Robust Multimodal Architecture: Towards an Executive System

Doctorant : Sébastien Grand
Laboratoire commun : RadaR-IO

IMT Mines Albi
École Mines-Télécom

EPSI
ENGINEERING AND PRODUCTS FOR SECURITY BY ITNI

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

IMT MINES ALBI

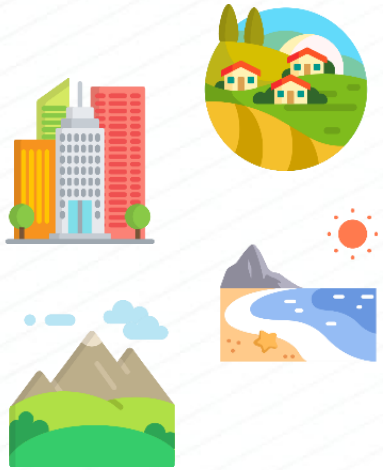
Aurélie Montarnal, *CGI Mines Albi*
Bruno Mériaux, *EPSI Radar*
Frédérick Bénaben, *ISyE Georgia Tech*
Guillaume Pouget, *CGI Mines Albi*
Charles Piffault, *CGI Mines Albi*

CONTEXT

Surveillance System

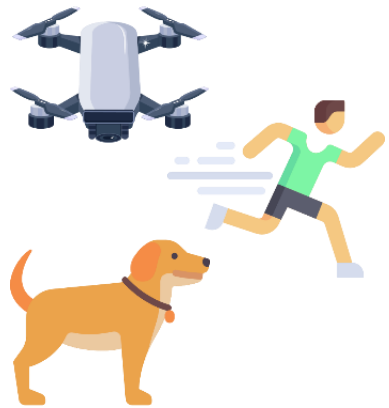
Radar

Different places



Different contexts

Different needs



Needs :

Generalisation capabilities under varying and dynamic contexts ;

Radar alone is not enough, need to use more sources ;

LIMITATIONS OF UNIMODAL MODELS

Radar
Unique data source

Only use one data source ;

If this source becomes unsable (sensors out-of-service, weather conditions, etc.), the model is not useful anymore ;

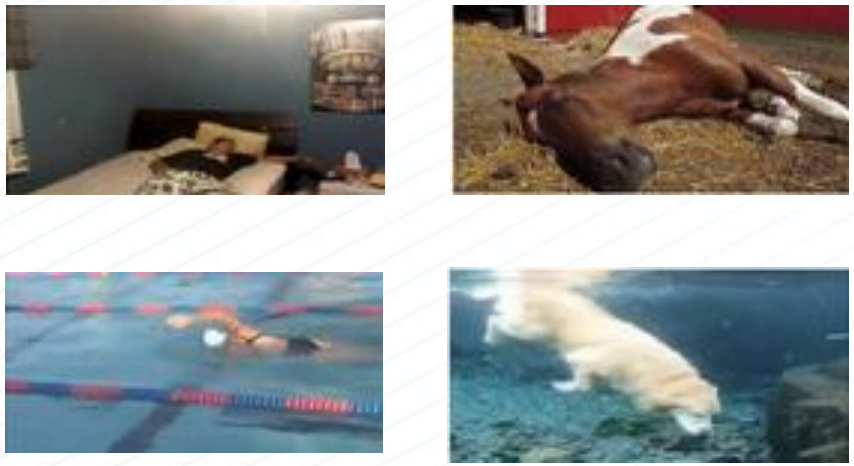
Unlike human brain that is multimodal, and can use multiple sources for an enhanced perception, unimodal model perceives the world from only one source ;

The case of radar data, interpretability is harder than RGB data ;

→ In short, a unimodal model does not fully capture the **overall context** and **information** present in the data. **Generalisation** to new cases (shifting from original distribution) is more difficult!

WHAT IS GENERALIZATION ?

Source domain
(humans / animals)



Training/Val

Target domain
(cartoon)

Test

What do we think to be a good approach :

- Use **other modalities** for an enhanced perception ;
- Learn rich and **domain invariant representations** using pretext task ;
- Add an **executive system** able to adaptively integrate the different representations ;

Source: Hao Dong and al. SimMMDG : A simple and effective framework for multi-modal domain generalization. Neurips, 2023.

GENERALIZE WITH MULTIPLE SOURCES

Radar

Camera

LiDAR

Complementary sensors:

- The camera provides rich visual detail;
- LiDAR provides 3D distance accuracy;
- Radar operates in all weather conditions;

Environmental robustness:

- The camera alone can be blinded by sunlight or darkness;
- LiDAR disrupted by rain;
- Radar sensitive to moving objects (vegetation).

Redundancy and reliability :

- If one sensor fails or is unreliable, the others can compensate.

PRETEXT TASKS TO GENERALIZE

A pretext task is an **auxiliary task**, often not directly useful in itself, but used to train a model to **learn useful representations**.

Learn to **align** different modalities.

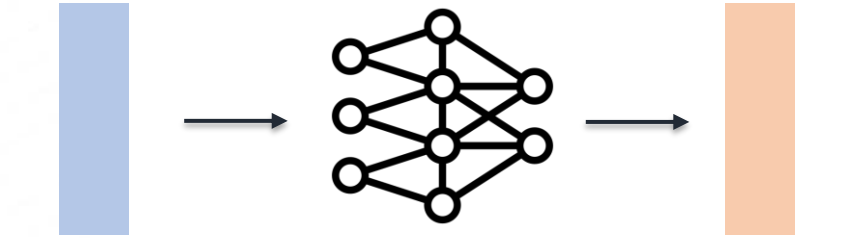
Learn **domain invariant** representations.

Better integrate **Multimodal Redundancy, Uniqueness and Synergy**.

