Multimodal Learning

Fusing knowledge from different modalities in an explainable way

Estimated duration : 25 min

UMONS

POLYTECH MONS

inforTech

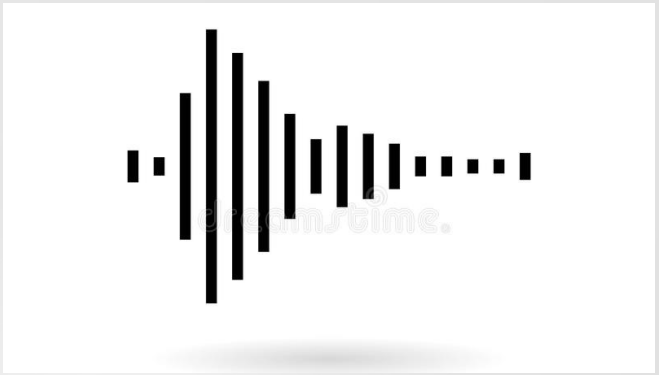
INSTITUT DE RECHERCHE
EN TECHNOLOGIES DE L'INFORMATION
ET SCIENCES DE L'INFORMATIQUE DE L'UMONS

Context

Multimodal learning suggests that when a number of our senses — visual, auditory, kinesthetic — are being engaged in the processing of information, we understand and remember more.

What about computers ?

Images



audio

Video

Lorem ipsum dolor

Utinam habemus assueverit et est. Elit pe

Ex eam nusquam commune. Vis eu perpet

Lorem ipsum dolor sit amet, te quaestio d

Utinam habemus assueverit et est. Elit pertinacia mea no. At eleife

Ex eam nusquam commune. Vis eu perpetua interesset. Utroque n

Lorem ipsum dolor sit amet, te quaestio dignissim repudiandae eo

Sed ut perspiciatis unde omnis iste natus error sit voluptatem acc

Text

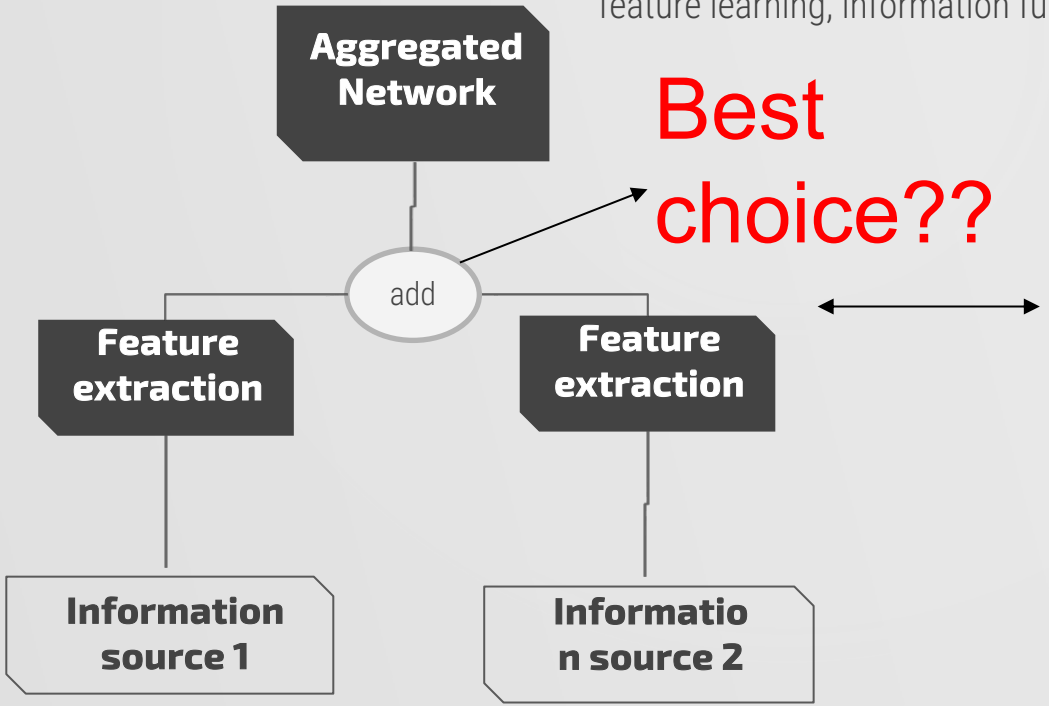
TABLE OF CONTENTS

01

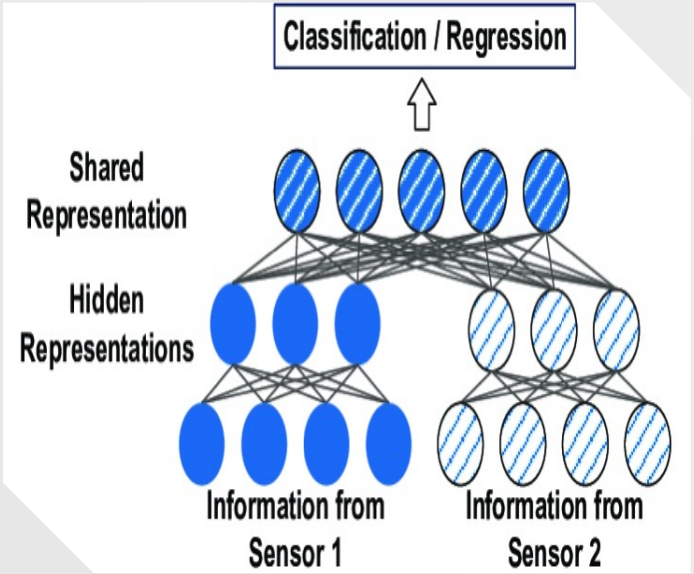
Problem statement

Problem statement

In multimodal learning, as its name suggests, we aim to do information fusion from different modalities to improve our network's predictive ability. The overall task can mainly be divided into three phases — individual feature learning, information fusion and testing.



Best choice??



Problem formulation

For an easy understanding, Let $\{x_1, x_2, x_3\}$ a set of **heterogeneous modalities**,

