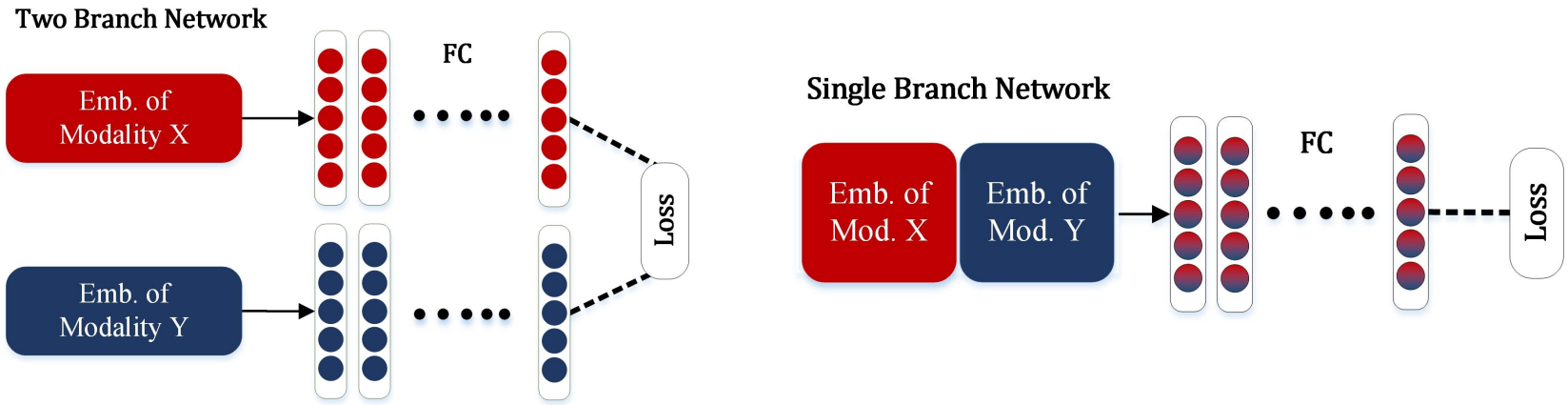
Face Embeddings

Voice Embeddings

Take away message
Embeddings extracted from modality-specific networks share many semantic similarities.
For example, the gender, nationality, and age of speakers are represented by their audio and visual signatures

Saeed, M. S., **Nawaz, S.**, Khan, M. H., Zaheer, M. Z., Nandakumar, K., Yousaf, M. H., & Mahmood, A. (2023, June). Single-branch network for multimodal training. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Common Input format (Embedding Level) - Motivation

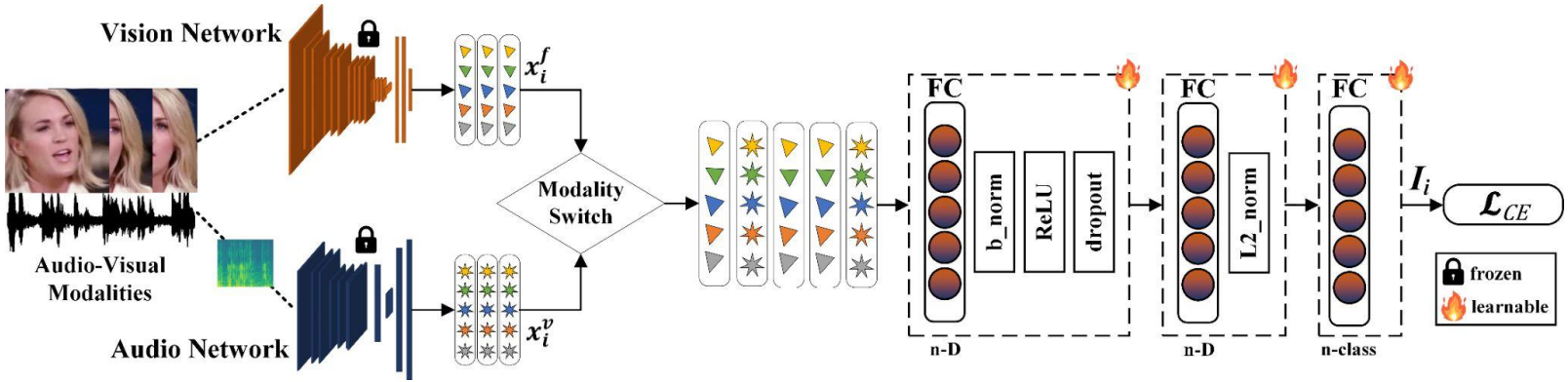


Take away message

The existing two-branch networks employ independent modality-specific branches to learn a joint representation from the embeddings of modality X and Y. In contrast, the proposed single-branch network leverages only one branch to learn similar representations.

Saeed, M. S., **Nawaz, S.**, Khan, M. H., Zaheer, M. Z., Nandakumar, K., Yousaf, M. H., & Mahmood, A. (2023, June). Single-branch network for multimodal training. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

Common Input format (Embedding Level) - Architecture



Saeed, M. S., **Nawaz, S.**, Zaheer, M. Z., Khan, M. H., Nandakumar, K., Yousaf, M. H., ... & Schedl, M. (2024). Modality Invariant Multimodal Learning to Handle Missing Modalities: A Single-Branch Approach. arXiv preprint arXiv:2408.07445.

