Dyadformer

A Multi-modal Transformer for Long-Range Modeling of Dyadic Interactions

David Curto, Albert Clapés, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, David Gallardo-Pujol, Georgina Guilera, David Leiva,Thomas B. Moeslund, Sergio Escalera, and Cristina Palmero

Presenter: Albert Clapés (postdoc AAU)

Slides by: Javier Selva

UNIVERSITAT DE BARCELONA

AALBORG UNIVERSITY

CVC Computer Vision Center

Motivation

- Growing interest in **social interaction understanding**.
- **Dyadic context** is crucial to understand both **individual behavior** and **interaction** characteristics.
- **Evaluation on UDIVA** v0.5!
- **Lacking** in competing methods:
  - **Long-range temporal** information for **self-reported personality** prediction.
  - **Joint** modeling of **both interlocutors.**
  - Use of **metadata** is uncommon.

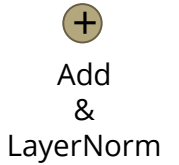
Source: https://pixabay.com/photos/people-girls-women-students-2557396/

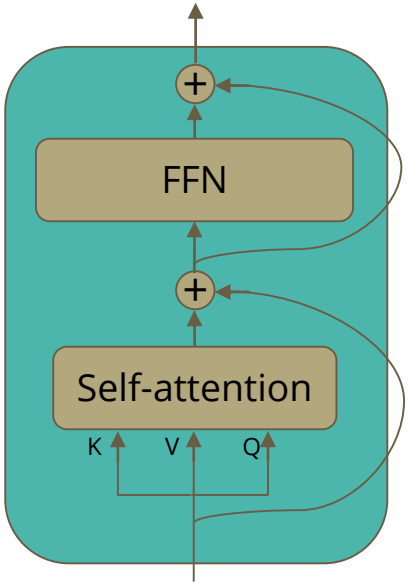
Contributions

- Purely data-driven method to **model** (and infer) **self-reported personality** in dyadic interactions.
- It is using larger temporal context w.r.t. previous works (**+30 seconds**).
- Leverage cross-attention to **jointly model both subjects** in the interaction.
- **New state-of-the-art results** on UDIVA v0.5.