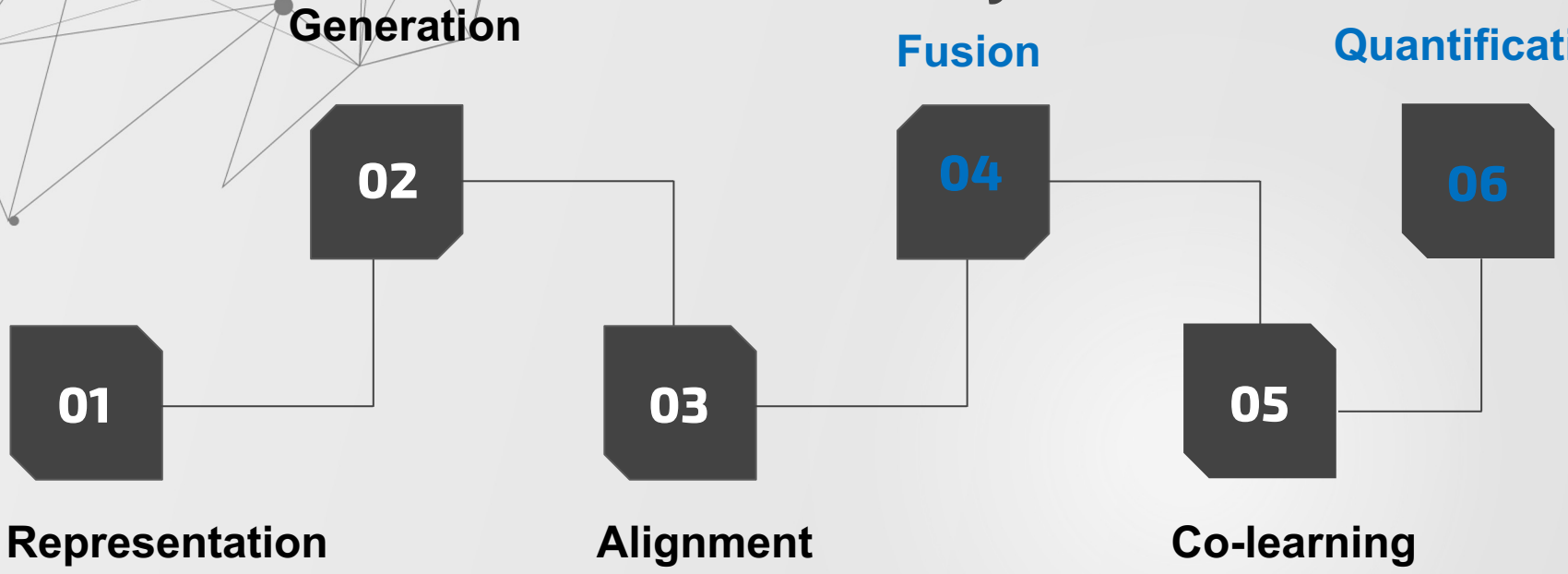
And f_1, f_2, f_3 **encoders** for each modality respectively,

The generated representations, denoted z, are generated as follows :
$z_1 = f(x_1)$, $z_2 = f(x_2)$, $z_3 = f(x_3)$

In order to generate a **latent representation** Z, we feed these generated features z_i to a **fusion operation** denoted $\oplus$ as follows :

$$Z = z_1 \oplus z_2 \oplus z_3$$

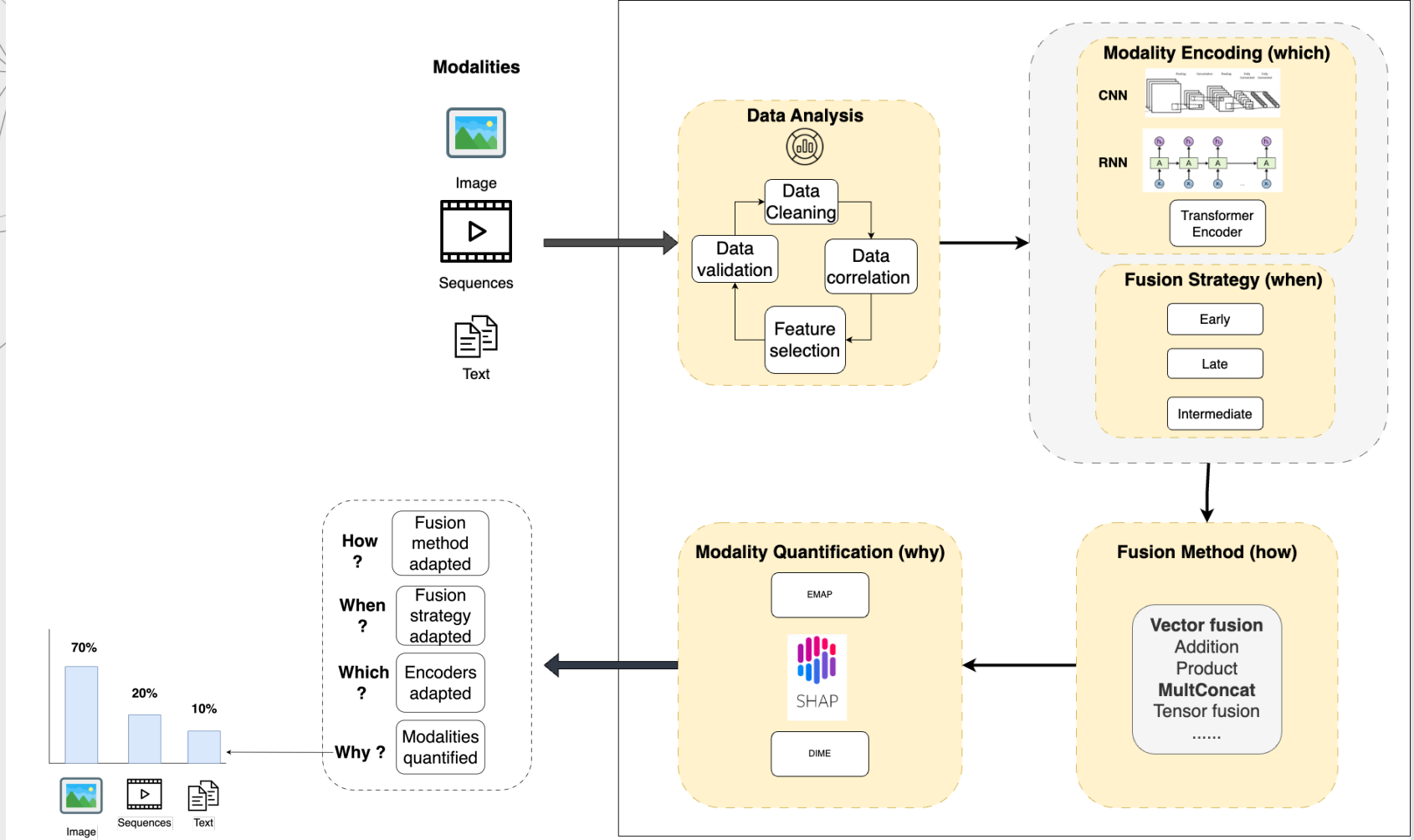
Multimodal Learning Challenges Taxonomy



* Our focus in this thesis will be on **Fusion** and **Quantification (XAI for multimodal)**

