Attending to Emotional Narratives

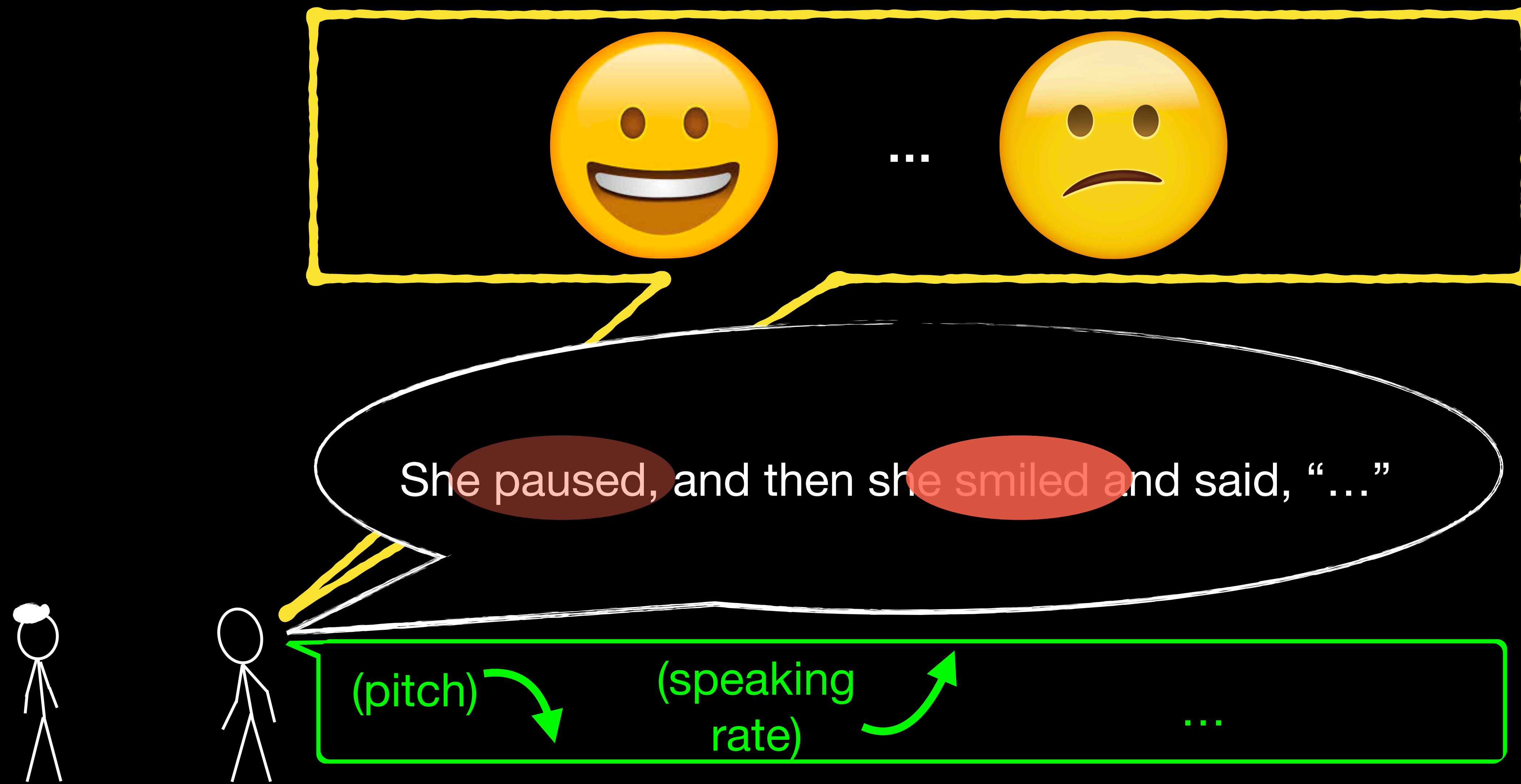
Desmond Ong
National University of Singapore
& A*STAR Singapore

Together with:
Zhengxuan Wu[1], Xiyu Zhang[1], Zhi-Xuan Tan[2], Jamil Zaki[1]
[1]Stanford University & [2]A*STAR

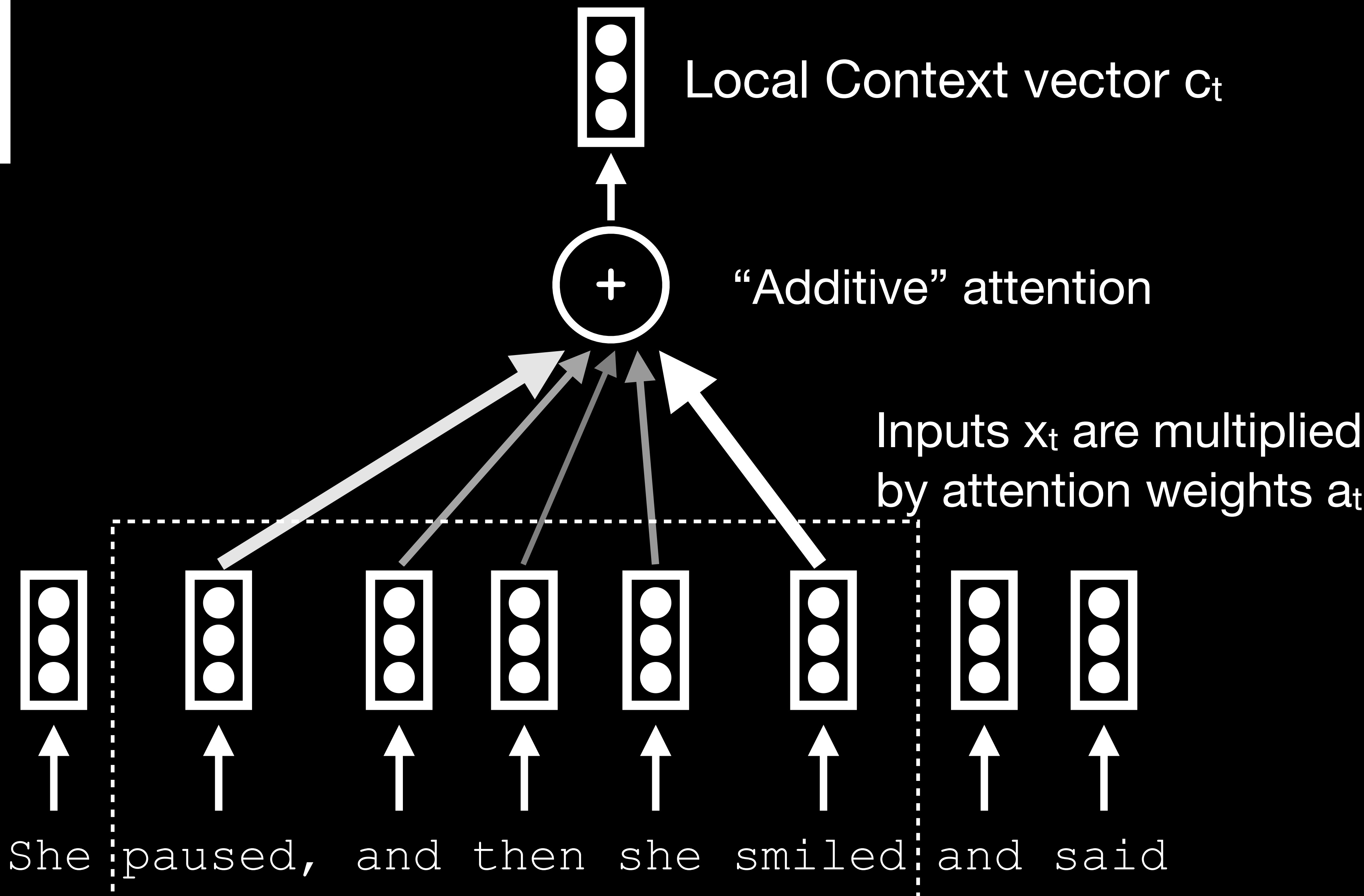
Agency for Science, Technology and Research
SINGAPORE

NUS National University of Singapore | School of Computing

Human emotion reasoning



Neural Network "Attention"



Research Question:

Can these attention mechanisms improve multimodal emotion recognition?

Luong, Pham & Manning (2015); Bahdanau, Cho & Bengio (2015)

The Stanford Emotional Narratives Dataset (SEND)

Volunteers describing emotional life events.

