Strong and Simple Baselines for Multimodal Utterance Embeddings

Paul Pu Liang*, **Yao Chong Lim***, Yao-Hung Hubert Tsai,
Ruslan Salakhutdinov and Louis-Philippe Morency

Human Language is often multimodal

Language
- Word choice
- Syntax
- Pragmatics

Acoustic
- Tone
- Prosody
- Phrasing

Visual
- Facial expressions
- Body language
- Eye contact
- Gestures

Sentiment
- Positive/Negative
- Intensity

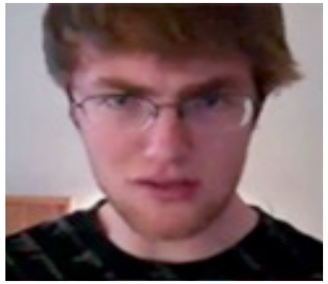
Emotion
- Anger
- Happiness
- Sadness
- Confusion
- Fear
- Surprise

Meaning
- Sarcasm
- Humor

Human Language is often multimodal

Sentiment Intensity

"This movie is great" +

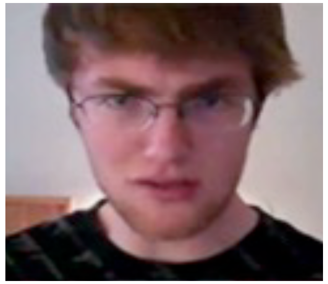


Neutral expression

Human Language is often multimodal

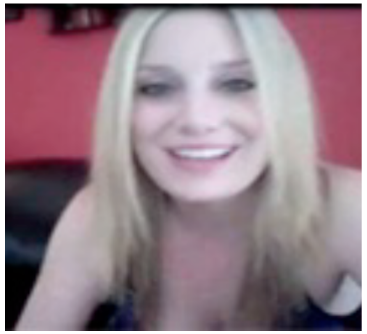
Sentiment Intensity

"This movie is great" +



Neutral expression

"This movie is great" +



Smile

Challenges in Multimodal ML

Challenges in Multimodal ML