03

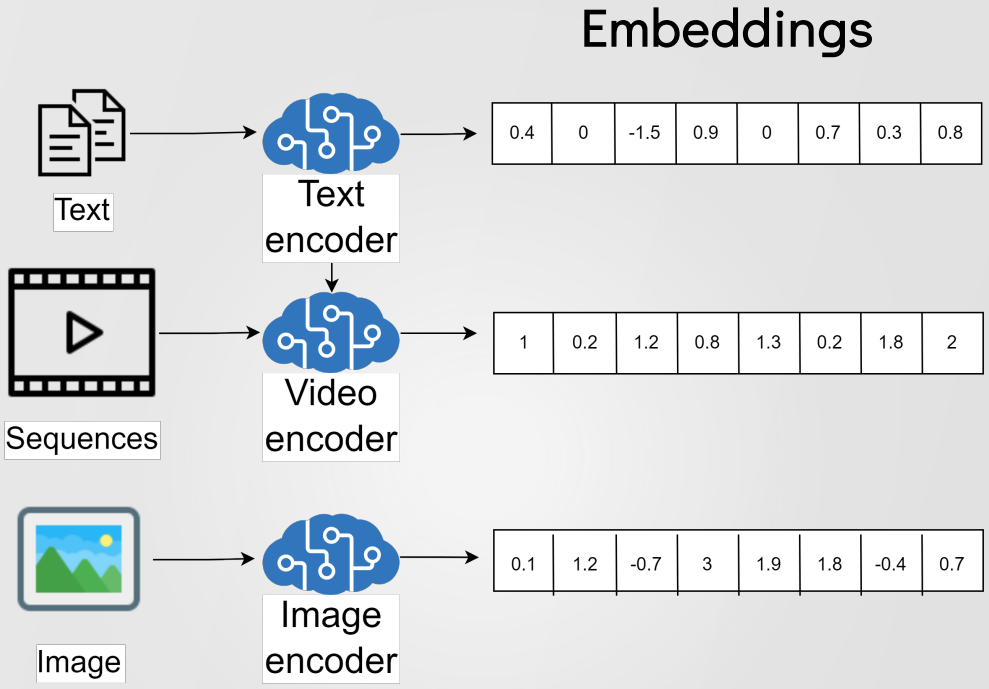
Multimodal Framework


Modalities
Image
Sequences
Text
Data Analysis
Data Cleaning
Data correlation
Feature selection
Data validation
Modality Encoding (which)
CNN
RNN
Transformer Encoder
Fusion Strategy (when)
Early
Late
Intermediate
Fusion Method (how)
Vector fusion
Addition
Product
MultConcat
Tensor fusion
.....
Modality Quantification (why)
EMAP
SHAP
DIME
How ?
Fusion method adapted
When ?
Fusion strategy adapted
Which ?
Encoders adapted
Why ?
Modalities quantified
70%
20%
10%
Image
Sequences
Text

Multimodal framework

Encoder choice (Which ?)

Modality	Deep learning architecture
Image	CNN , Transformers
Text	Transformers, LSTM
Audio	CNN
Video	CNN
Time-series	LSTM
Tabular data	MLP



Multimodal framework

Fusion Strategy (When ?)

Fusion Strategy	Pros	Cons
Early Fusion	- only requires to train a single model - less processing per modality	- sensitive to noise and missing data - requires synchronized data
Late Fusion	- combines outputs from individual models - robust to missing modalities - flexible in handling different data types	- loss of low-level feature interactions - potentially redundant computations - requires overall high unimodal accuracy for voting schemes.
Intermediate Fusion	- captures more low-level features than early fusion	- may require dimensions or values normalization

Remark :

- Fusion strategy choice will also determine which fusion method to use (for example , late fusion is adapted with voting schemes methods)

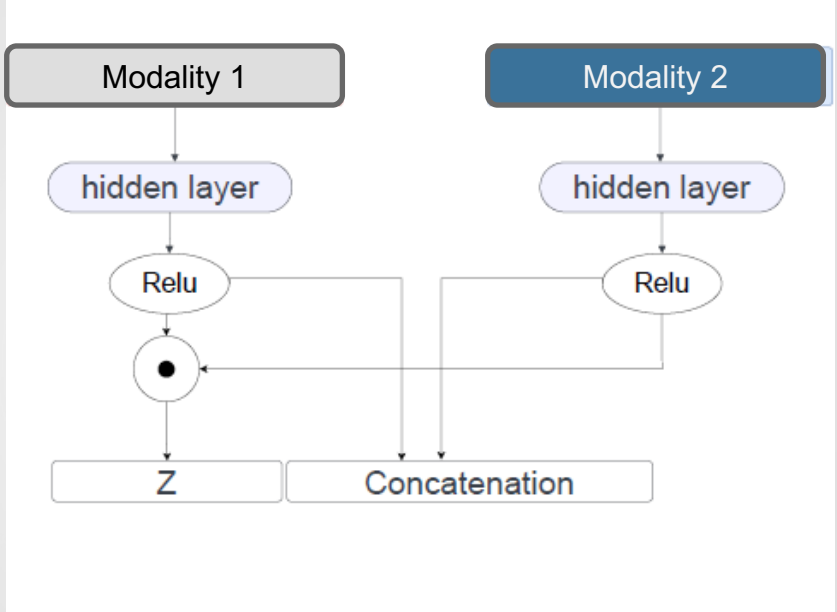
Multimodal framework

Fusion method (how ?)

Proposed approach : MutlConcat Fusion

https://arxiv.org/abs/2406.04349

- It is based on the integration of **two terms**: a concatenated representation of both modalities in a shared sub-space, and an element-wise multiplication of each of them.
- The intuition behind this separation is to **preserve modality-specific** features while concurrently extracting **cross-modal features**.
- Element-wise product between vectors preserves similarities (big * big = bigger) or the discrepancies (good review * bad review = bad review)



Multimodal framework

Fusion method (how ?)

Proposed approach : MutlConcat Fusion

https://arxiv.org/abs/2406.04349

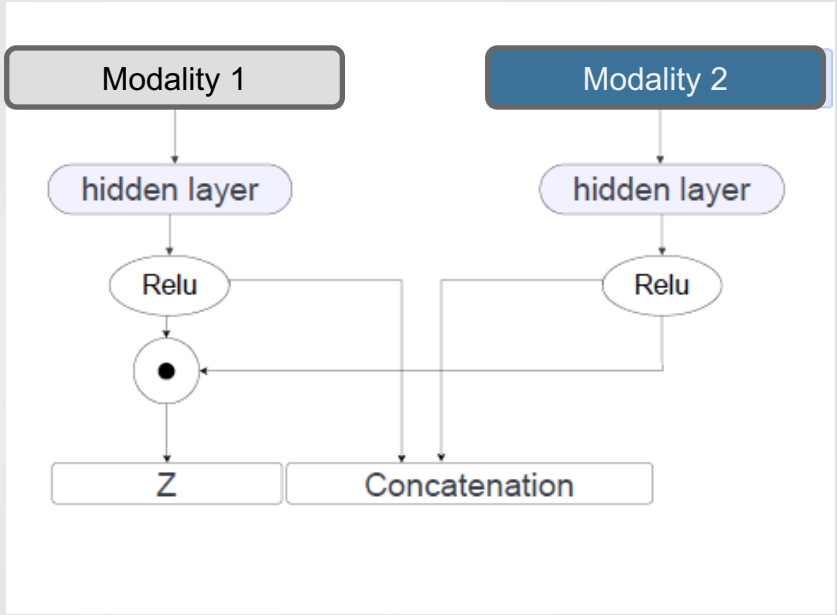
$$out_i = \mathrm{ReLU}(W_i \times M_i + b_i)$$

$$\mathrm{MultConcat} = C \parallel Z$$

$$C = \|_{i=1}^{N} out_i \qquad Z = \odot_{i=1}^{N} out_i$$

Such that :

- M_i represents Modality embedding
- Z represents cumulative element-wise products of N modalities



Multimodal framework

Fusion method (how ?)