This **first release (SENDv1)** contains:

- 49 unique "targets"
- N=193 video clips,
- ~2 mins each, total 7 hrs 15 mins.
- 60:20:20 Train/Valid/Test spilt

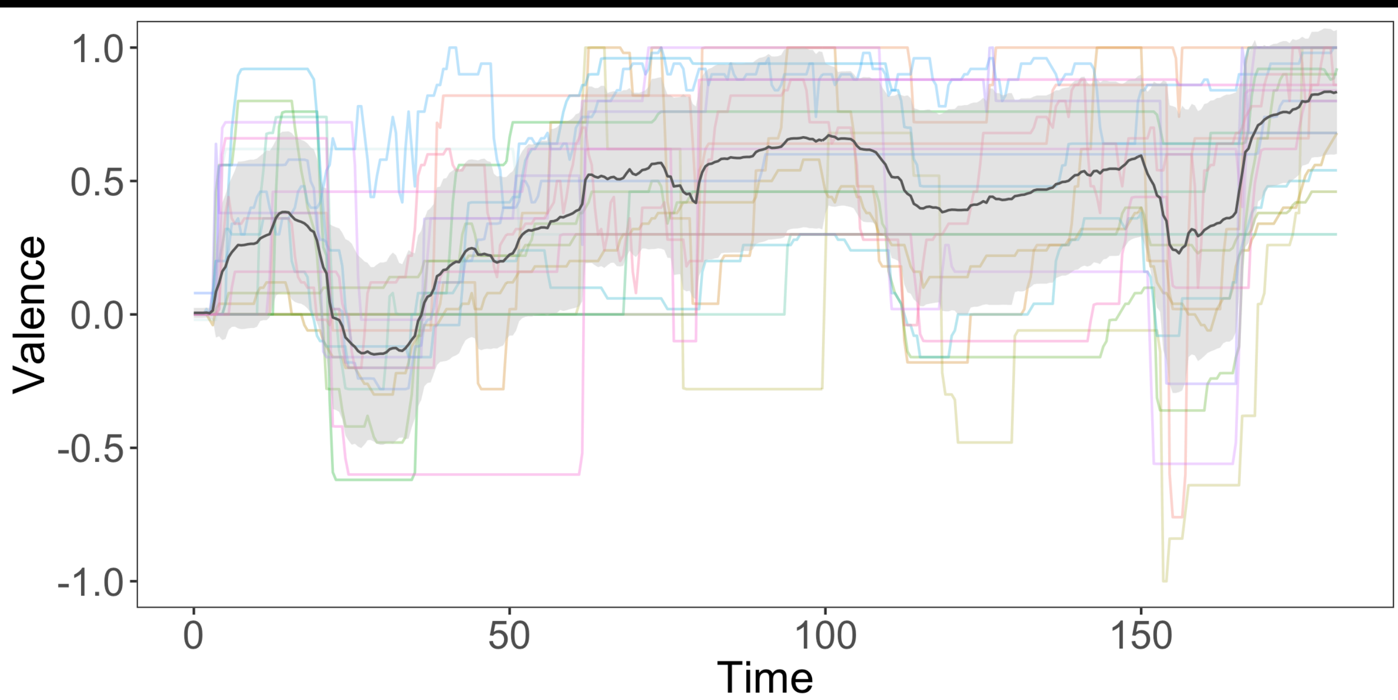
Each clip annotated by ~20 observers (on Amazon Mechanical Turk)
Annotated for emotional valence, sampled every 0.5s.

Gold-Standard Labels: **Evaluator-Weighted Estimator** (Grimm et al, 2007)

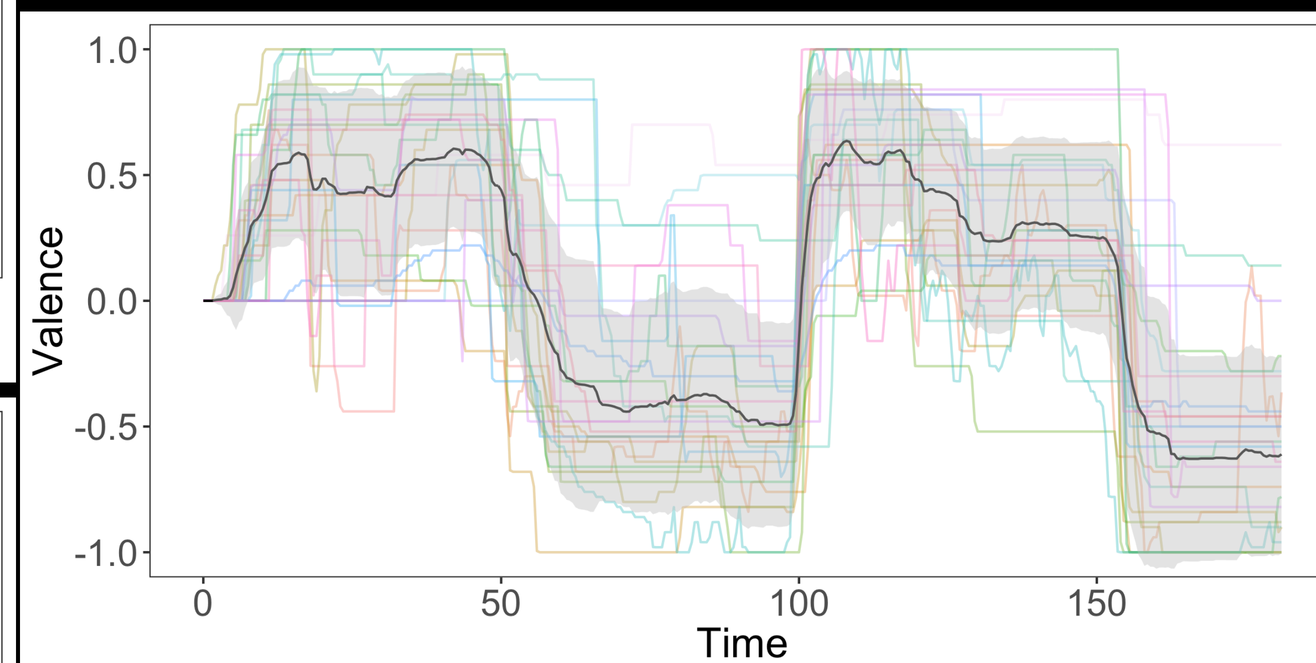
Evaluation Metric: Concordance Correlation Coefficient with the EWE

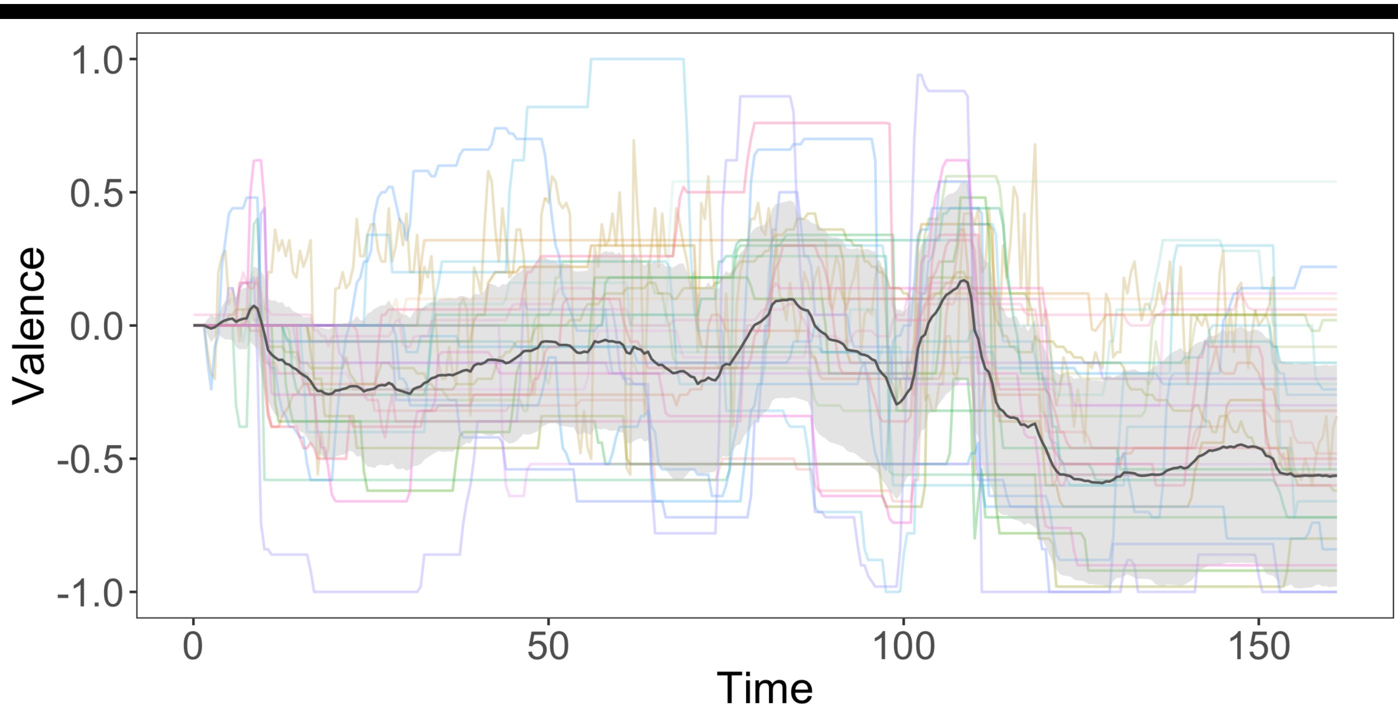
Ong, D. C., Wu, Z., Zhi-Xuan, T., Reddan, M., Kahhale, I., Mattek, A., & Zaki, J. (invited revision). Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset.