1. Intramodal interactions

Smile + Head nod vs. Smile + Head shake

Challenges in Multimodal ML

1. Intramodal interactions

Smile + Head nod vs. Smile + Head shake

2. Crossmodal interactions

Bimodal “This movie is great” + Smile

Challenges in Multimodal ML

1. Intramodal interactions

Smile + Head nod vs. Smile + Head shake

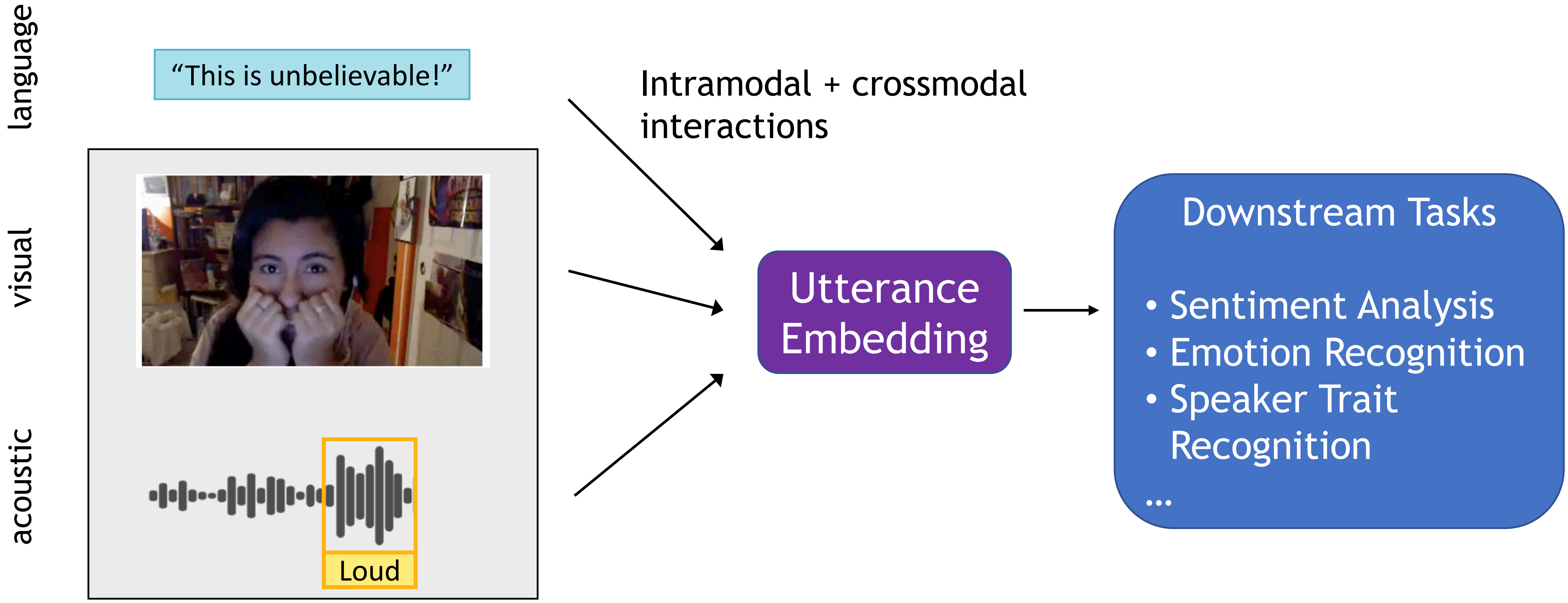
2. Crossmodal interactions

Bimodal “This movie is great” + Smile

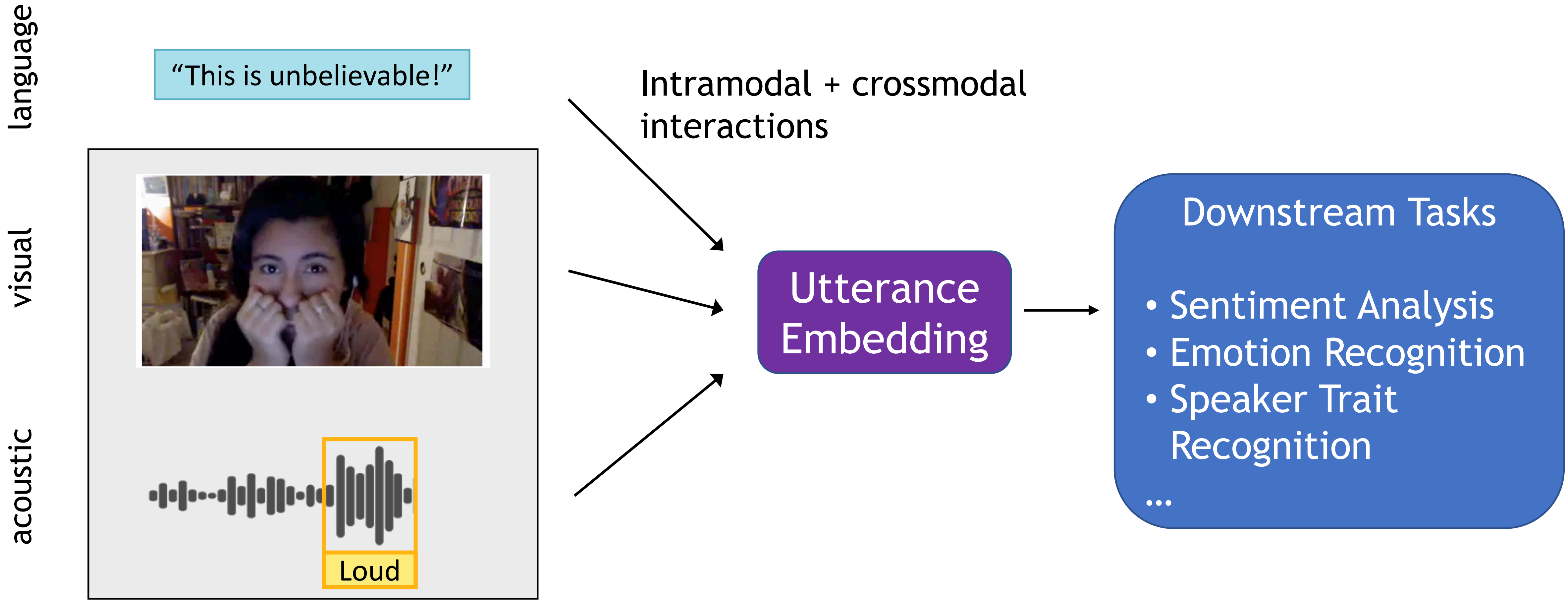
Trimodal “This movie is GREAT” + Smile + “great” is emphasized, drawn-out

(Sarcasm)

Multimodal Language Embedding



Multimodal Language Embedding



Why fast models?

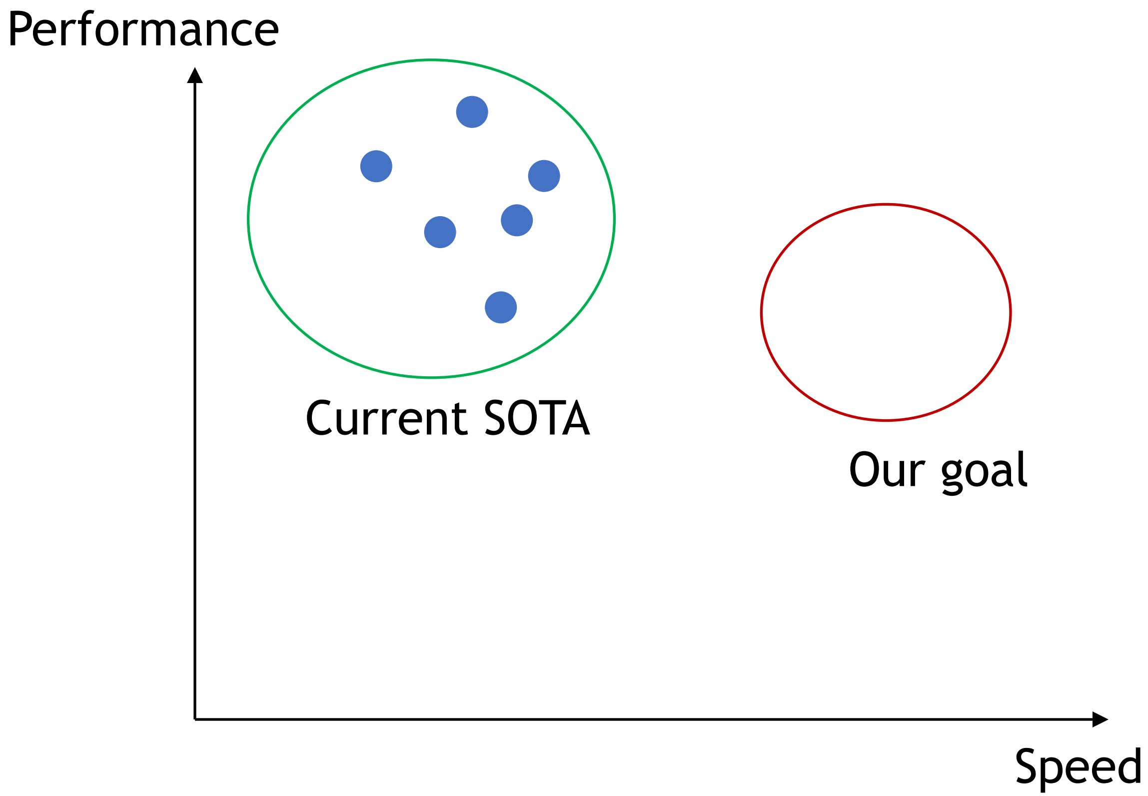
- Applications
- Robots, virtual agents, intelligent personal assistants
- Processing large amounts of multimedia data

Research Question

Can we make principled but simple models for multimodal utterance embeddings that perform competitively?