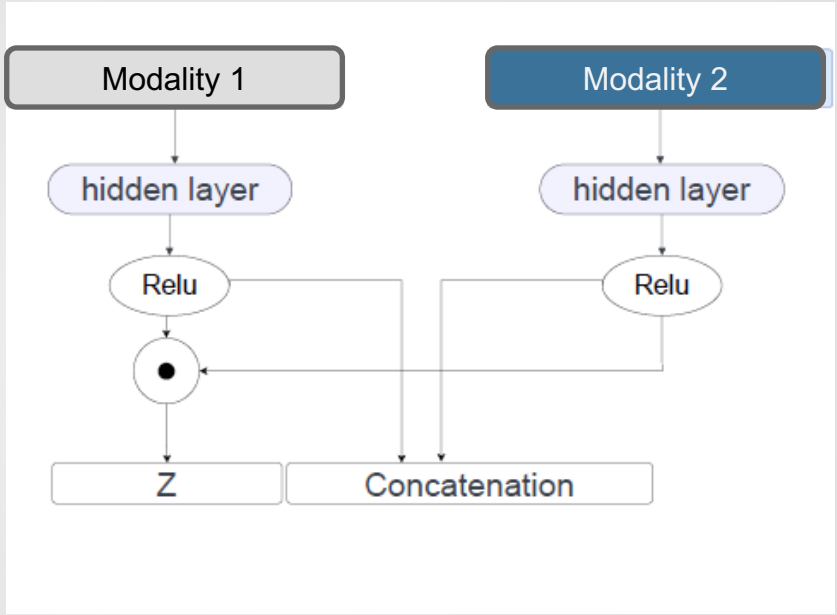
Proposed approach : MutlConcat Fusion

https://arxiv.org/abs/2406.04349

$$out_i = \text{ReLU}(W_i \times M_i + b_i)$$

$$\text{MultConcat} = C \parallel Z$$

$$C = \|_{i=1}^{N} out_i \qquad Z = \odot_{i=1}^{N} out_i$$



Advantages :

- Model-agnostic approach
- Can be adapted to multiple modalities
- Easy to implement

Multimodal framework

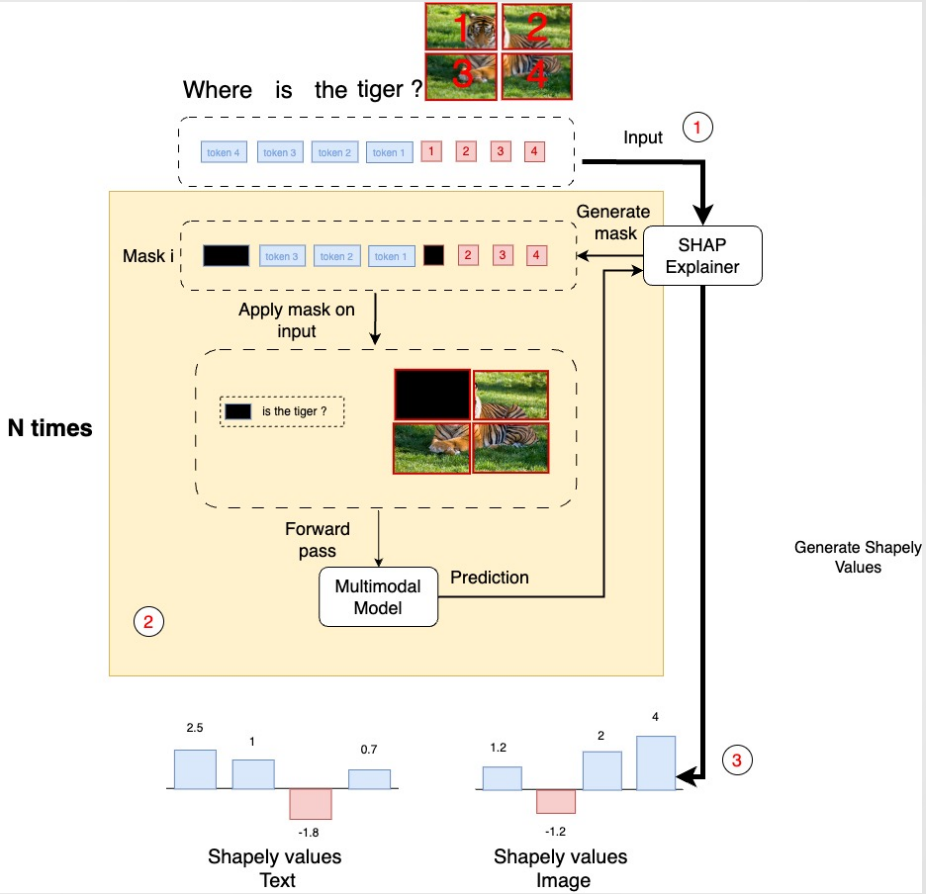
Modality Quantification (why ?)

MM SHAP

(L Parcalabescu,2023)

https://arxiv.org/abs/2212.08158

- An overview of SHAP applied to Vision-language models



04

Use case 1 : Multimodal HS Code Prediction

https://arxiv.org/abs/2406.04349

Use case 1

Use case 1 : multimodal HS code prediction

- In collaboration with eOrigin, the goal of the study was to develop smart tools to automate declaration validation
- Usually unimodal and treats a single data type at once
- Explore the effect of including other modalities on performance such as product image

Modalities

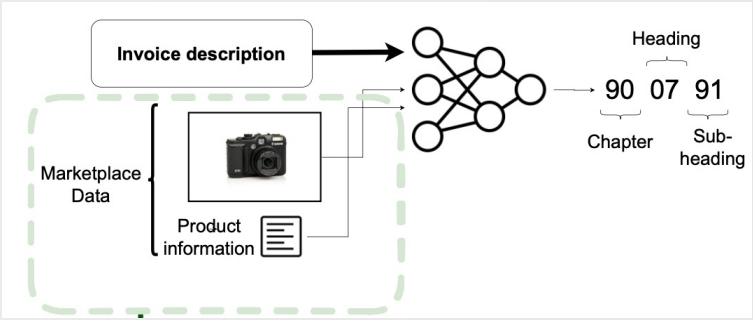
Images

Text(Custom déclarations, marketplace information)

Use case 1

Use case 1 : multimodal HS code prediction

Dataset : a total of 7593 customs declarations supplied by e-Origin having 16 unique HS6.



Modalities :

Invoice description (D) : the description given to a product

Product image (I): corresponding image in the marketplace

Title (T): product title in the market place

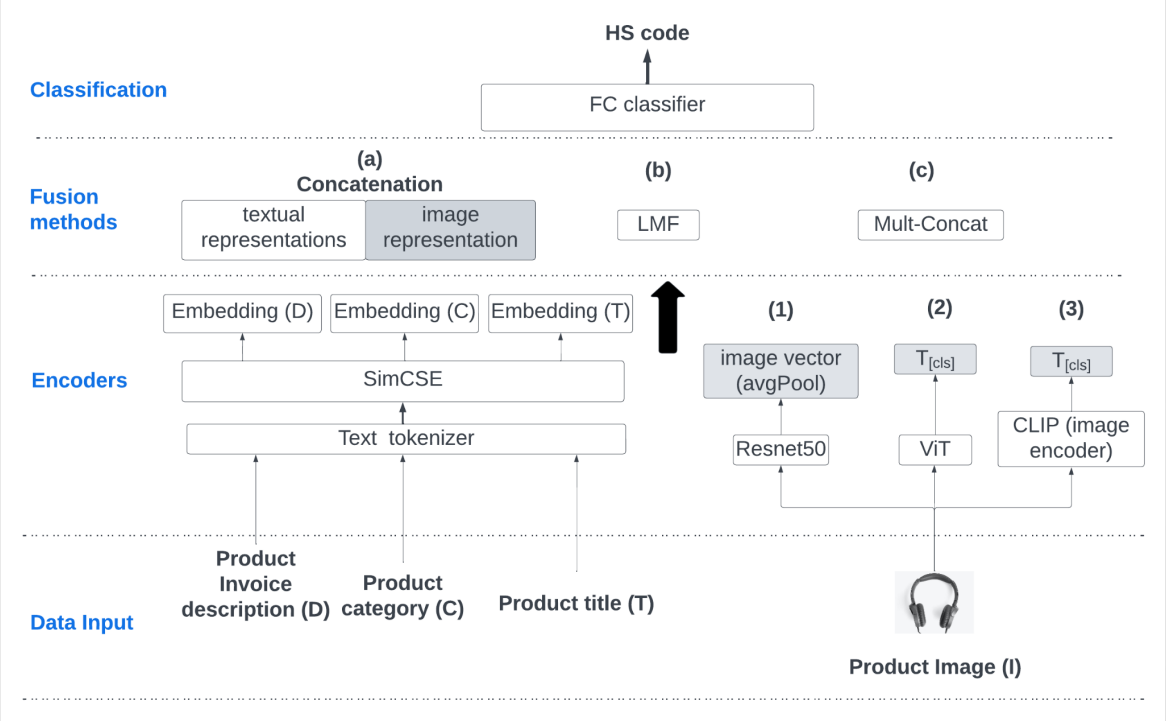
Category (C) : product category (low level / most specific) in the market place