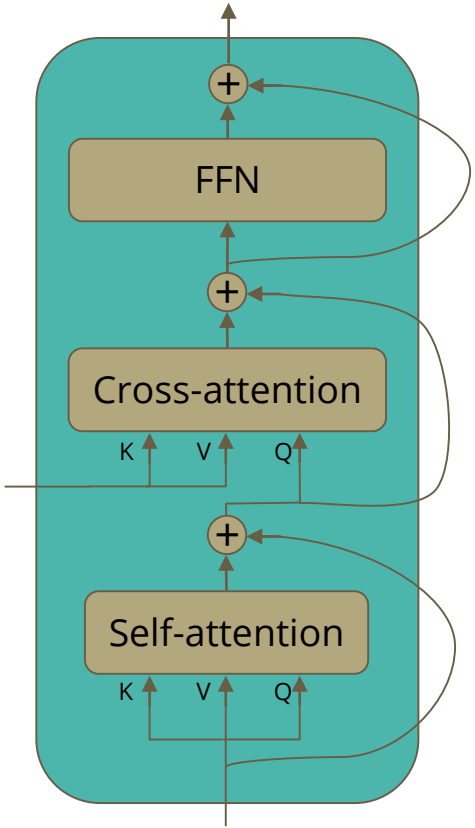
Self-attention and Cross-attention

- **Transformer**-based architecture
- Q, K, V **Attention**





Self-attention



Cross-attention

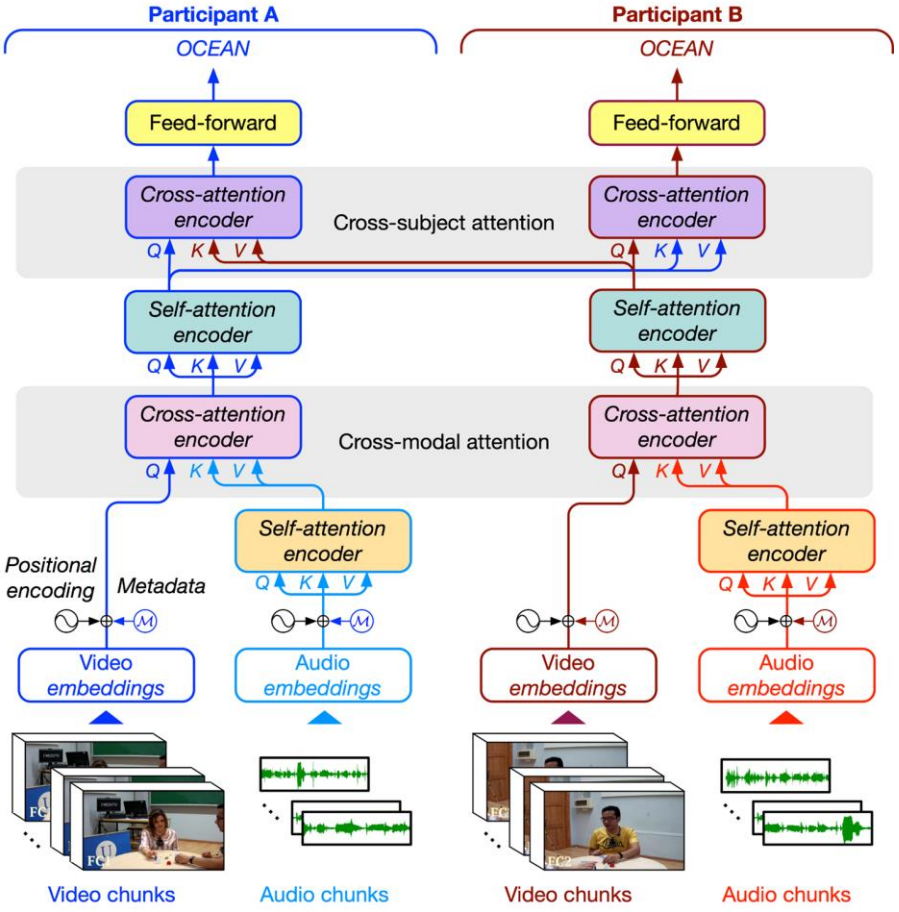
The Dyadformer

Input: sequences of **video & audio clip-level features.**

Embeddings: **pretrained R(2+1)D and VGGish.**

Add both **positional** and **metadata** information.

Weight sharing across **streams & intra-encoder layers**.



Input → Output

- Pre-segmented ~**3-second clips** of aligned **audio-video**.
- **T-clips sequences** as input (T ∈ {6,12}) resulting in **up to ~30 seconds**.
  - 67.5/10.9/5.6 hours (Train / Validation / Test)
- **Metadata**:
  - **Participant** (age, gender, cultural background, session index of participant, and pre-session mood/fatigue).
  - **Session** (task order within the session and its difficulty).
  - **Dyadic information** (relationship between interactants).

Self-reported OCEAN values as z-scores (standardized values) clipped the range [-3,3]:

- **O**pen-mindedness
- **C**onscientiousness
- **E**xtraversion
- **A**greeableness
- **N**egative emotionality

for **both subjects simultaneously.**

Dataset: UDIVA v0.5

Talk Lego Animals Ghost

2 views

- **Audiovisual** recordings of **face-to-face dyadic interactions**.
- **134 participants** (17-75 y.o., 55.2% male)
- Participant and session **metadata**
- **145** interaction **sessions** (4 tasks each)

- **Multilingual**: Spanish (73.1%), Catalan (17.25%) and English (9.65%)
- Includes **self-reported** OCEAN **personality scores**.
- Other labels.

Source: [2] **Palmero, C.** et al. "*Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset*". In WACVW (2021)

Evaluation

Both **loss** and **evaluation metric** are computed at sequence level and based on $\mathbf{MSE_{seq}}$.