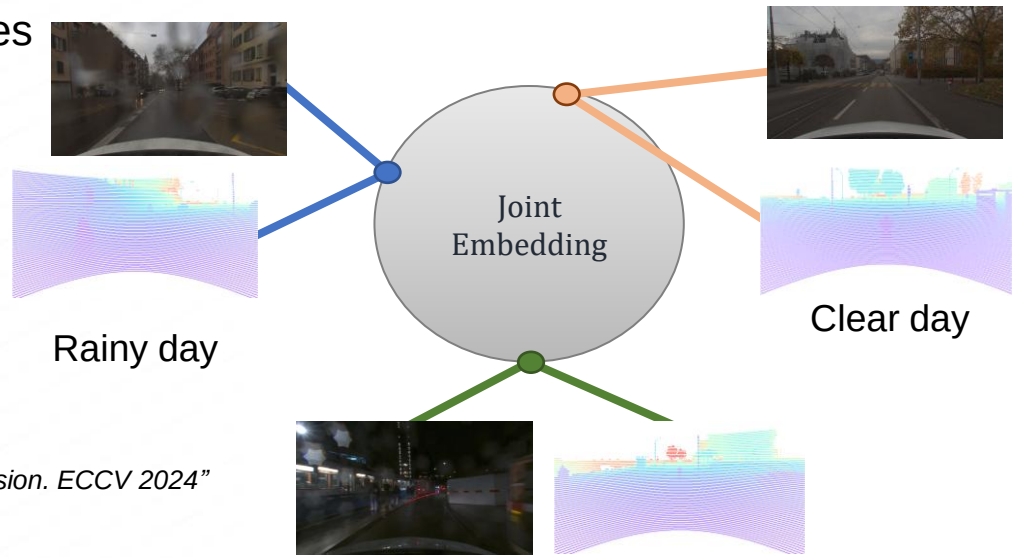
Greater generalization to new, never-before-seen areas (no need to re-train each time).

What are those tasks ?

Predicting the embedding of one modality from the embedding of another



Contrasting alignment of different modalities in a common space



Source : *Dong, Hao et al. "Towards Multimodal Open-Set Domain Generalization and Adaptation through Self-supervision. ECCV 2024"*
Dufumier, Benoit et al. "What to align in multimodal contrastive learning?" ICLR 2025

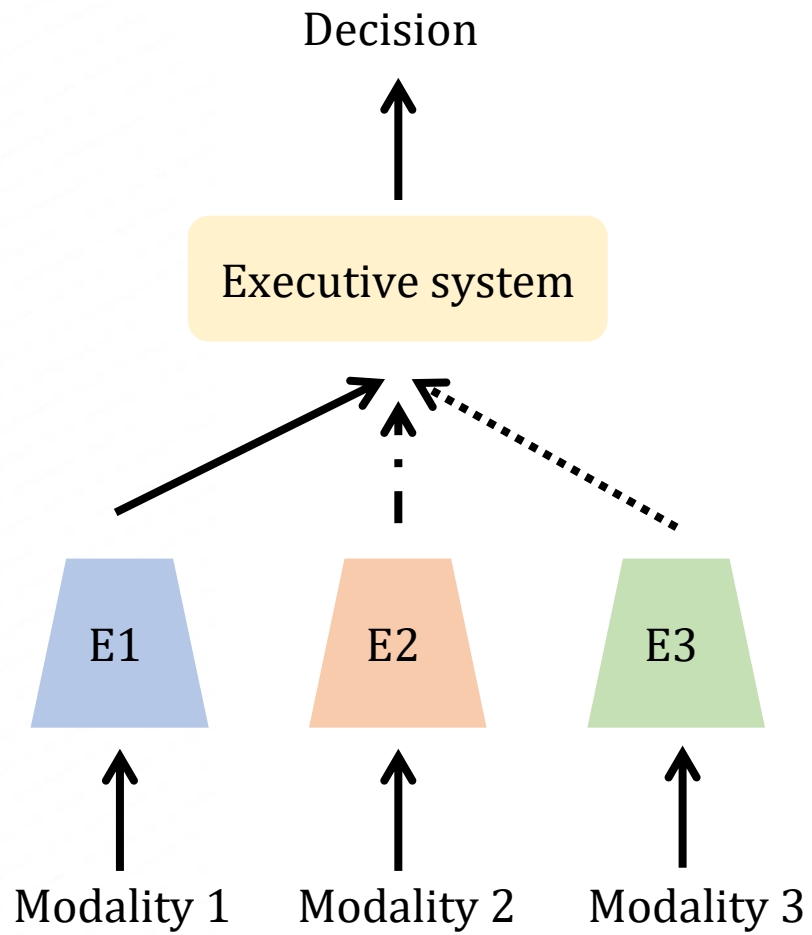
ADAPTIVELY FUSE CONTRIBUTIONS

An **executive system** assesses the **reliability** of each modality according to the context.
It **dynamically weights** their contributions and can **inhibit** some of them if they are judged to be irrelevant.

By night: radar and LiDAR provide a more reliable perception than the camera.
By day: the camera becomes the dominant modality thanks to its visual richness.

The executive system adapts the weight given to each modality in real time, directly influencing the final decision.

Key challenge: **avoiding unimodal bias**, when the system relies excessively on a single modality, to the detriment of overall robustness.



FUTURE WORKS

Analyse which pretext tasks leads to better domain generalization :

- Contrastive tasks ;
- Reconstruction tasks ;
- Predictive tasks ;

Analyse different executive system architectures :

- Symbolic rules (based on domain knowledge) ;
- Attention based fusion ;
- Uncertainty based fusion ;
- Transformer based fusion;

MERCI DE VOTRE ATTENTION

IMT Mines Albi
École Mines-Télécom

CENTRE
GÉNIE INDUSTRIEL