Remark :

Some annotations are not validated : the model might be prone to data bias

Use case 1

Use case1: methodology architecture



Use case 1

e-Origin

Use case1: Results

Fusion method	Encoder		Modality	Top-k		
	Image	Text		k=1	k=3	k=5
MultConcat	ViT	SimCSE	I,T,D,C	**0.653**	0.929	0.977
Concat			I,T,D,C	0.624	0.924	0.977
LMF			I,T,D,C	0.088	0.188	0.347
MultConcat	ResNet50	SimCSE	I,T,D,C	0.612	**0.935**	**0.982**
Concat			I,T,D,C	0.571	0.924	0.977
LMF			I,T,D,C	0.047	0.182	0.241
MultConcat	CLIP	SimCSE	I,T,D,C	0.629	0.918	0.977
Concat			I,T,D,C	0.624	0.924	0.977
LMF			I,T,D,C	0.277	0.359	0.477
MultConcat	/	SimCSE	T,D,C	0.647	0.930	0.970
MultConcat	RestNet50	SimCSE	I,D	0.582	0.870	0.924
baseline (unimodal models)						
/	/	SimCSE	D	0.500	0.829	0.906
/	ViT	/	I	0.394	0.729	0.847
/	RestNet50	/	I	0.388	0.688	0.806
/	CLIP	/	I	0.482	0.806	0.894

- Suboptimal results with unimodal solutions (product description only)

- Our multimodal fusion approach outperforms the classical state of the art methods.

+10.6% of Top3 accuracy
compared to the unimodal model.

Use case 1

Use case1: modality quantification(why?)

Results of applied MM SHAP

Pair Type	Modality Scores		Correct Predictions		Incorrect Predictions	
	Text	Image	Text	Image	Text	Image
Image-Category	0.8887	0.1113	0.8892	0.1108	0.8816	0.1184
Image-Invoice Description	0.8807	0.1193	0.8839	0.1161	0.8295	0.1705
Image-Title	0.8759	0.1241	0.8752	0.1248	0.8864	0.1136

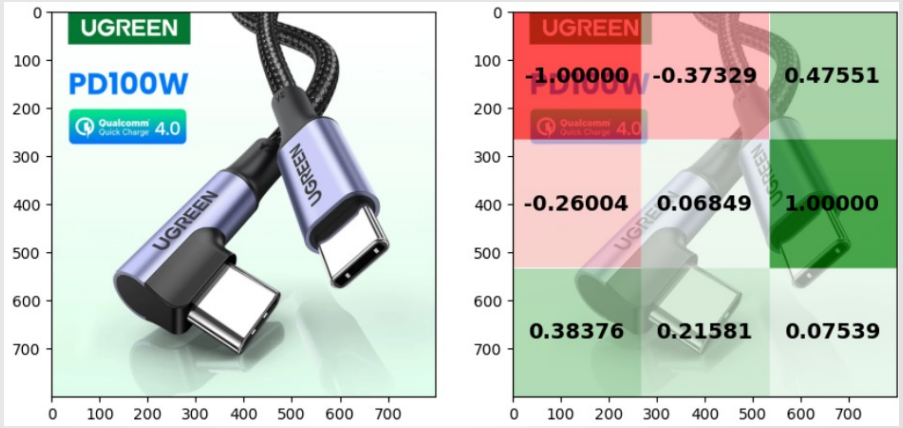
- Text has a higher contribution than images across all image-text pairs
- image modality can sometimes mislead the model
- It proves the effectiveness of the scoring methods since it goes in line with the obtained results

Remark :

- Only the specified pairs will be masked by mmshap since it is adapted for two inputs. The other modalities remain the same.

Use case 1

Use case1: modality quantification(why?)

	Invoice description(text)	Image product
	usb type usb cable	
Modality contributions	$\phi_{text} = 0.6160$	$\phi_{img} = 0.0245$
Contribution proportions	96%	4%

05

Use case 2 : RGB-D Action Recognition

https://doi.org/10.3390/electronics13122294

Use case 2

RGB-D Dangerous Action recognition

https://doi.org/10.3390/electronics13122294

- In collaboration with Infrabel, the goal of the study was to develop smart tools to monitor construction sites
- Deep learning base methods are promising methods
- Usually unimodal and treats a single data type at once
- Explore the effect of including other modalities on performance such as depth maps

Modalities

RGB Images

Depth maps

Use case 2

Use case 2 : Dataset

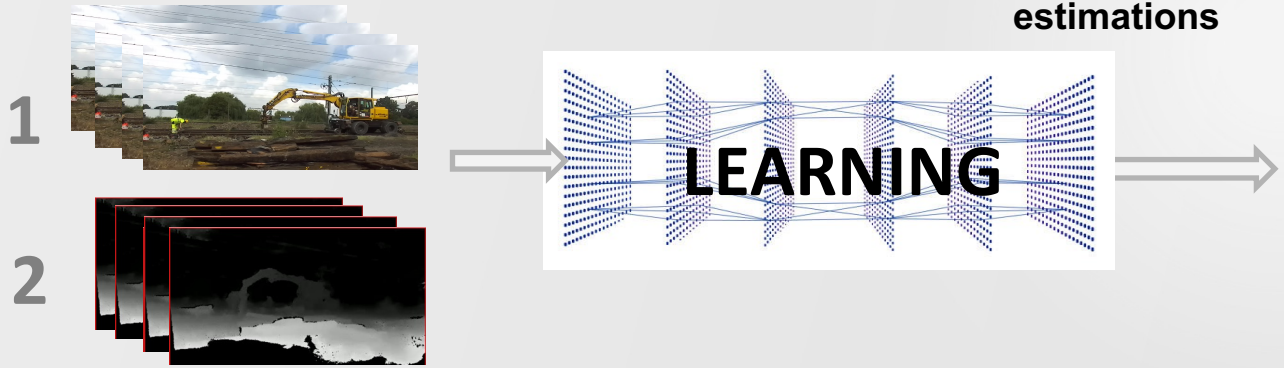
Dataset : a total of 1632 simulated dangerous actions in railways construction site provided by **Infrabel**.