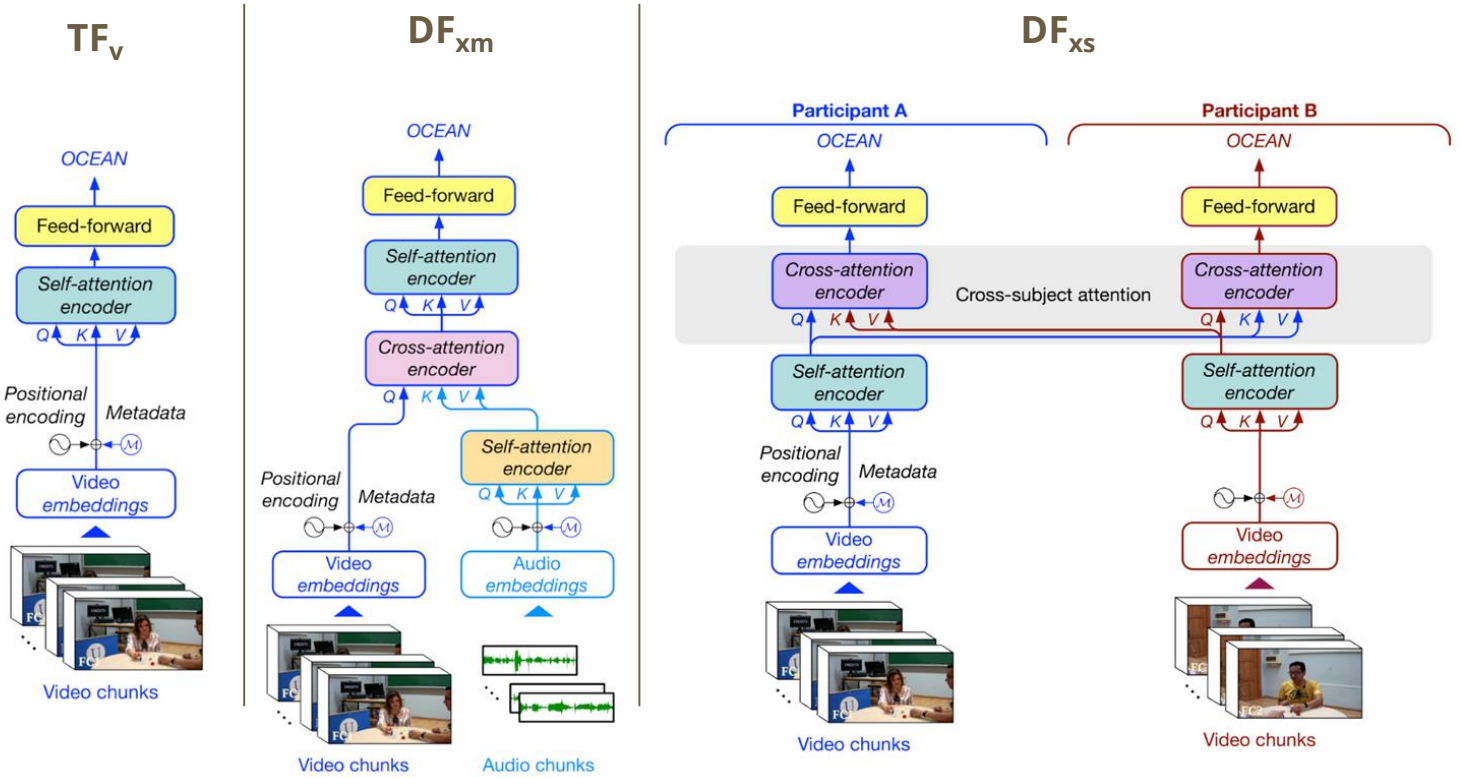
We also report $\mathbf{MSE_{part}}$ – the MSE at participant level – by aggregating (i.e. median) their OCEAN predictions before computing the error.

Pearson Correlation is measure correctness of predictions' distribution.

Experiments:

- **Ablation:** #layers, types of cross-attention, and temporal receptive field.
- **Comparison to SOTA**

Ablation


TF$_v$
OCEAN
Feed-forward
Self-attention encoder
Q K V
Positional encoding
Metadata
Video embeddings
Video chunks
DF$_{xm}$
OCEAN
Feed-forward
Self-attention encoder
Q K V
Cross-attention encoder
Q K V
Positional encoding
Metadata
Video embeddings
Self-attention encoder
Q K V
Audio embeddings
Video chunks
Audio chunks
DF$_{xs}$
Participant A
Participant B
OCEAN
OCEAN
Feed-forward
Feed-forward
Cross-attention encoder
Cross-subject attention
Cross-attention encoder
Q K V
Q K V
Self-attention encoder
Q K V
Self-attention encoder
Q K V
Positional encoding
Metadata
Video embeddings
Video embeddings
Video chunks
Video chunks

Ablation

Ablated:

- **L** = #layers in the encoders (sharing weights).
- **T** = #clips in the audiovisual sequences.
- **Arch.** = model architectures.

Findings:

- The more **temporal context** the better (T = 12).
- **Cross-modal & cross-subject (DF$_{xm,xs}$)** improve prediction.