The Stanford Emotional Narratives Dataset (SEND)



- heterogeneous
- complex emotional trajectories





Ong, D. C., Wu, Z., Zhi-Xuan, T., Reddan, M., Kahhale, I., Mattek, A., & Zaki, J. (invited revision). Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset.

The Stanford Emotional Narratives Dataset (SEND)

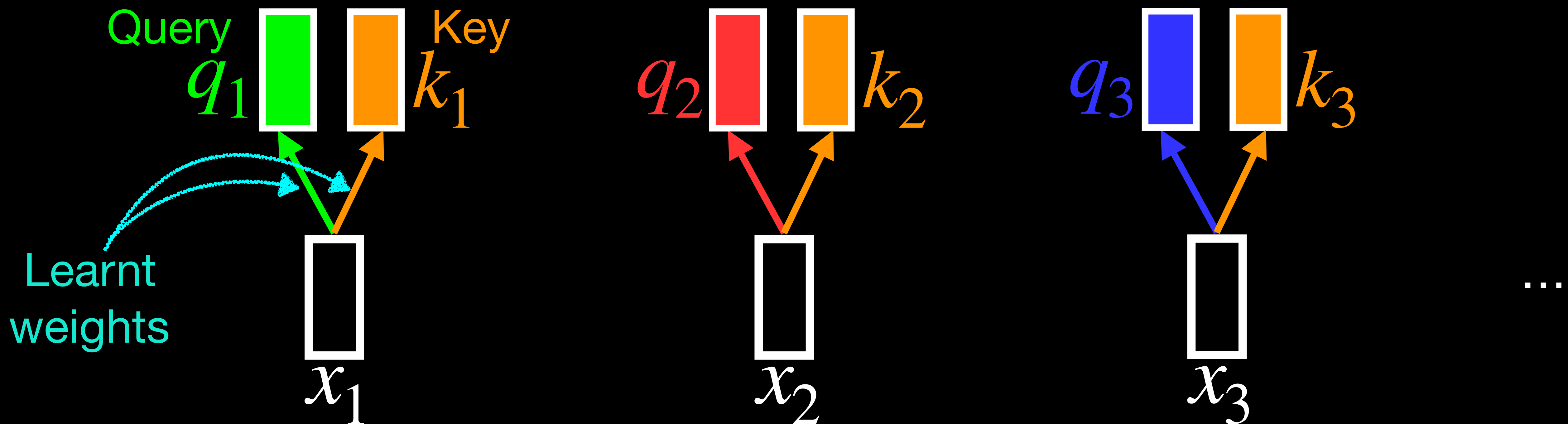
Summary

- Multimodal (Video, Audio, Text)
- Unscripted, naturalistic expressions
- Large diversity of stories
- Continuous over time (time-series)
- Dimensional labels
- Multiple annotations —> calculate reliable estimate

Ong, D. C., Wu, Z., Zhi-Xuan, T., Reddan, M., Kahhale, I., Mattek, A., & Zaki, J. (invited revision). Modeling emotion in complex stories: the Stanford Emotional Narratives Dataset.

Transformers (1)

State-of-the-art in NLP
Introduces the concept of "self-attention"

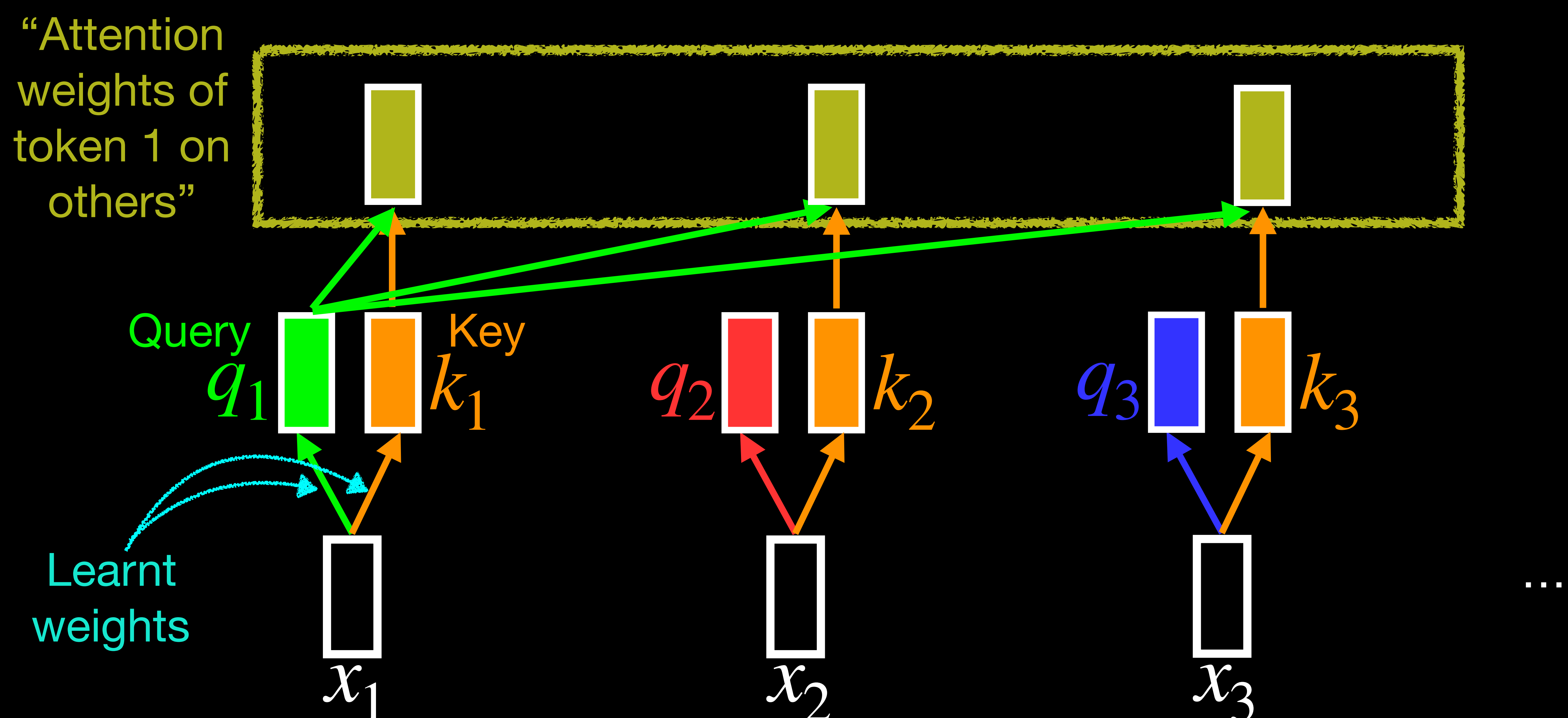


Input token (Feature vector at time t_1)

Vaswani et al (2017) Attention is all you need, *NeurIPS*

Transformers (1)