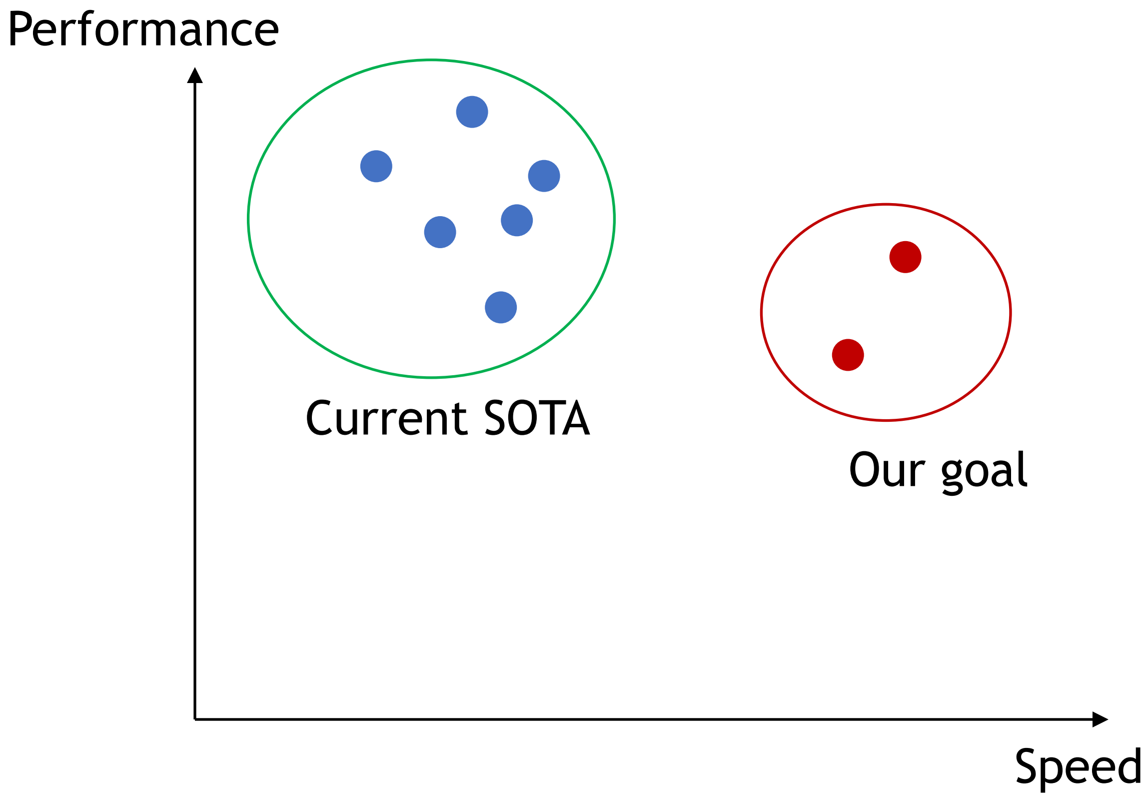
Research Question

Can we make principled but simple models for multimodal utterance embeddings that perform competitively?



Research Question

Can we make principled but simple models for multimodal utterance embeddings that perform competitively?

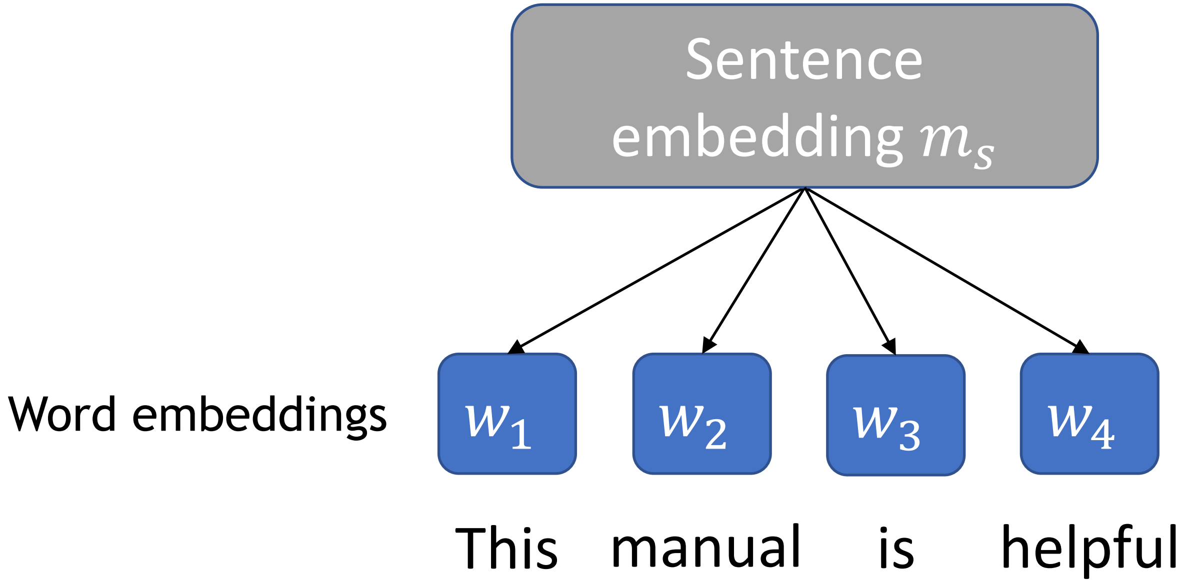


Our models:

- Fewer parameters
- Has a closed-form solution
- Linear functions
- **Competitive with SOTA!**

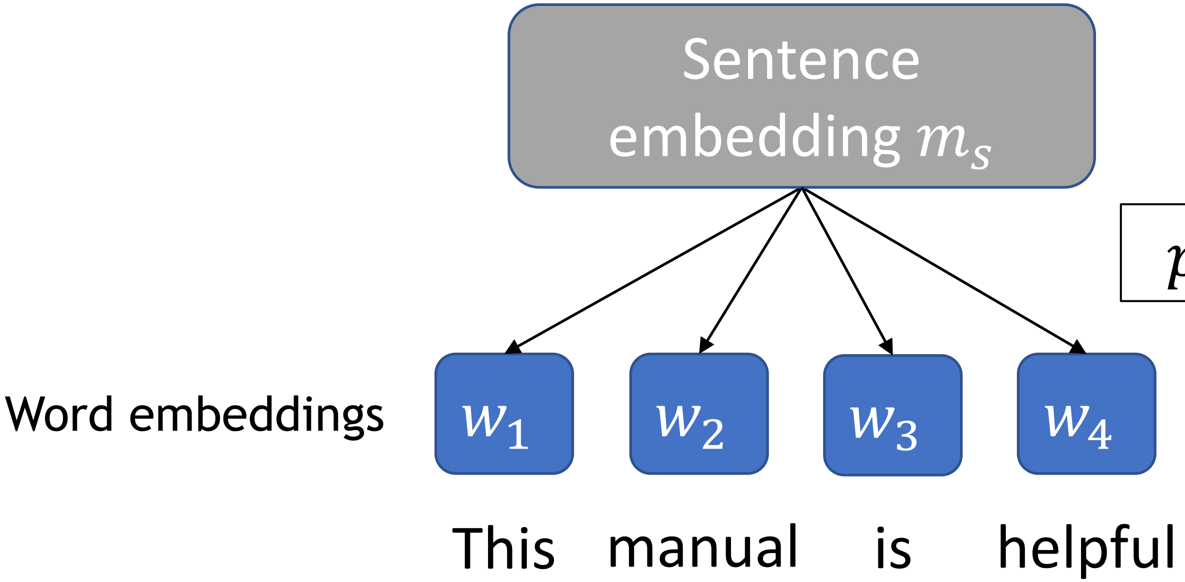
A language-only solution

Arora et al. (2016, 2017):