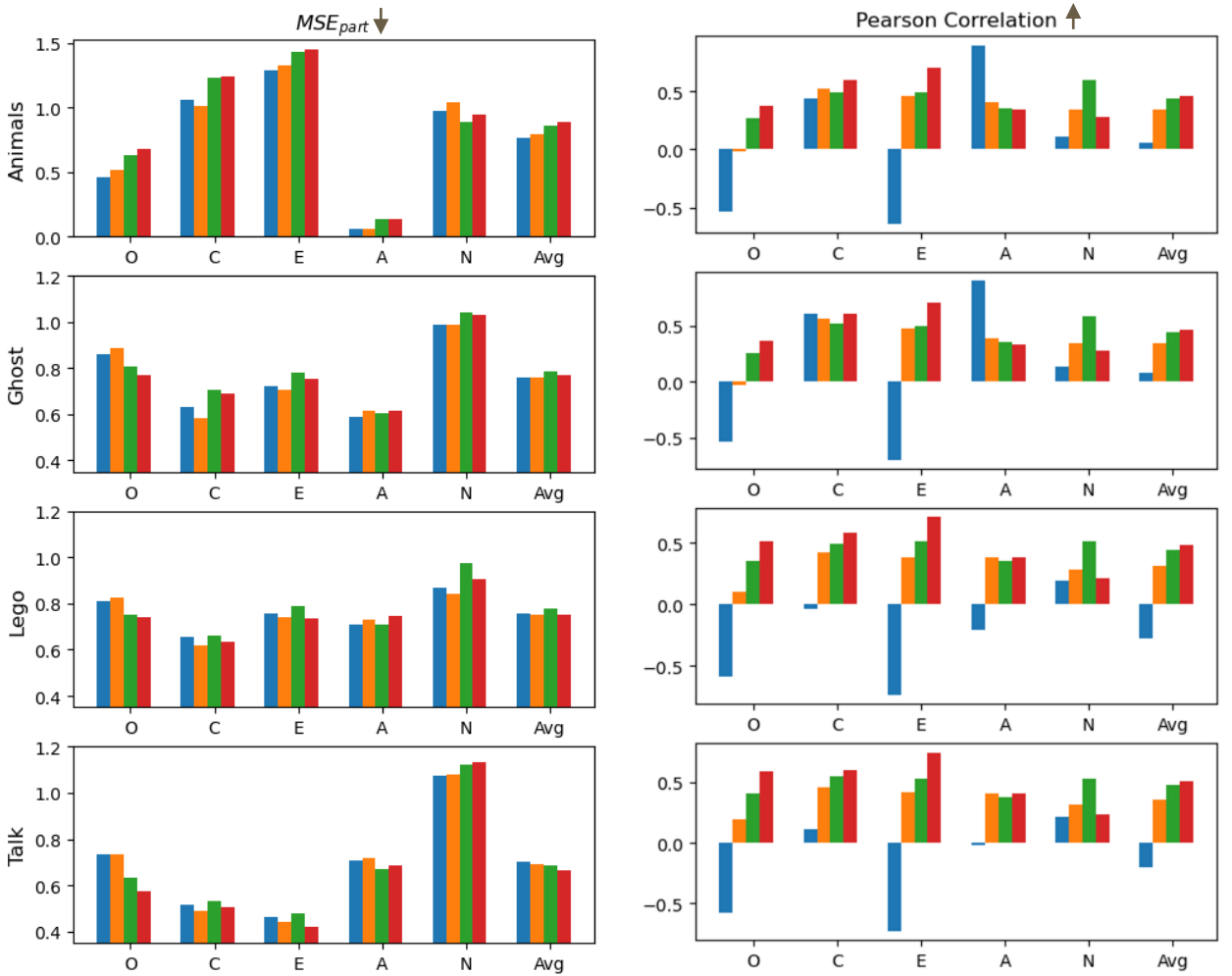
Arch.	L		MSE$_{seq}$	MSE$_{part}$	Params
				$T = 12$	
TF$_v$	2			0.732	10.0M
	4			0.744	
	6			0.807	
	8			0.804	
	L_{xm}	L_{xs}		$T = 12$	
DF$_{xm}$	1	-		0.732	19.4M
	2	-		**0.722**	
	3	-		0.762	
DF$_{xs}$	-	1		0.745	19.4M
	-	2		0.738	
	-	3		0.743	
DF$_{xm,xs}$	1	1		0.741	36.0M
	1	2		0.748	
	2	1		**0.722**	
	2	2		0.737	