- The samples were generated using Unity Game engine

Modalities :

RGB stream(sequence frame): A sequence of images represented in RGB

Depth map stream(generated) : A sequence of depth map frames containing depth estimations



Use case 2

Use case 2 : Dataset examples

Bucket-Worker	Cabin-Worker	Safe action	Track-Excavator
			
Worker moving under the bucket, In danger of getting hit, or materials may fall from the bucket	Worker moving too close to the cabin while the excavator is being operated	No dangerous actions	The excavator moving forward to the tracks (active railway line or electric wires)

Use case 2

Use case 2 : Depthmap generation example

Monodepth2 is a powerful tool to method to generate depth maps samples

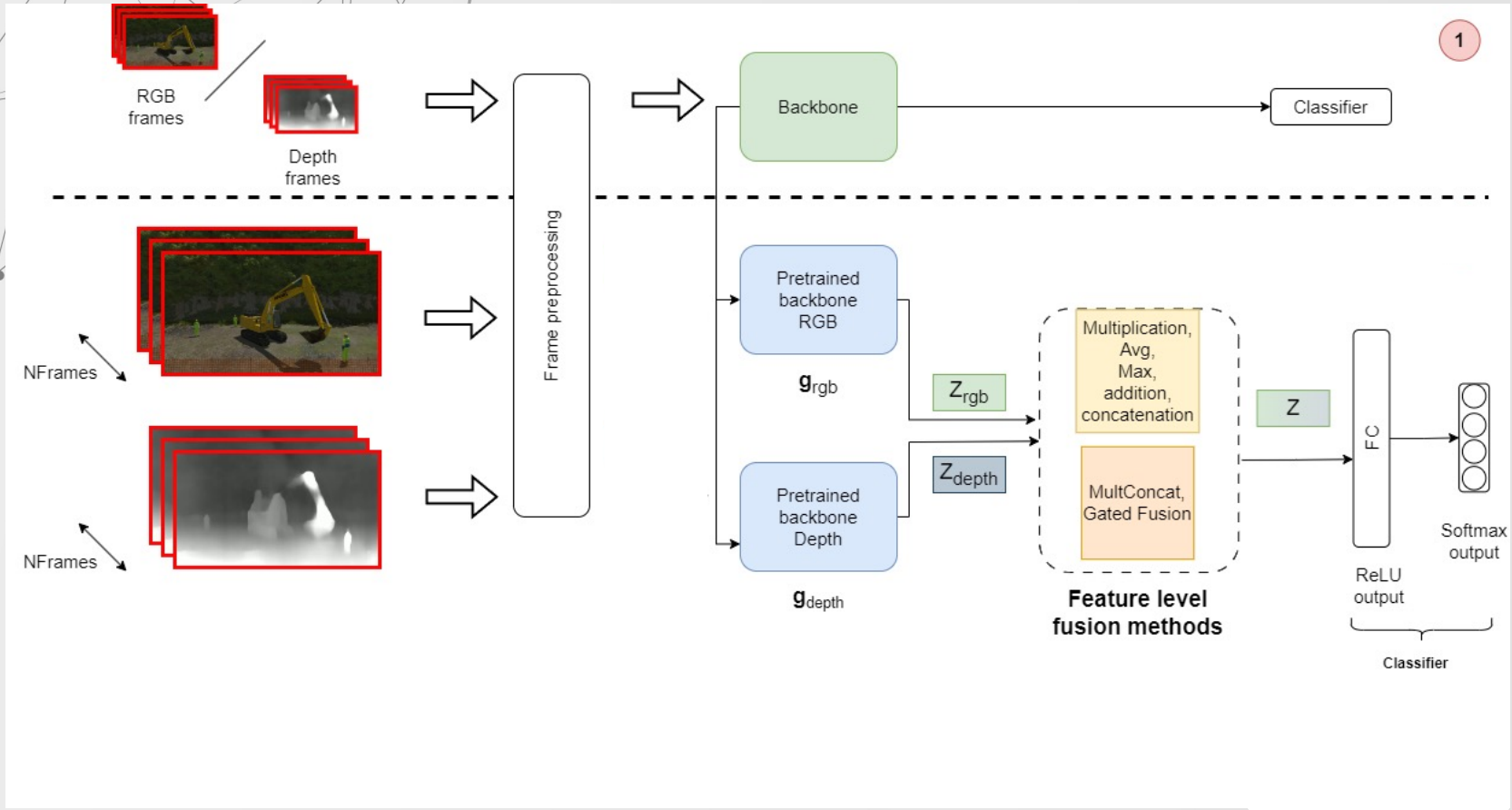
RGB frames

Generate using
monodepth2

Depth frames

Use case 2

Use case 2 : Methodology



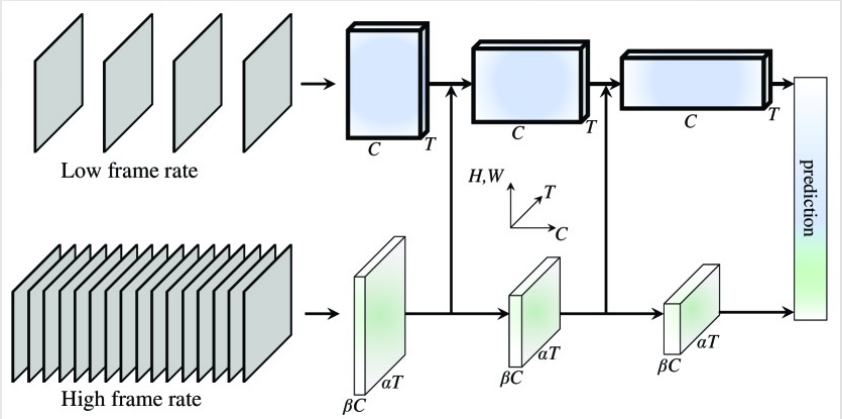
Use case 2

Use case 2 : Encoder used

Slowfast

•Two main components: Slow and Fast pathways for video recognition.

•Slow pathway captures spatial semantics at a low frame rate, while the Fast pathway captures fine temporal motion at a high frame rate.



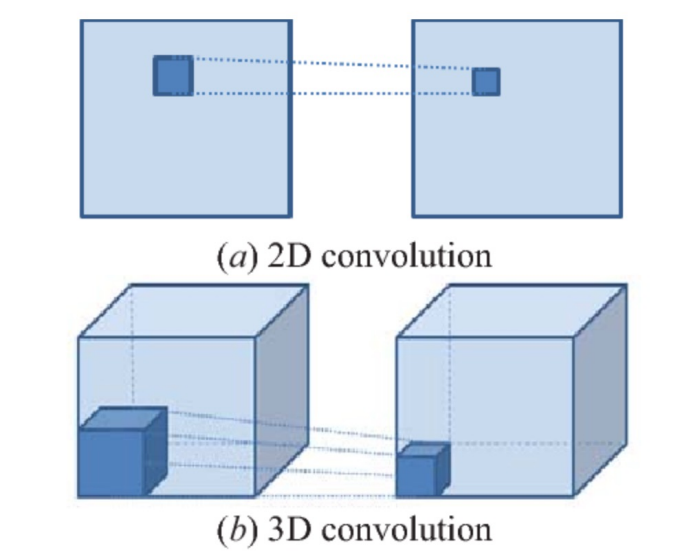
Use case 2

Use case 2 : Encoder used

C3D

•Uses **3D convolutional layers** (3D ConvNets) to capture spatial and temporal information.

• Straightforward architecture with repeated blocks of 3D convolutions followed by pooling layers, designed for fixed-length video segments, typically 16 frames.