A language-only solution

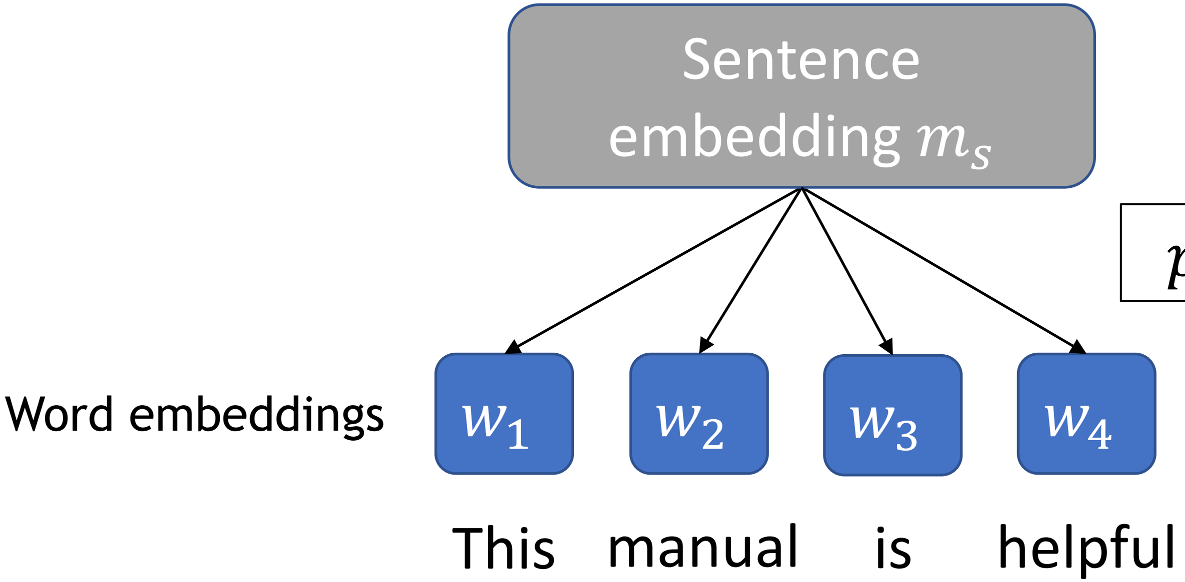
Arora et al. (2016, 2017):



$$\boxed{p(w_i | m_s) \propto \exp(w_i \cdot m_s)}$$

A language-only solution

Arora et al. (2016, 2017):

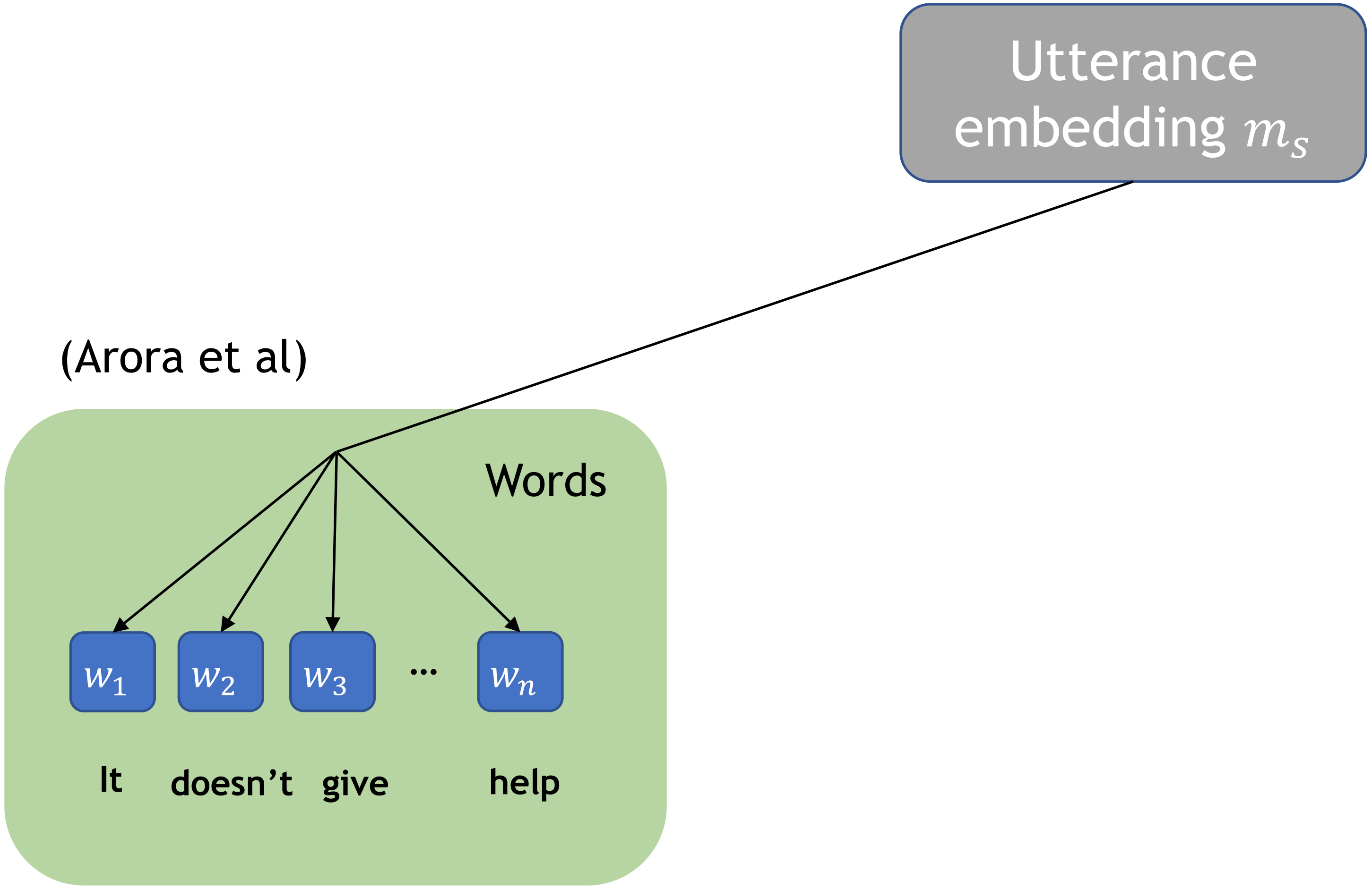


$$p(w_i|m_s) \propto \exp(w_i \cdot m_s)$$

Fast: No learnable parameters.