T = #clips in the sequences; L = #layers in the encoders

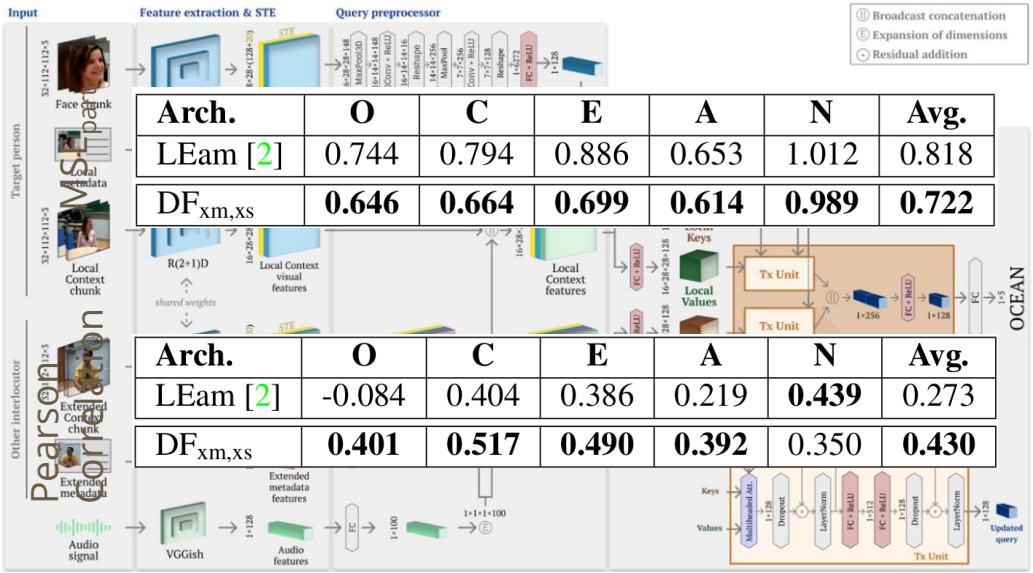
Ablation results per OCEAN trait and task

Different tasks enact **different traits**.

Models using **cross-subject interactions** show **better** Pearson **Correlation**.



SOTA COMPARISION



Arch.	O	C	E	A	N	Avg.
LEam [2]	0.744	0.794	0.886	0.653	1.012	0.818
$DF_{xm,xs}$	**0.646**	**0.664**	**0.699**	**0.614**	**0.989**	**0.722**

Mean Absolute Error

Arch.	O	C	E	A	N	Avg.
LEam [2]	-0.084	0.404	0.386	0.219	**0.439**	0.273
$DF_{xm,xs}$	**0.401**	**0.517**	**0.490**	**0.392**	0.350	**0.430**

Pearson Correlation

- In [2], we worked with **3-second chunks** (here, ~30 seconds).
- **Multi-modal fusion** was done by **concatenating** (here, self-/cross-attention).
- Regresses **personality of the target subject** (here, both simultaneously).

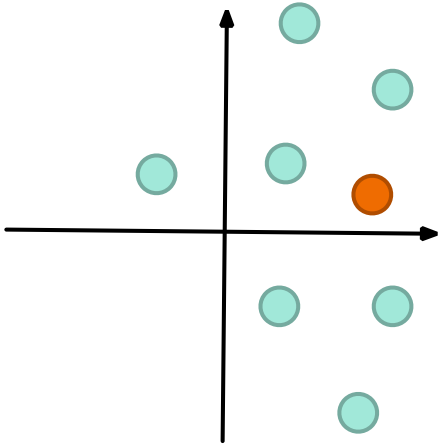
Source: [2] **Palmero, C.** et al. "*Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset*". In WACVW (2021)

Summary

- Successful data-driven architecture for **predicting self-reported personality on observed dyadic interactions**.
- **Multi-modal fusion** via cross-attention, i.e. cross-modal attention.
- Explicitly modeling **both subjects jointly**, i.e. cross-subject attention.
- Relatively **long temporal receptive fields (+30 secs)**.
- **SOTA on UDIVA v0.5**.

Future work

- Exploiting UDIVA v0.5 **dialogue transcripts**.
- Designing **self-supervised pre-text tasks**.
- **Transfer perceived personality** prediction model and/or **multi-task**.
- **Perceived personality as a proxy for self-reported personality**, Then, learn a mapping from perceived -> self-reported.
- **Contrasting interaction behaviors** of the same person with different interlocutors.



Questions?

Open to collaboration ☺

mailto: alcl@create.aau.dk

Other interests:

- Video-related analyses and/or multi-modal (incl. video) -> sequential models (e.g., Transformers).
- XAI and active supervision in vision through language.
- Learning from small/noisy data (i.e., self-supervision, self-learning, knowledge distillation, etc).
- **Applications**: Affective, assistive, edge, and personalized AI.