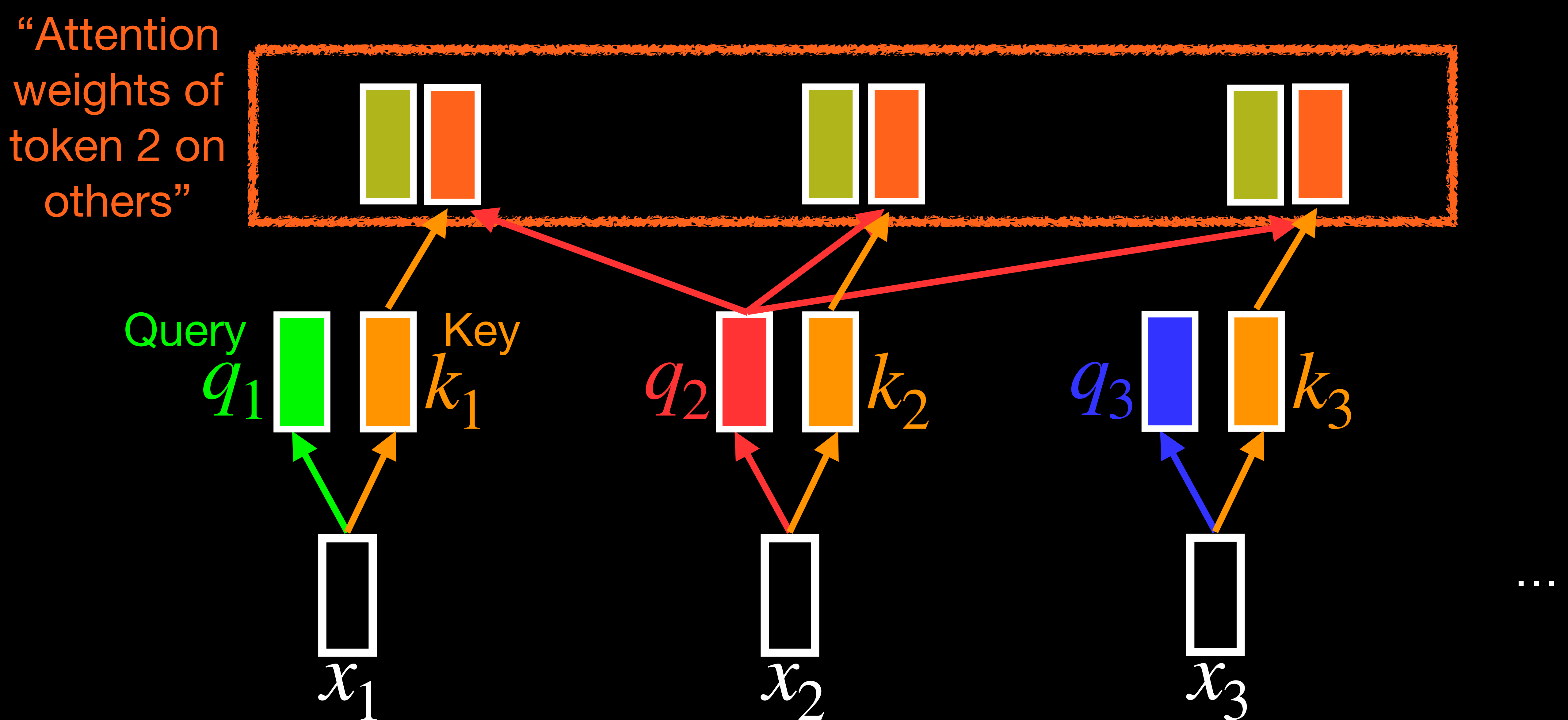
State-of-the-art in NLP
Introduces the concept of "self-attention"



Input token (Feature vector at time t_1)

Vaswani et al (2017) Attention is all you need, *NeurIPS*

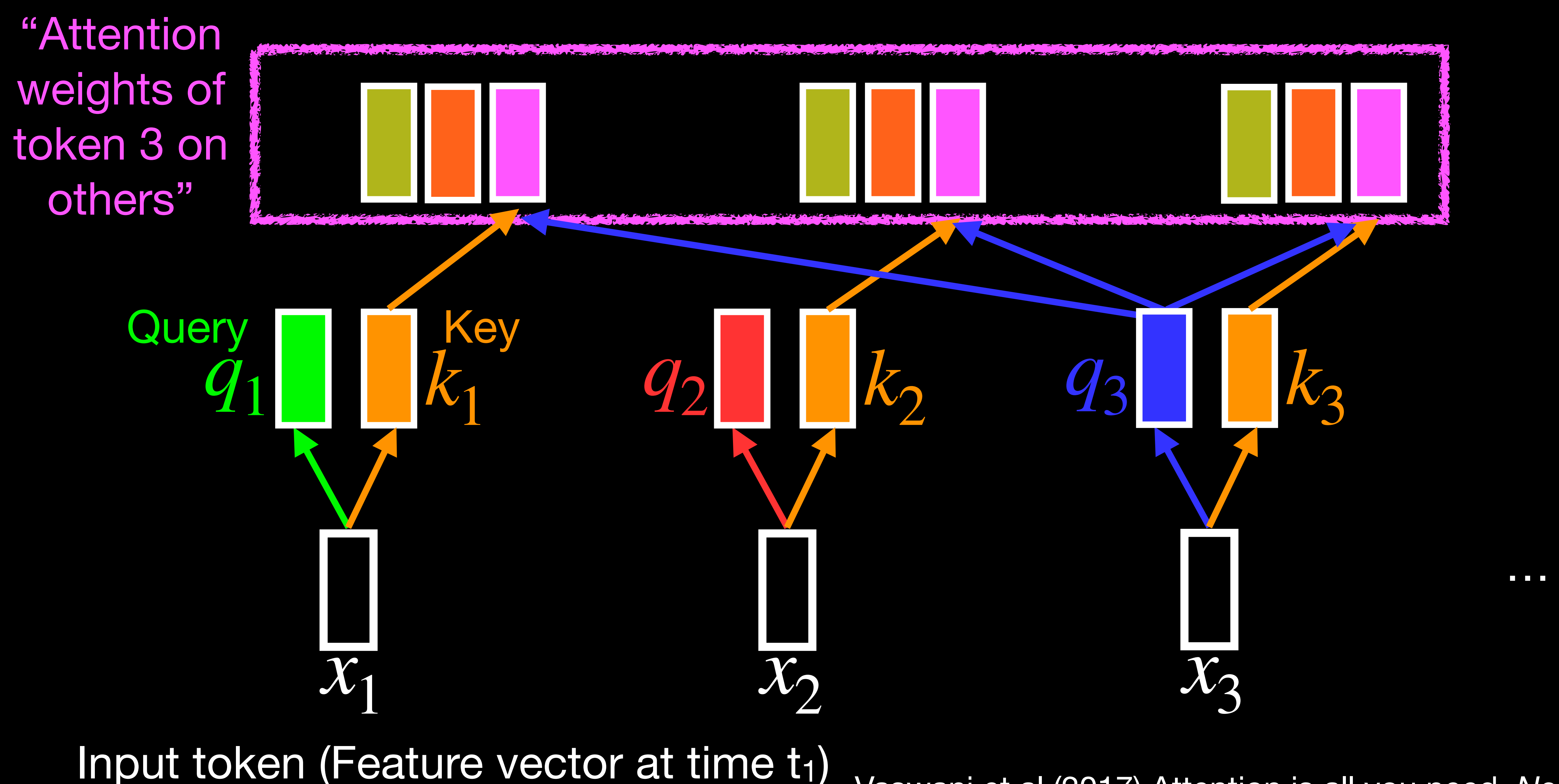
Transformers (2)



Input token (Feature vector at time t_1)

Vaswani et al (2017) Attention is all you need, *NeurIPS*

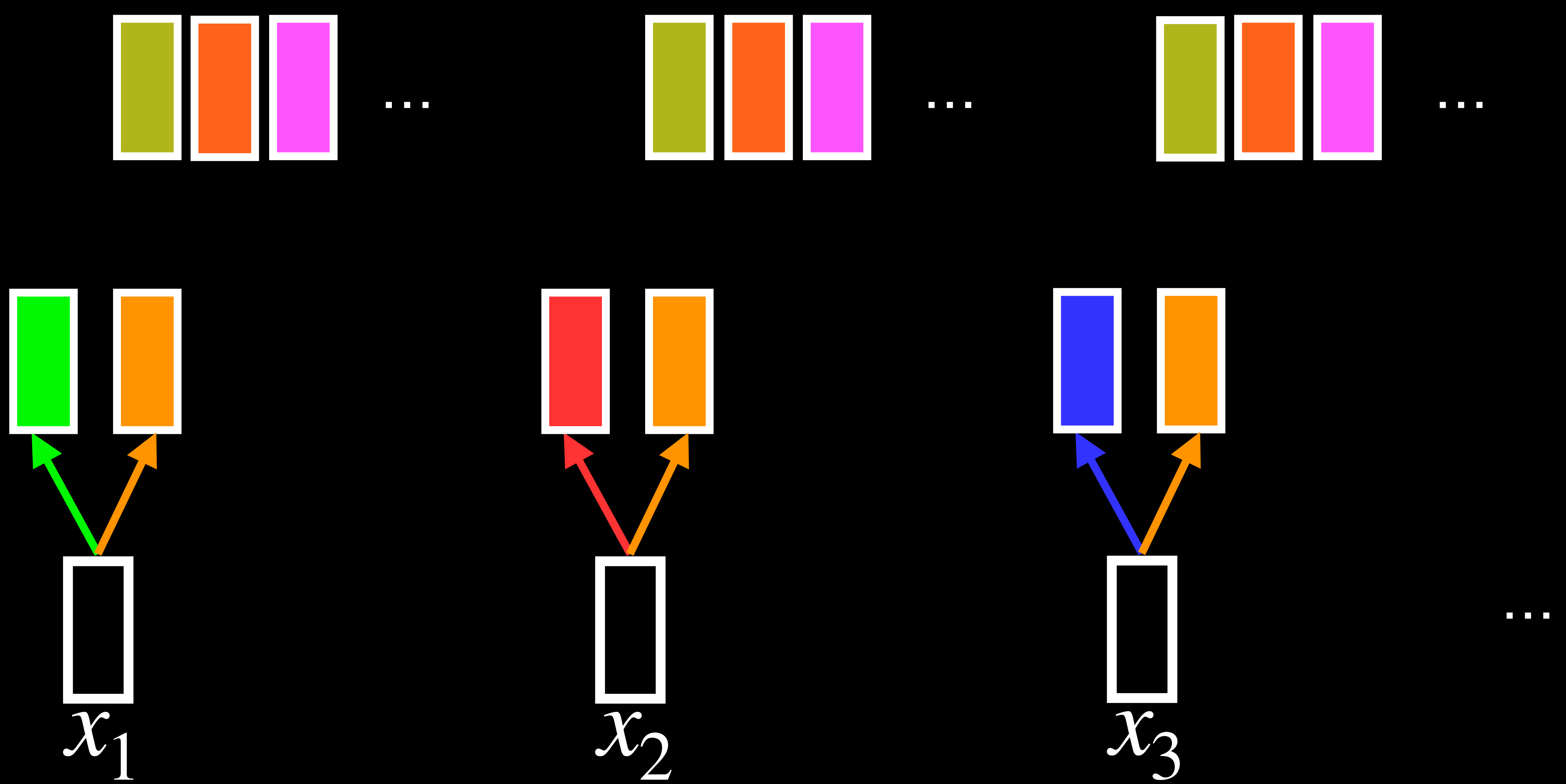
Transformers (3)