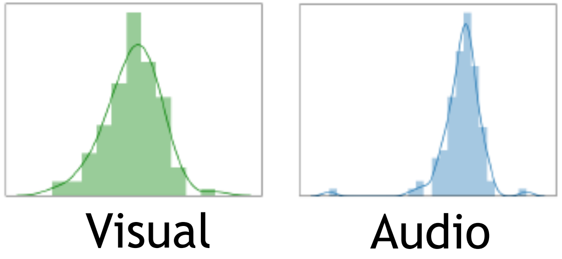
MMB1: Representing intramodal interactions

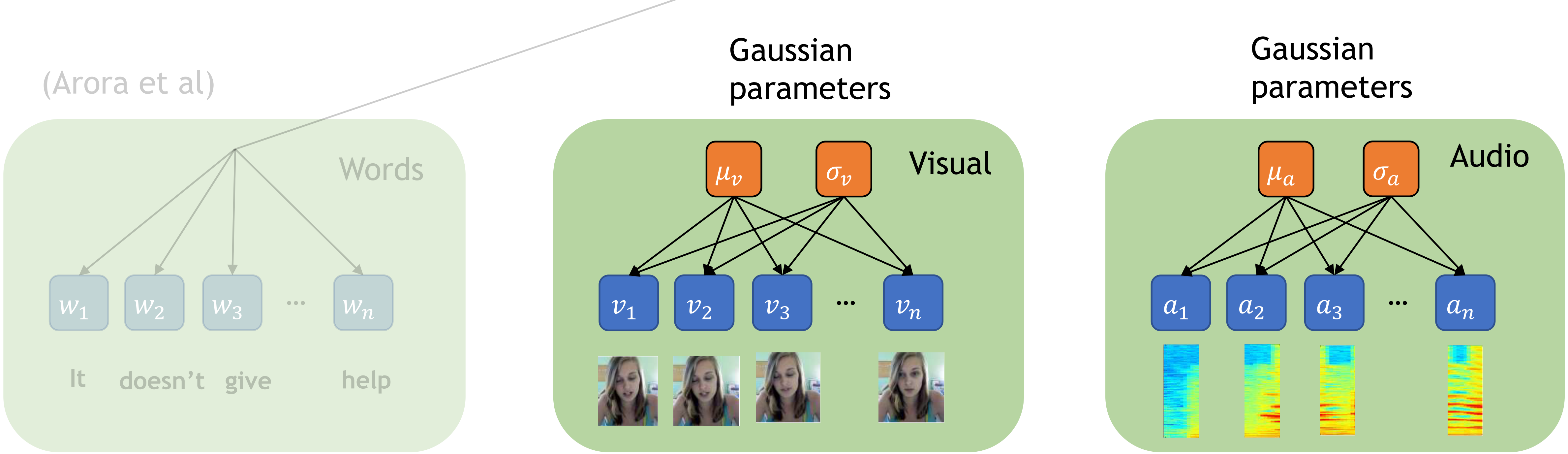
MMB1: Representing intramodal interactions



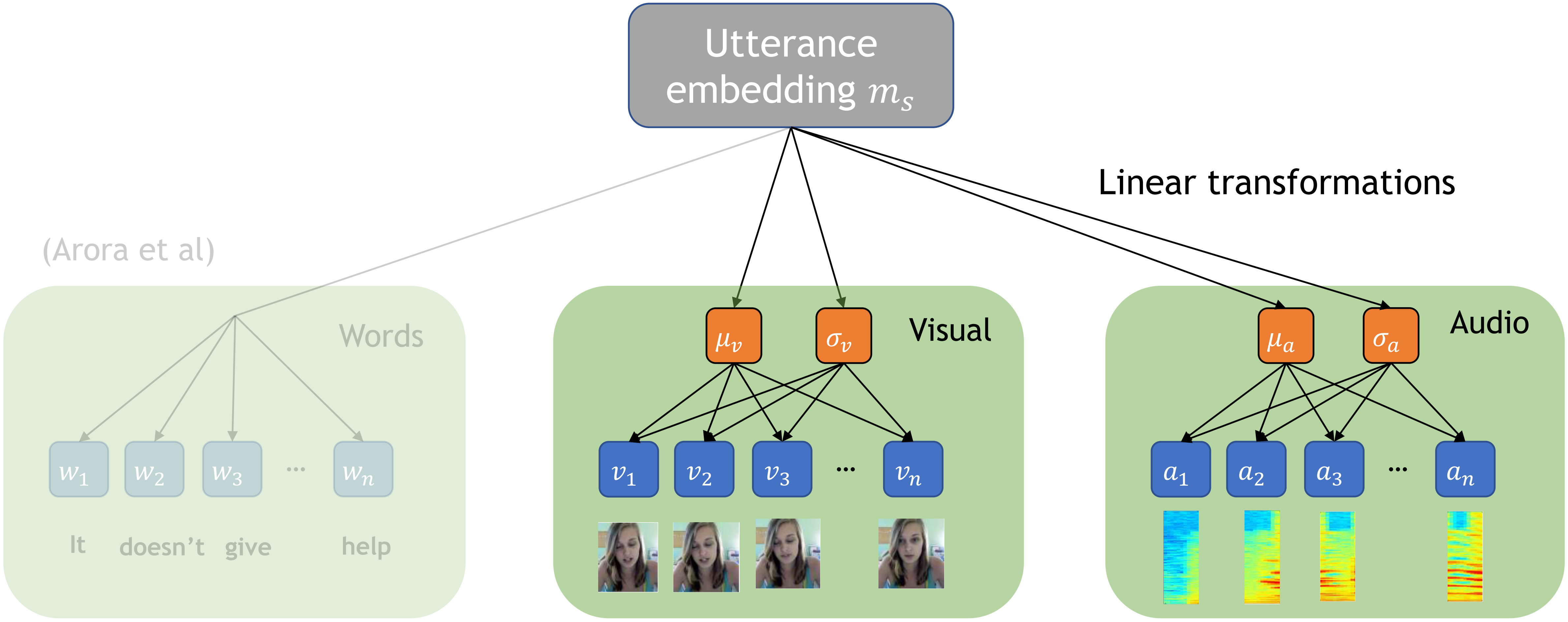
MMB1: Representing intramodal interactions