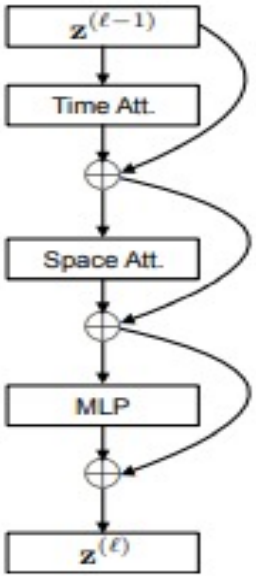
Use case 2

Use case 2 : Encoder used

Timesformer

•**Convolution-free** video classification using self-attention over space and time.

•**Transformer-based** architecture enables spatiotemporal feature learning from frame-level patches.



Divided Space-Time Attention (T+S)

Use case 2

Use case 2 : Results & Discussion

- SlowFast encoder outperforms C3D and TimeSformer, with MultConcat achieving the best results.

- MultConcat suits with SlowFast encoder but less with C3D and TimeSformer due to encoder dependency.

- RGB vs. Depth: Models perform better with RGB-only settings, indicating the importance of detailed appearance information.

https://doi.org/10.3390/electronics13122294

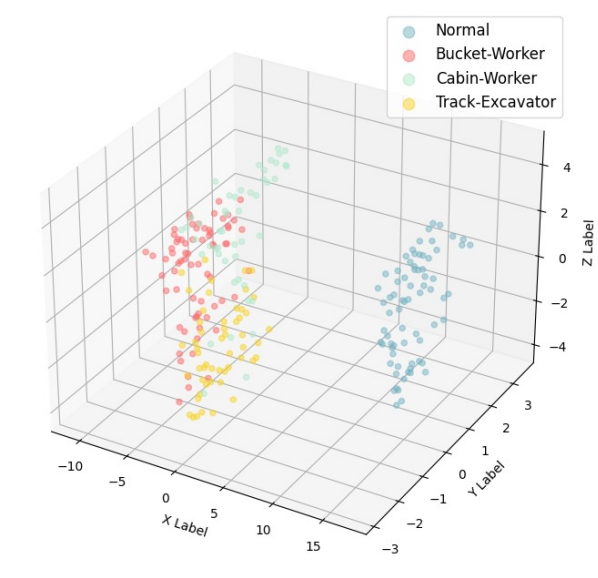
backbone	fusion method	modalities	test accuracy
C3d	Addition	RGB-D	0.741
	Concatenation	RGB-D	0.728
	Max	RGB-D	0.663
	Product	RGB-D	0.683
	Average	RGB-D	0.720
	GatedFusion [82]	RGB-D	0.22
	/	rgb only	0.806
	/	depth only	0.786
	MultConcat [81]	RGB-D	0.675
Timesformer	Addition	RGB-D	0.296
	Concatenation	RGB-D	0.296
	Max	RGB-D	0.399
	Product	RGB-D	0.428
	Average	RGB-D	0.383
	GatedFusion [82]	RGB-D	0.3868
	/	depth only	0.7572
	/	rgb only	0.7901
	MultConcat [81]	RGB-D	0.465
Slowfast	Addition	RGB-D	0.868
	Concatenation	RGB-D	0.860
	Max	RGB-D	0.860
	Product	RGB-D	0.860
	Average	RGB-D	0.876
	GatedFusion [82]	RGB-D	0.8477
	/	depth only	0.7984
	/	rgb only	0.8601
	MultConcat [81]	RGB-D	**0.893**

Use case 2

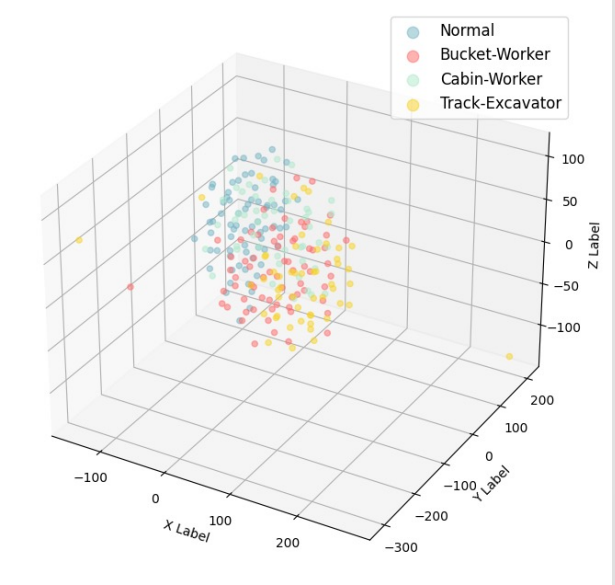
Use case 2 : T-SNE visualization

- RGB embeddings show overlapping clusters, making it hard to distinguish between dangerous action categories.

- Combining RGB and depth maps with MultConcat improves cluster separation (separated cluster for safe actions)



RGB-D embeddings



RGB embeddings

Use case 2

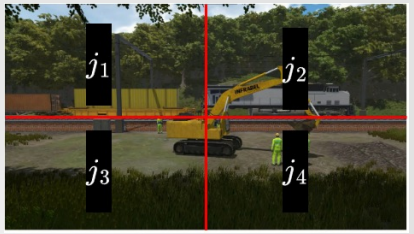
Modality Quantification (why ?)

INFR/A SECURE

Adaptation of MM SHAP to Rgb-d input (use-case 2)

$$\phi_j = \sum_{S \subseteq \{1,\ldots,n\} \setminus \{j\}} \frac{\text{val}(S \cup \{j\}) - \text{val}(S)}{\gamma}$$

Patches example :



Such that :

- S is a subset of all patch ids excluding j
- val(S) is the model's prediction with only the patches in subset S, in other words, the regions not masked.
- $\gamma = \frac{|S|!(n-|S|-1)!}{n!}$ is the normalizing factor accounting for all possible combinations of choosing subset S

Modality Contributions :

$$\Phi_{\text{rgb}} = \sum_{j=1}^{n_{\text{rgb}}} |\phi_j| \quad ; \quad \Phi_{\text{depth}} = \sum_{j=1}^{n_{\text{depth}}} |\phi_j|$$

Modality Proportions (in %) :

$$\text{rgb-SHAP} = \frac{\Phi_{\text{rgb}}}{\Phi_{\text{rgb}} + \Phi_{\text{depth}}} \quad ; \quad \text{depth-SHAP} = \frac{\Phi_{\text{depth}}}{\Phi_{\text{rgb}} + \Phi_{\text{depth}}}$$

Use case 2

Use case 2 : Modality quantification

Results of the adapted MM SHAP on Rgb-d input (use-case 2)

Class	Modality scores		Correct predictions		Incorrect predictions	
	RGB	Depthmaps	RGB	Depthmaps	RGB	Depthmaps
Bucket-Worker	0.5470	0.4530	0.5491	0.4509	0.5114	0.4886
Cabin-Worker	0.5585	0.4415	0.5638	0.4362	0.5445	0.4555
Track-Excavator	0.5538	0.4462	0.5505	0.4495	0.57	0.43
Safe Action (Normal)	0.5619	0.4381	0.5619	0.4381	/	/