Vaswani et al (2017) Attention is all you need, *NeurIPS*

Transformers (4)

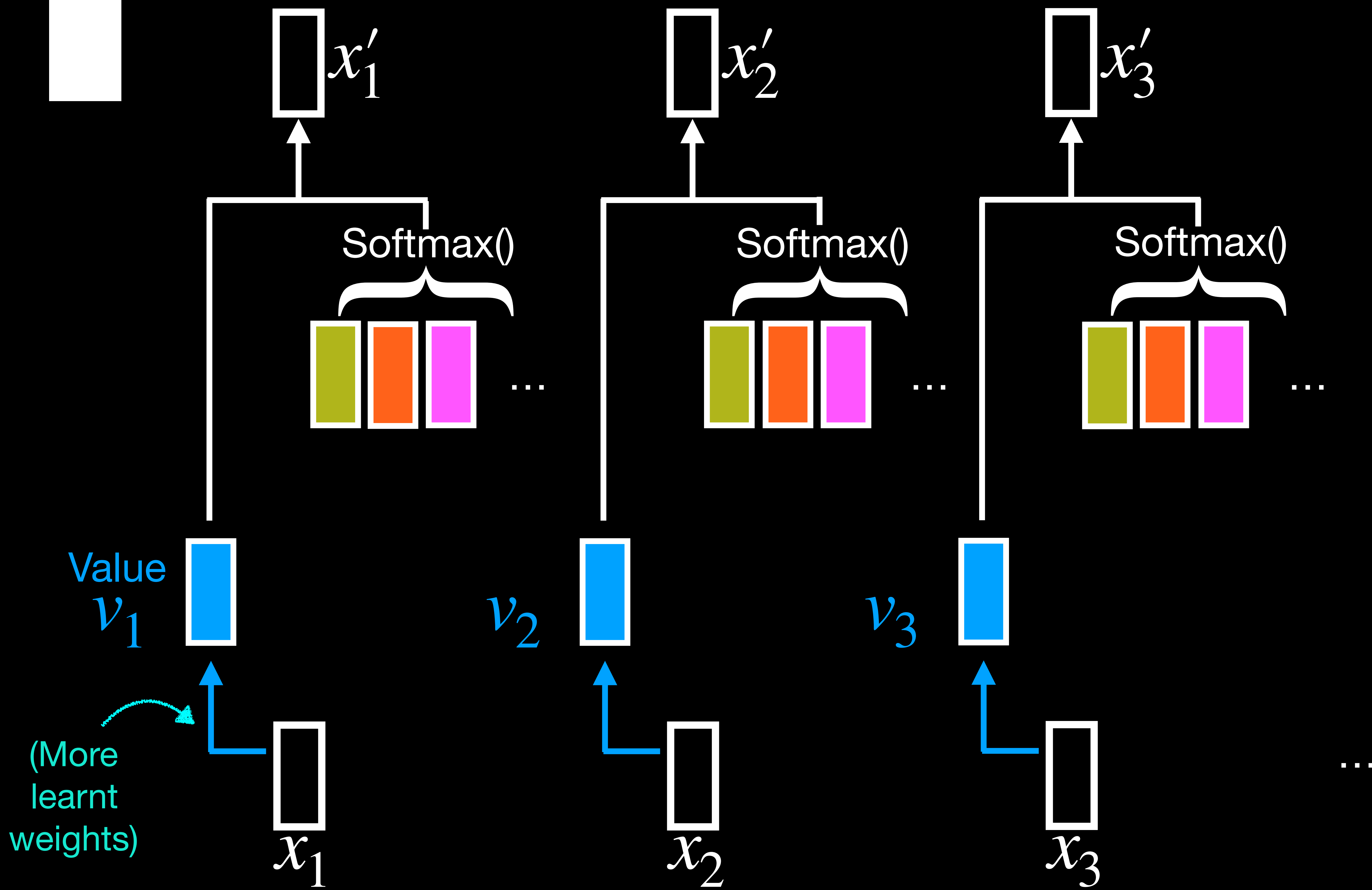
“Attention weights of all tokens on all others”



Input token (Feature vector at time t_1)

Vaswani et al (2017) Attention is all you need, *NeurIPS*

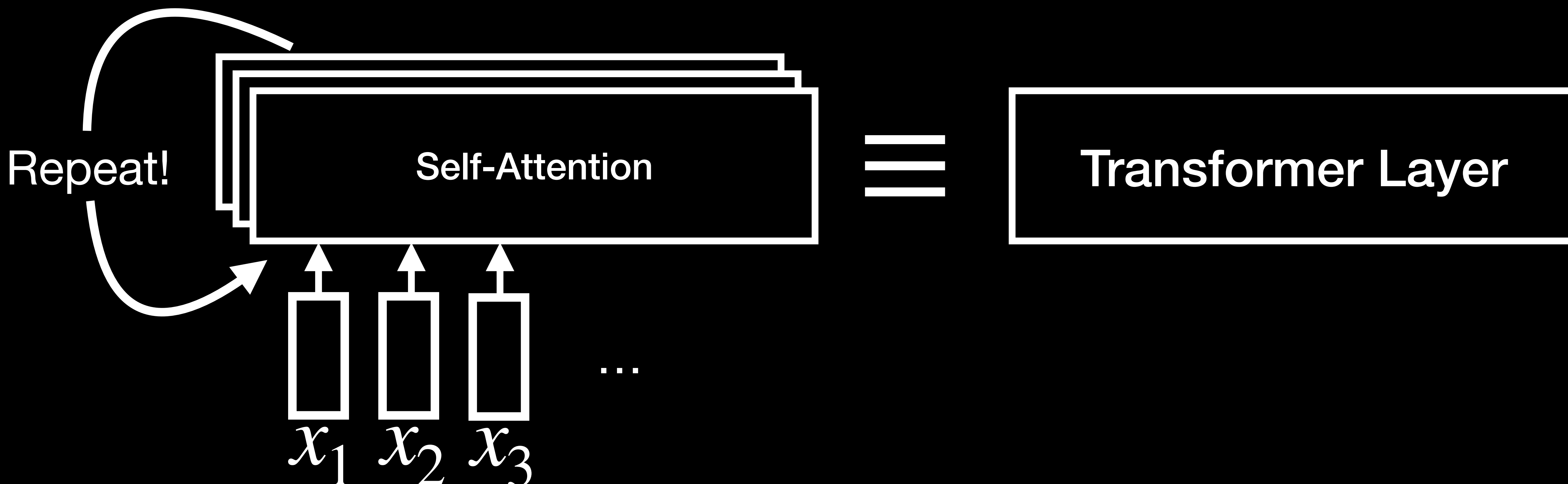
Transformers (5)



Input token (Feature vector at time t_1)