Utterance-level feature distributions:

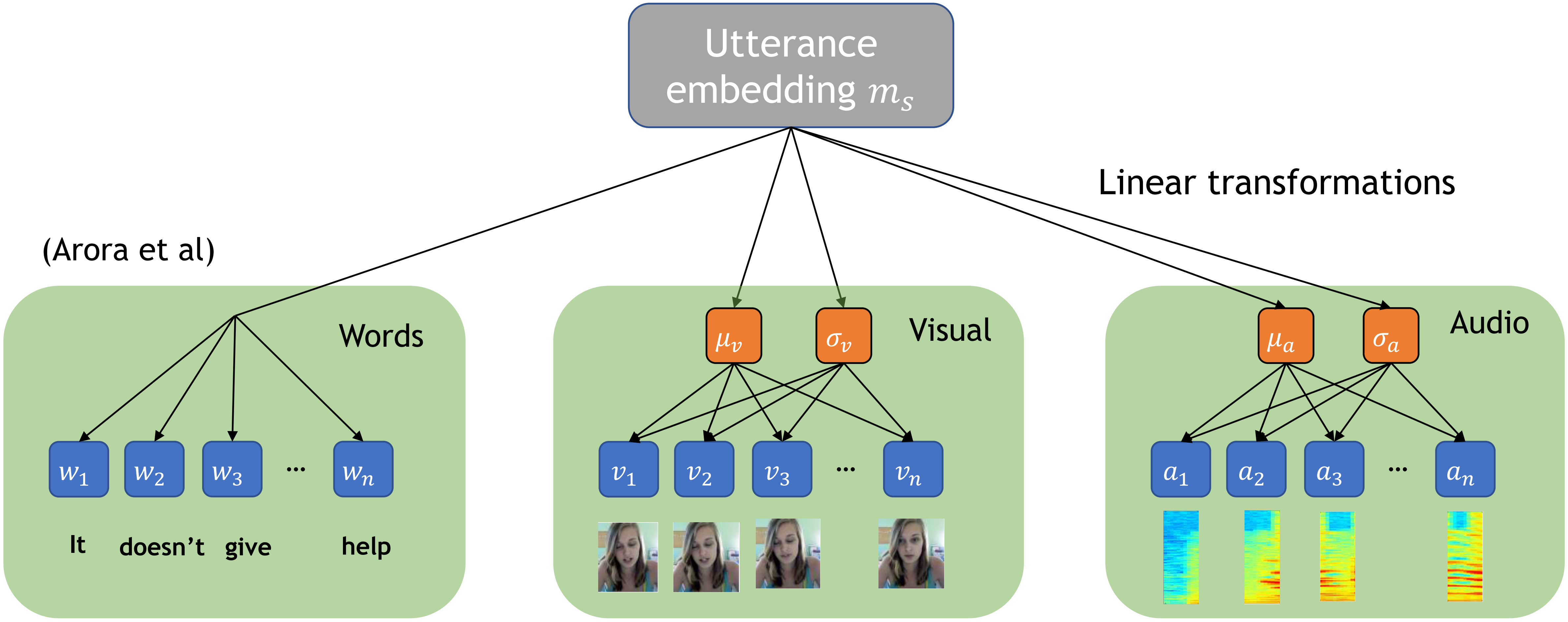




MMB1: Representing intramodal interactions

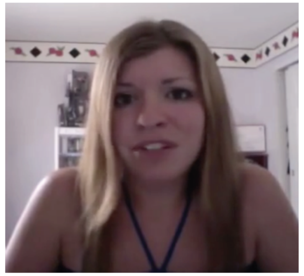
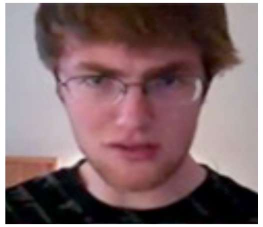