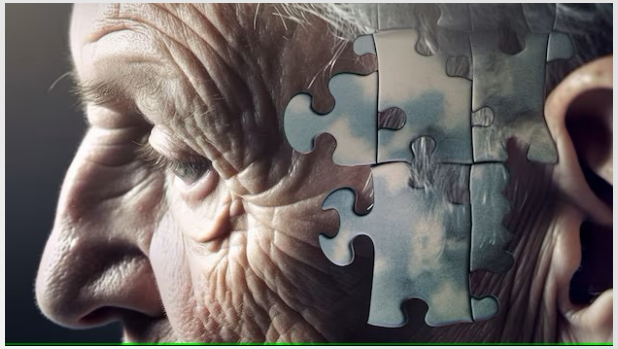
- RGB modality has a bit higher contribution than depthmaps across all classes
- Depth modality can sometimes mislead the model **(higher contribution for incorrect predictions)**
- It proves again the effectiveness of the scoring methods since it goes in line with the obtained results

06

AI4Brain use case overview

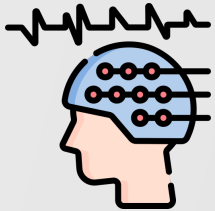
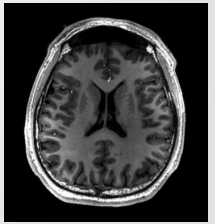
Use case 3 : AI4Brain-MedResyst

Alzheimer diagnosis using Deep learning



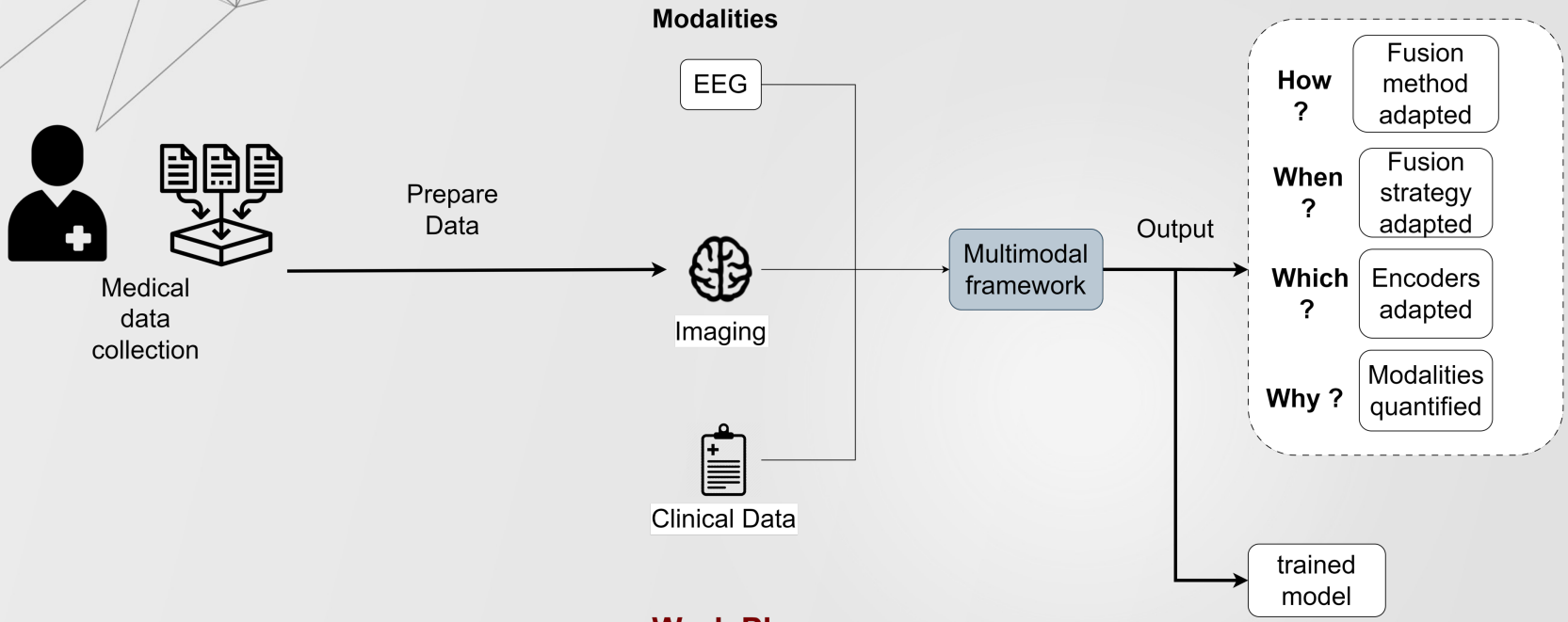
In collaboration with ISIA, Neuroscience department of UMons and CHU Ambroise Paré, the goal of the study is to develop smart tools to diagnose Alzheimer pathology based on patients' historical data

Modalities

EEG	Clinical data	MRI Imaging
		

Use case 3 : AI4Brain-MedResyst

Alzheimer diagnosis using Deep learning



Work Plan

07

Conclusion & Perspective

Conclusion

- Multimodal learning algorithms are of paramount importance to leverage the diversity of data types and sources.

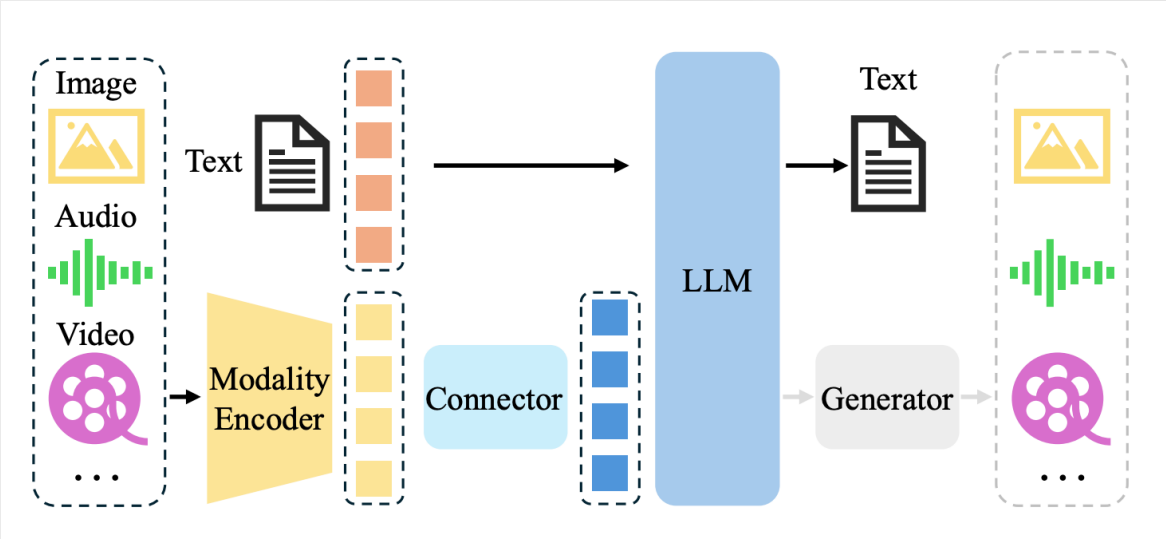
- Our proposed fusion method was proven to be effective for two use cases.

- As proven, it remains interesting to investigate modality contributions even if multimodal scores are better.

- The proposed multimodal framework (answering the which when how and why questions) offer enough flexibility for developing multimodal fusion models.

Perspective

- Adapt the provided framework to VLMS and multimodal LLM.



Yin, Shukang, et al. "A survey on multimodal large language models." *arXiv preprint arXiv:2306.13549* (2023).

THANKS

Do you have any questions?