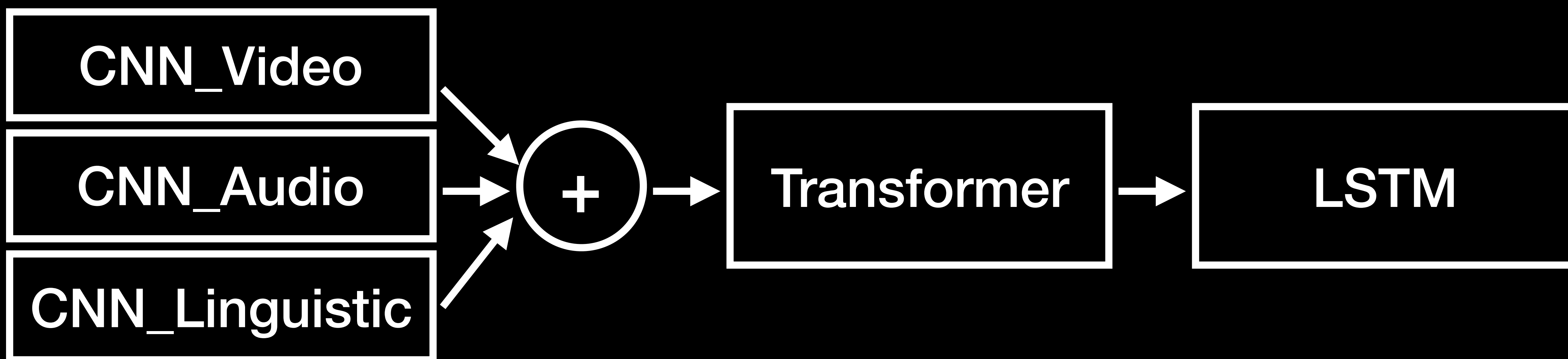
Vaswani et al (2017) Attention is all you need, *NeurIPS*

Transformers (6)



Vaswani et al (2017) Attention is all you need, *NeurIPS*

Simple Fusion Transformer + Results



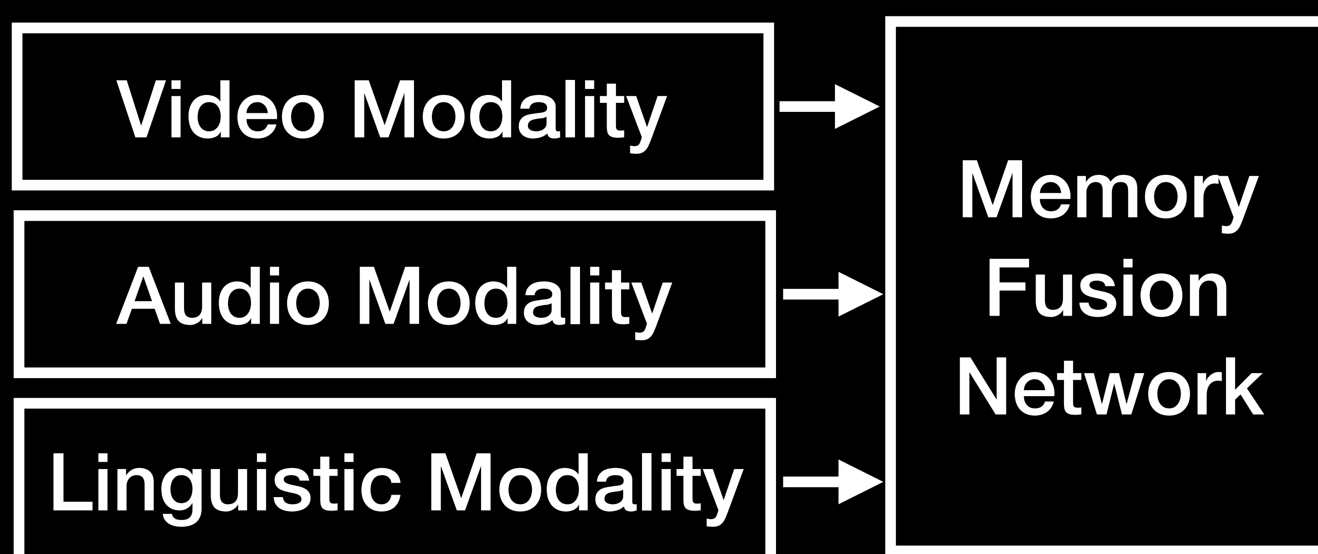
Concordance Correlation on Test Set

	Best Unimodal	Best Bimodal	Trimodal
SFT	.34	.35	.14
Human	-	-	.50

Memory Fusion Transformer

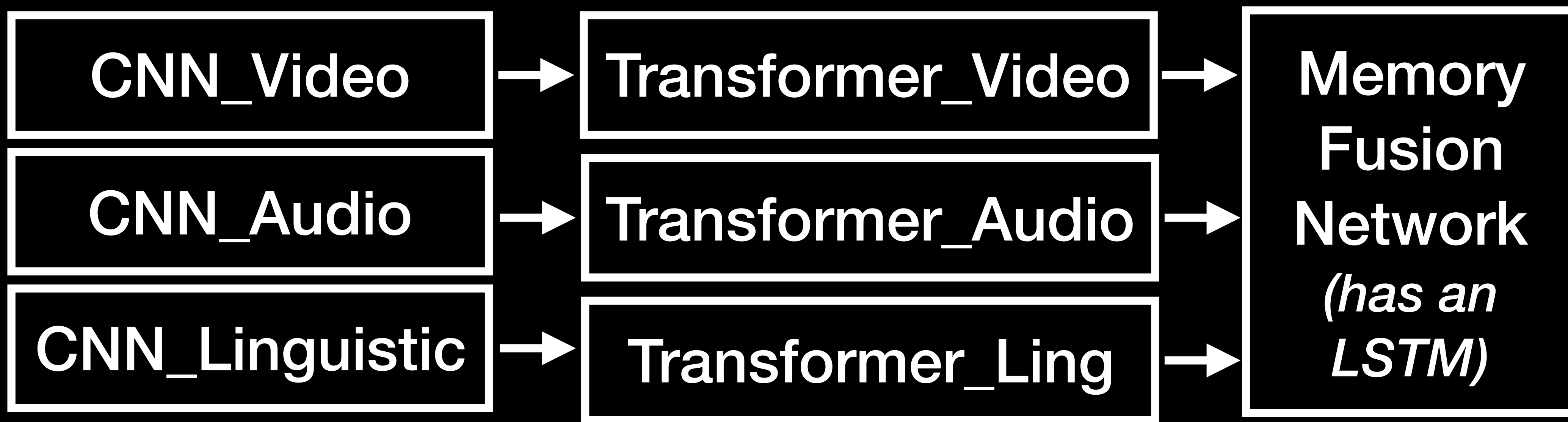
Our Simple Fusion Transformer [“**Self-Attention**”]
- Does well on Linguistic input
- And on Linguistic + Visual
- But performs poorly on trimodal input

Decided to also implement Memory Fusion Network (Zadeh et al, 2018), which learns “**cross-modality attention**”



Zadeh et al (2018), Memory Fusion Network for Multi-view Sequential Learning. *AAAI*