MMB1: Representing intramodal interactions

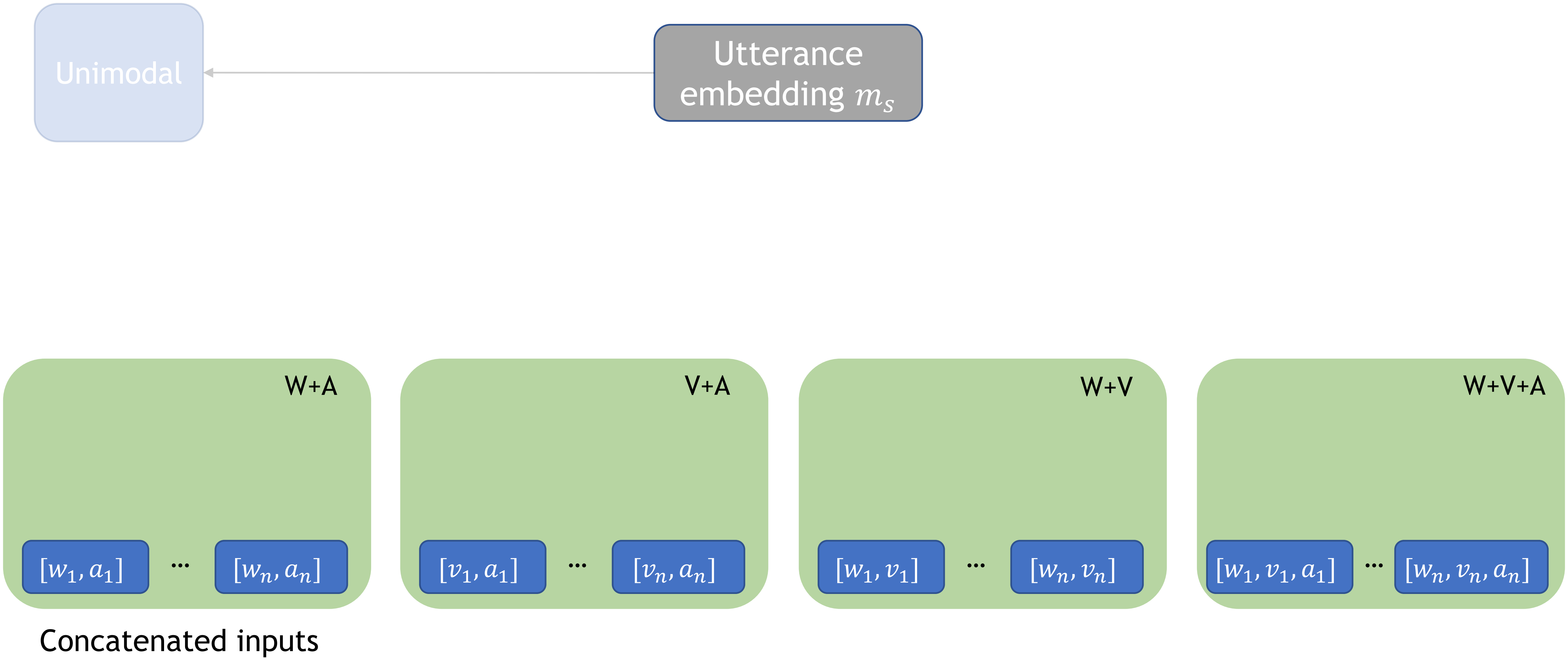


Small number of additional parameters!

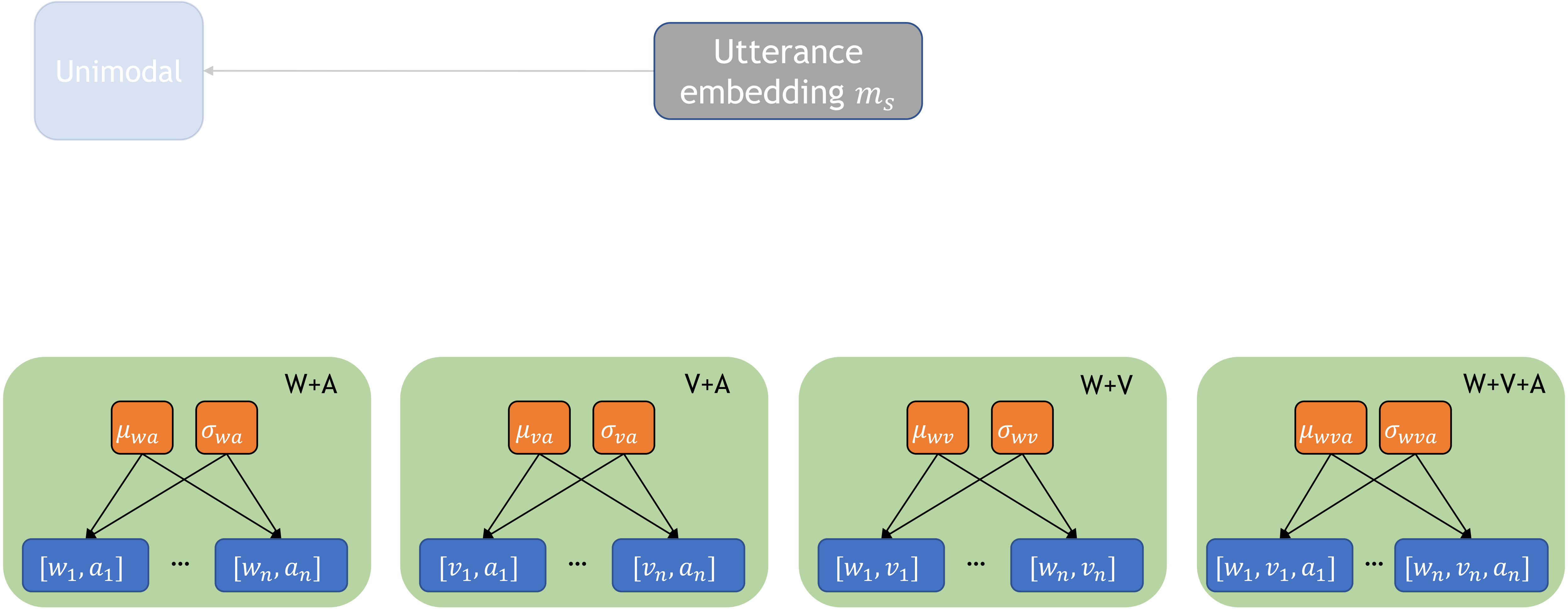
Crossmodal interactions

						Emotion
“It didn’t help”	+	 Neutral face	+	**Stable** voice		Disappointment
“It didn’t help”	+	 Sad face	+	**Shaky** voice		Sadness