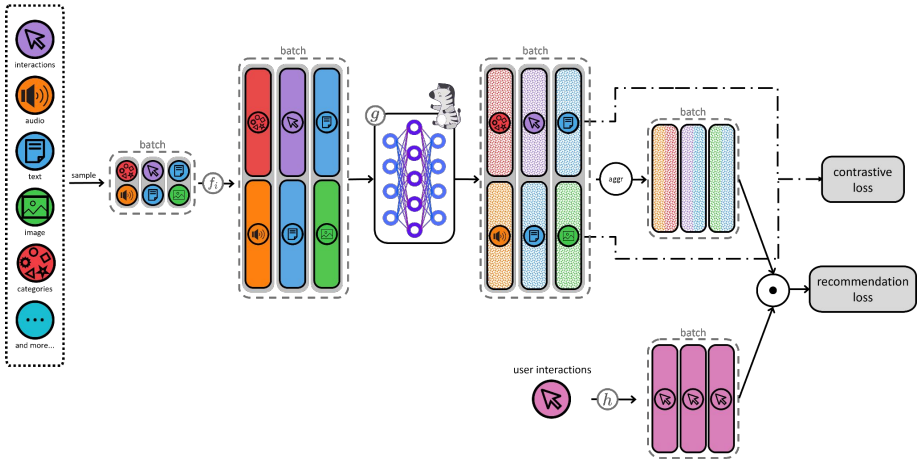
Common Input format (Embedding Level) - Results

Dataset	Methods	Settings	Training		Testing		Accuracy	Δ ↓
			Image	Text	Image	Text		
UPMC Food-101	ViLT	Complete Modalities	100%	100%	100%	100%	91.9	-
		Missing Modality	100%	100%	100%	30%	65.9	28.3%
	SRMM	Complete Modalities	100%	100%	100%	100%	**94.6**	-
		Missing Modality	100%	100%	100%	30%	**84.8**	**12.3%**

Saeed, M. S., **Nawaz, S.**, Zaheer, M. Z., Khan, M. H., Nandakumar, K., Yousaf, M. H., ... & Schedl, M. (2024). Modality Invariant Multimodal Learning to Handle Missing Modalities: A Single-Branch Approach. arXiv preprint arXiv:2408.07445.

SiBraR – Single-Branch for Recommendation

- SiBraR is a multimodal recommender system using a single-branch architecture to reduce the modality gap.



Ganhör, C., Moscati, M., Hausberger, A., **Nawaz, S.**, & Schedl, M. (2024, October). A Multimodal Single-Branch Embedding Network for Recommendation in Cold-Start and Missing Modality Scenarios. In Proceedings of the 18th ACM Conference on Recommender Systems (pp. 380-390).

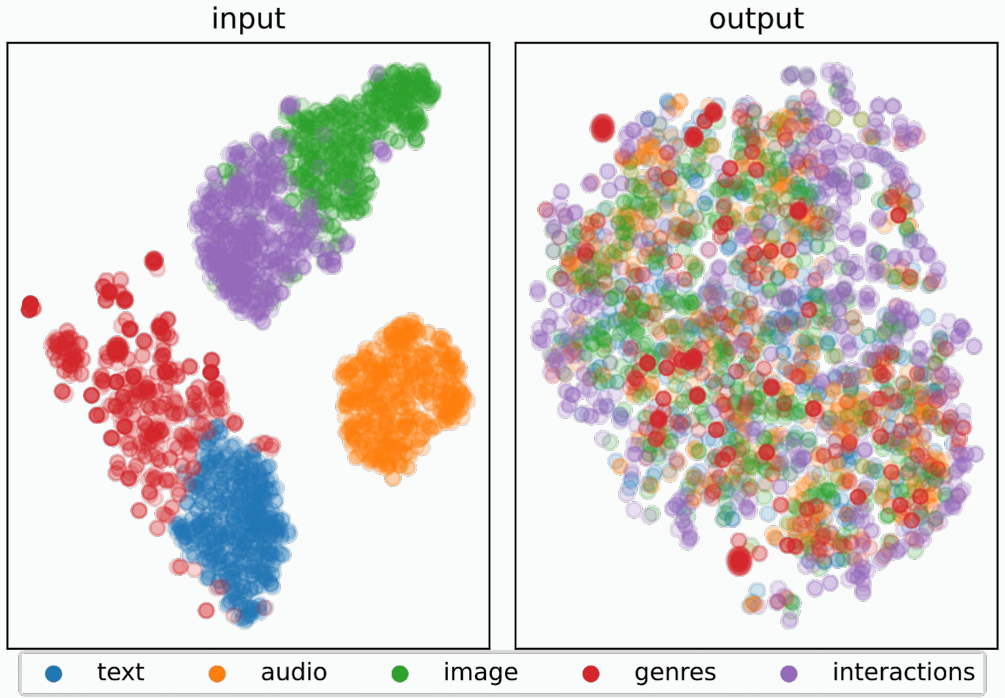
SiBraR – Multimodal Single-Branch Recommender

Training techniques:

- weight-sharing
- contrastive loss

Effect:

Reduce modality gap, thus interchangeable modalities.



Ganhör, C., Moscati, M., Hausberger, A., **Nawaz, S.**, & Schedl, M. (2024, October). A Multimodal Single-Branch Embedding Network for Recommendation in Cold-Start and Missing Modality Scenarios. In Proceedings of the 18th ACM Conference on Recommender Systems (pp. 380-390).

Collaborators

Muhammd Saad Saeed

Muhammad Zaigham Zaheer

Muhammad Haris Khan

Karthik Nandakumar

Arif Mehmood

Muhmmad Haroon Yousaf

Marta Moscati

Markus Schedl

Christian Ganhör

Hassan Sajjad

Ignazio Gallo

Alessio Del Bue

Muhammad Irzam Liaqat

Rohan Kumar

JOHANNES KEPLER UNIVERSITY LINZ

Questions