Memory Fusion Transformer

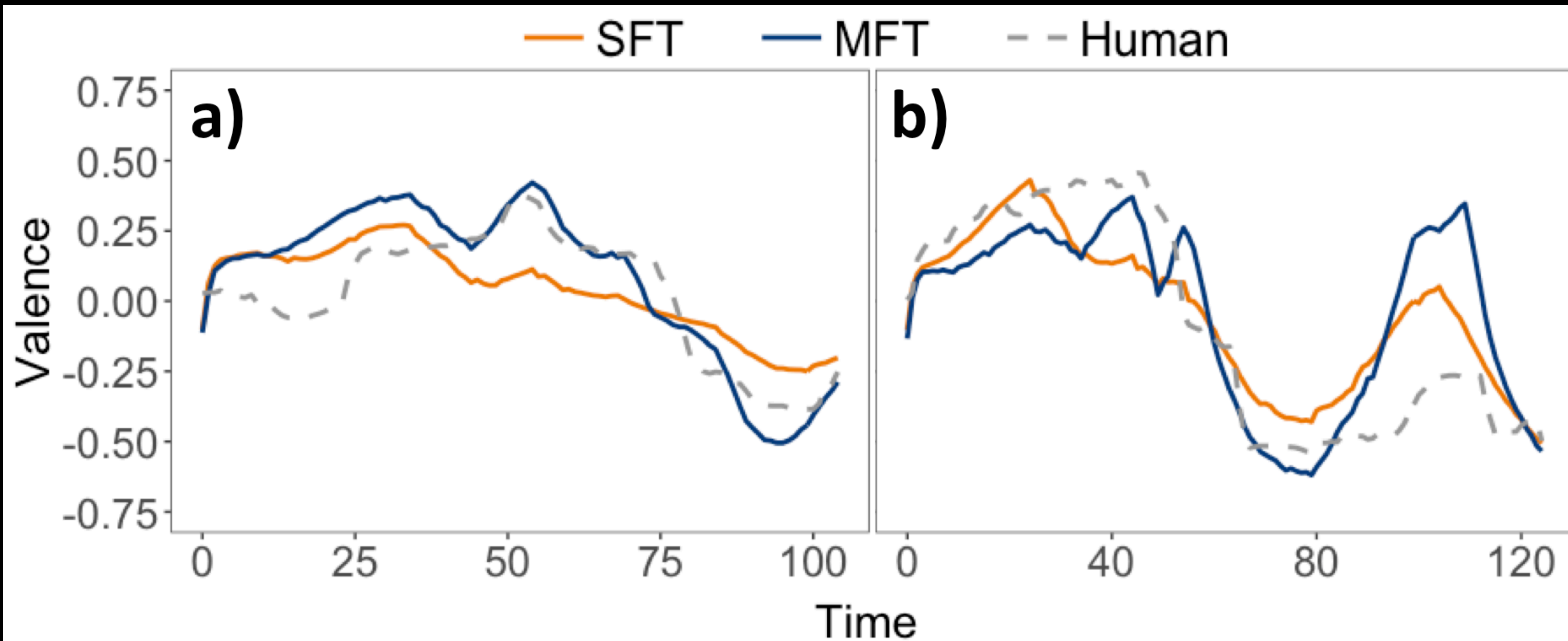


Concordance Correlation on Test Set

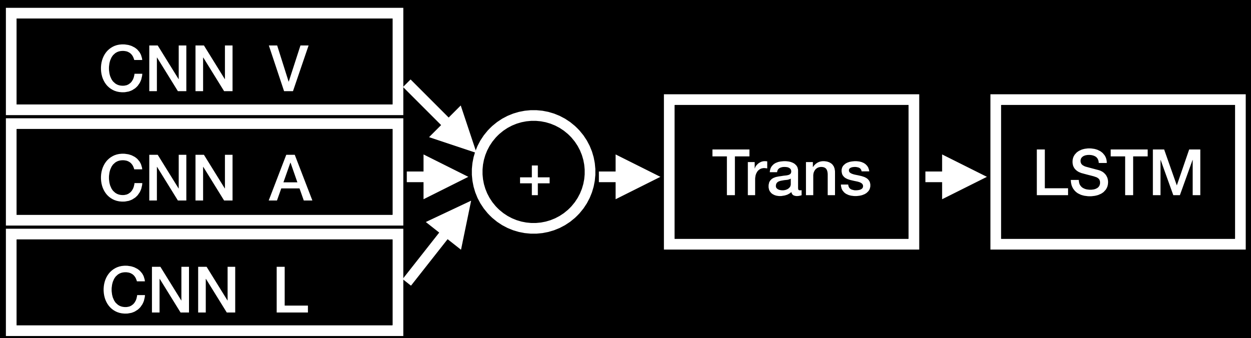
	Best Unimodal	Best Bimodal	Trimodal
SFT	.34	.35	.14
MFT	-	.36	.44
Human	-	-	.50

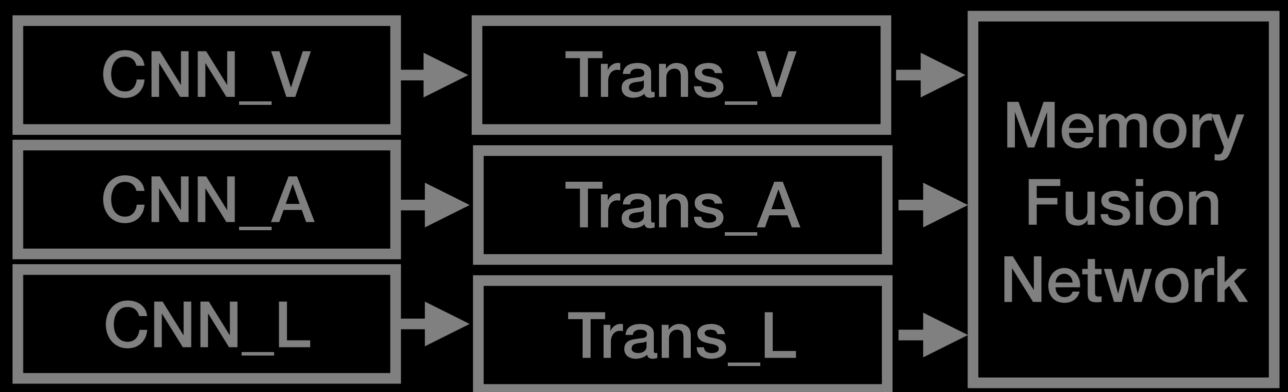
(n.s)

Results


SFT
MFT
Human
a)
b)
Valence
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
0
25
50
75
100
0
40
80
120
Time

Lesion Experiments (1)



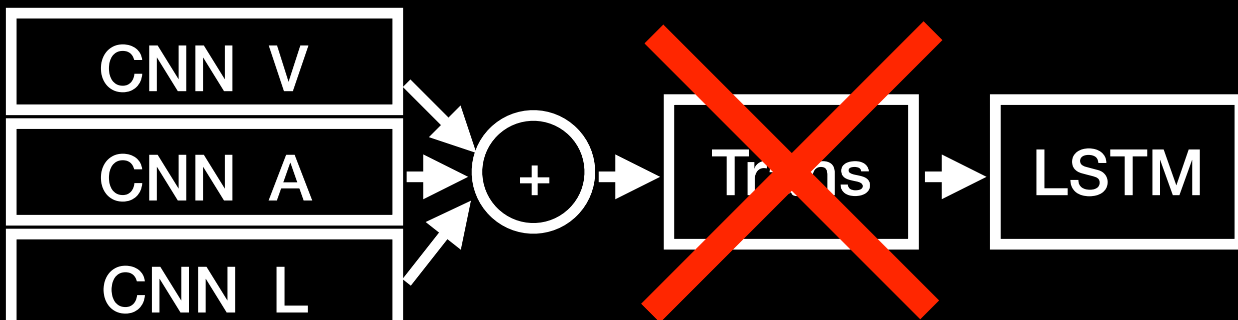


Simple Fusion Transformer

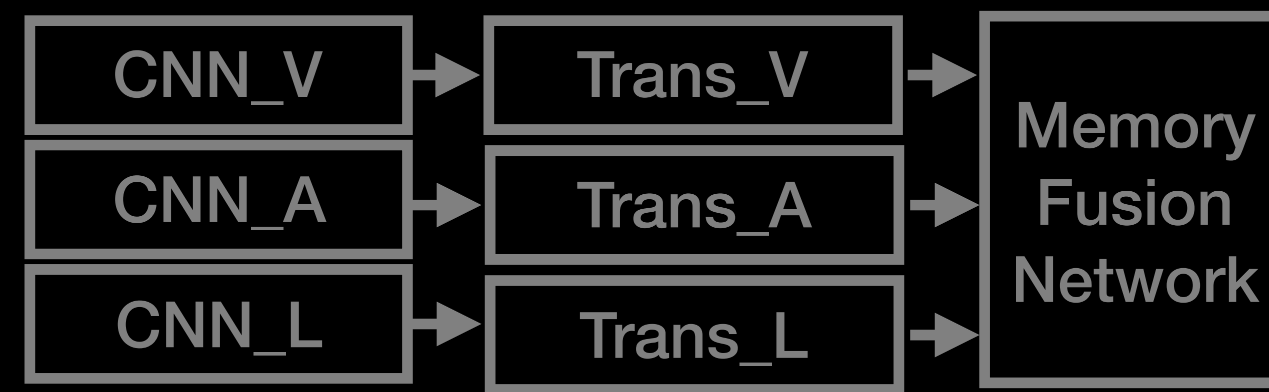
Concordance Correlation on Test Set

	Best Unimodal	Best Bimodal	Trimodal
SFT	.34	.35	.14
LSTM-only			
Trans-only			
MFT	-	.36	.44
MFN-only			
Human	-	-	.50

Lesion Experiments (1)



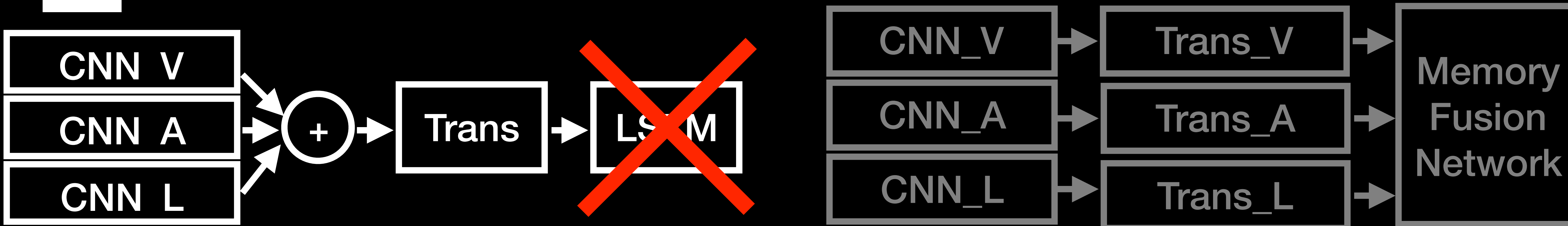
Simple Fusion Transformer



Concordance Correlation on Test Set

	Best Unimodal	Best Bimodal	Trimodal
SFT	.34	.35	.14
LSTM-only	.21	.17	-.02
Trans-only			
MFT	-	.36	.44
MFN-only			
Human	-	-	.50

Lesion Experiments (2)

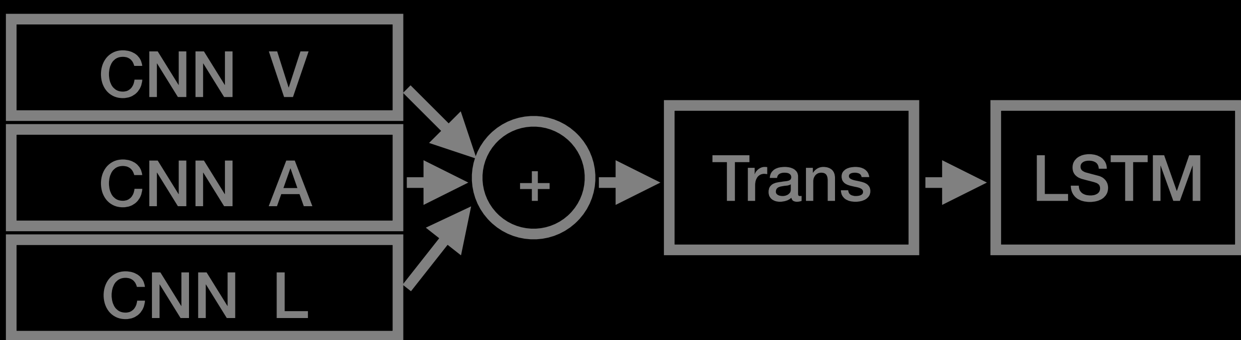


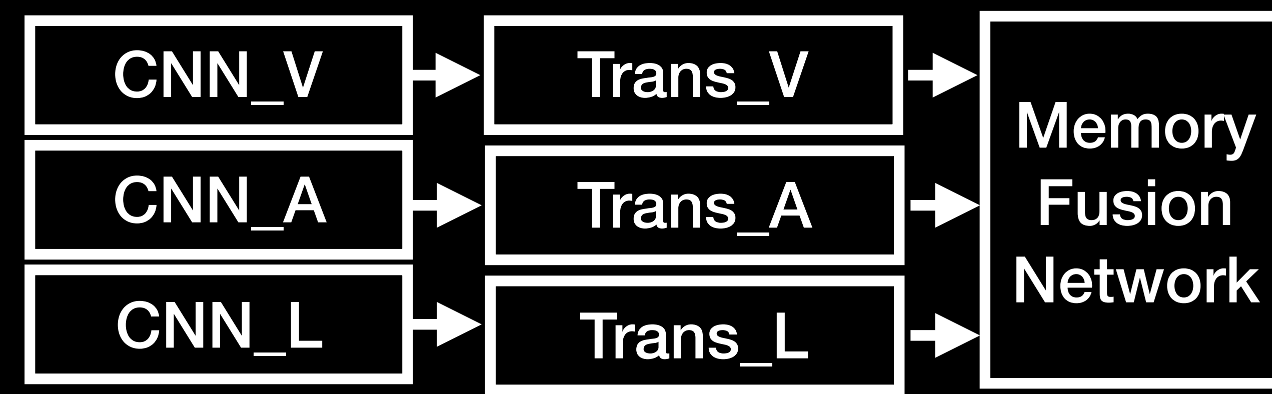
Simple Fusion Transformer

Concordance Correlation on Test Set

	Best Unimodal	Best Bimodal	Trimodal
SFT	.34	.35	.14
LSTM-only	.21	.17	-.02
Trans-only	.05	.05	.00
MFT	-	.36	.44
MFN-only			
Human	-	-	.50

Lesion Experiments (3)



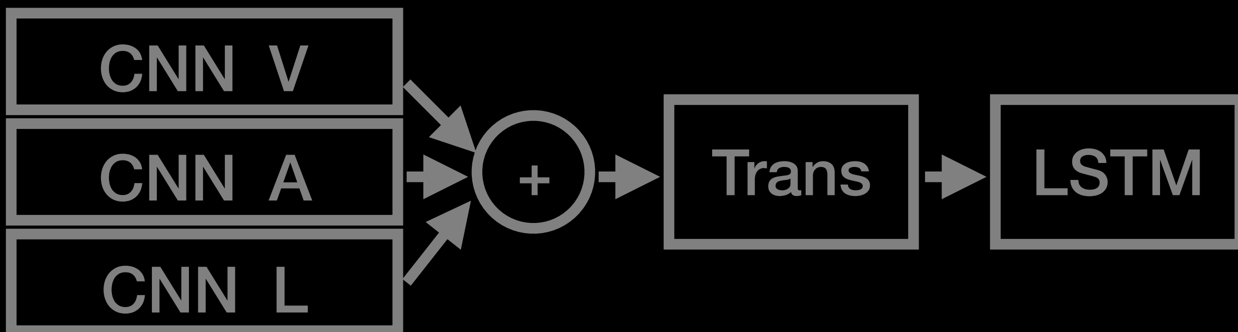


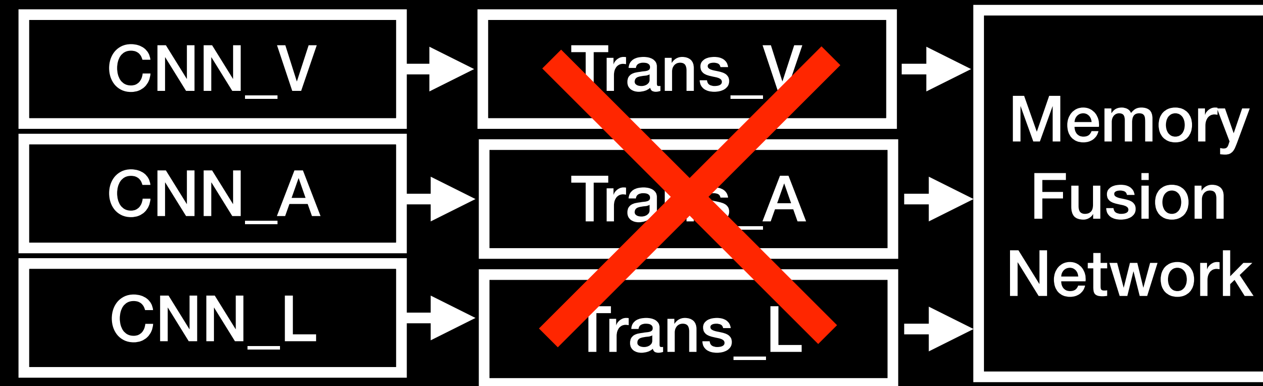
Memory Fusion Transformer

Concordance Correlation on Test Set

	Best Unimodal	Best Bimodal	Trimodal
SFT	.34	.35	.14
LSTM-only	.21	.17	-.02
Trans-only	.05	.05	.00
MFT	-	.36	.44
MFN-only			
Human	-	-	.50

Lesion Experiments (3)





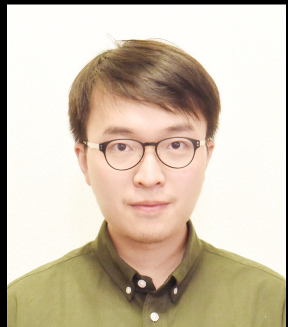
Memory Fusion Transformer

Concordance Correlation on Test Set

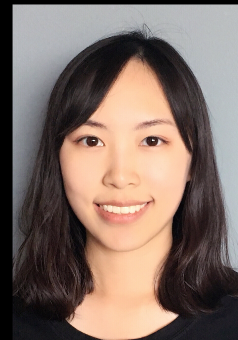
	Best Unimodal	Best Bimodal	Trimodal
SFT	.34	.35	.14
LSTM-only	.21	.17	-.02
Trans-only	.05	.05	.00
MFT	-	.36	.44
MFN-only	-	.33	.28
Human	-	-	.50

Summary, Limitations and Future Directions

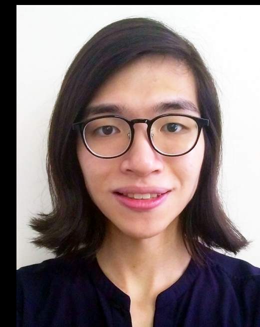
- Showed that neural network attention mechanisms (self-attention, cross-modality attention) can improve multimodal emotion recognition.
  - Lesioned experiments suggest that different types of attention contribute to better performance.
- Current work: probing attention weights
- Could serve as a way to build explainable affective computers
- More work on SEND: More diverse demographics, cross-cultural…



Zhengxuan Wu



Xiyu Zhang



Zhi-Xuan Tan

Jamil Zaki

Other collaborators on the SEND

Marianne Reddan
Xi Jia Zhou
Isabella Kahhale
Alison Mattek

Anat Perry (@ HUJI)

Thanks!

dco@comp.nus.edu.sg
web.stanford.edu/~dco

Paper: https://arxiv.org/abs/1907.04197

Code: https://github.com/frankaging/ACII2019-transformer