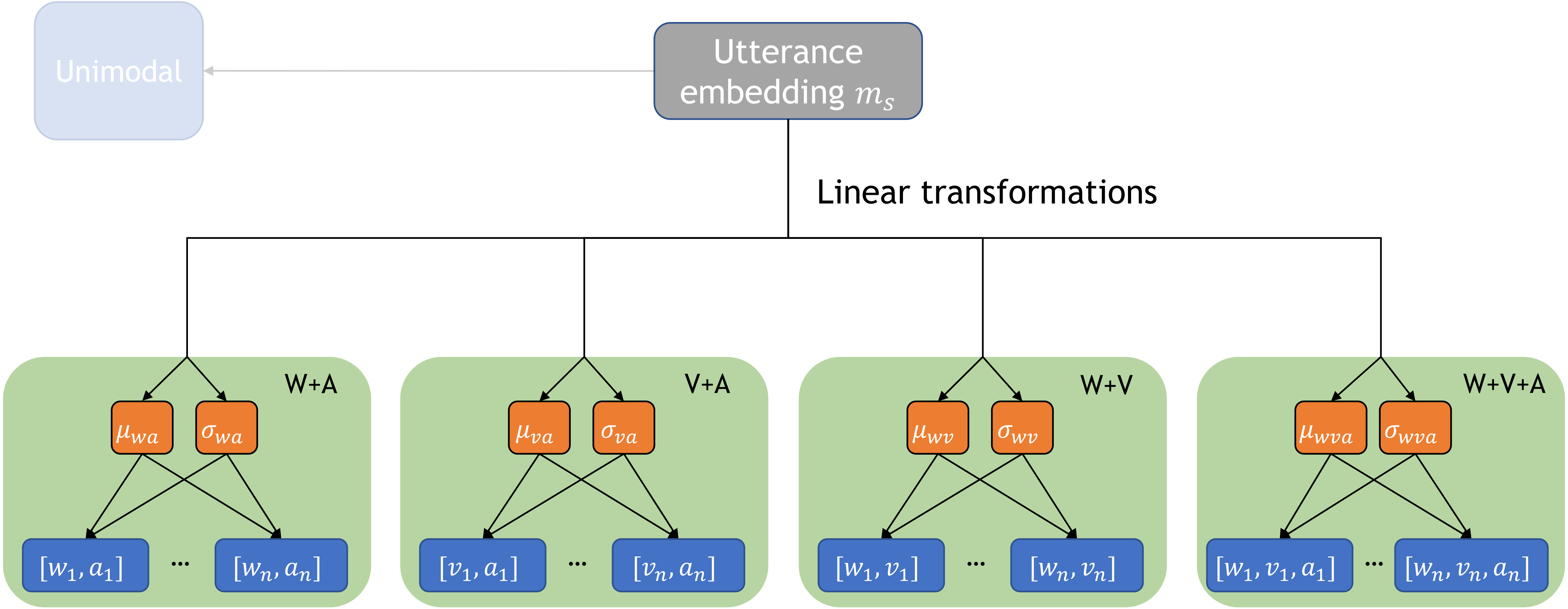
MMB2: Incorporating crossmodal interactions



MMB2: Incorporating crossmodal interactions

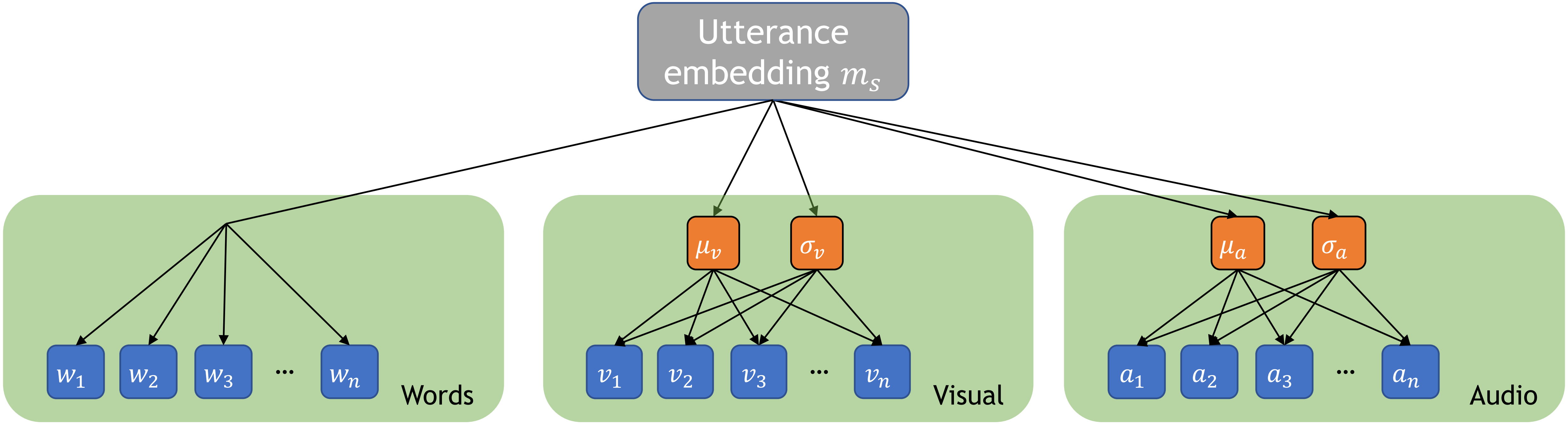


MMB2: Incorporating crossmodal interactions



How do we optimize the model?

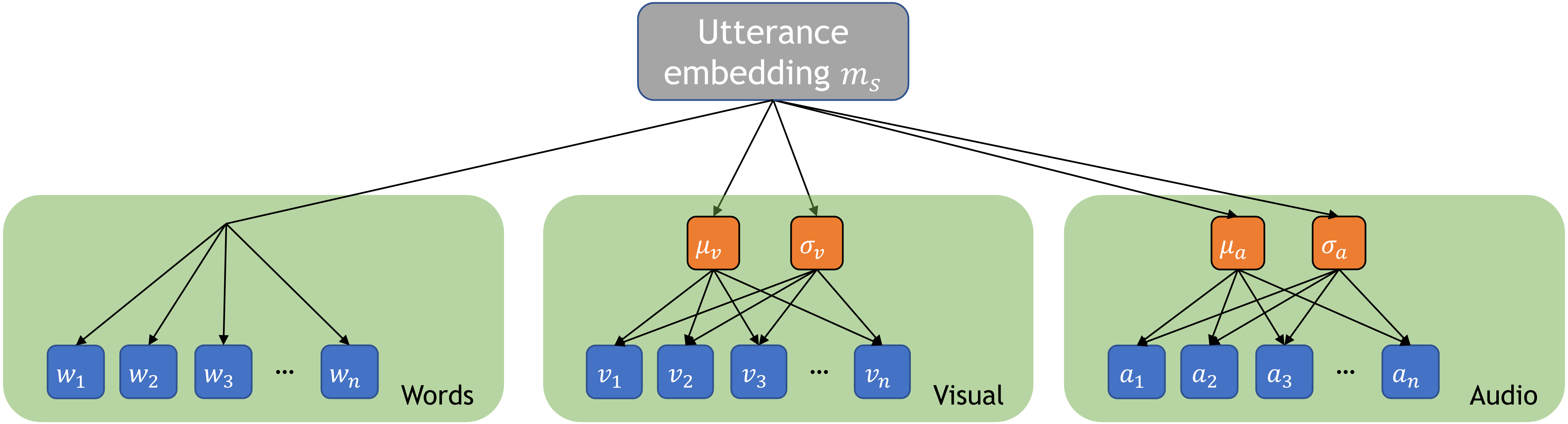
Coordinate ascent-style



How do we optimize the model?

Two steps each iteration:

Coordinate ascent-style

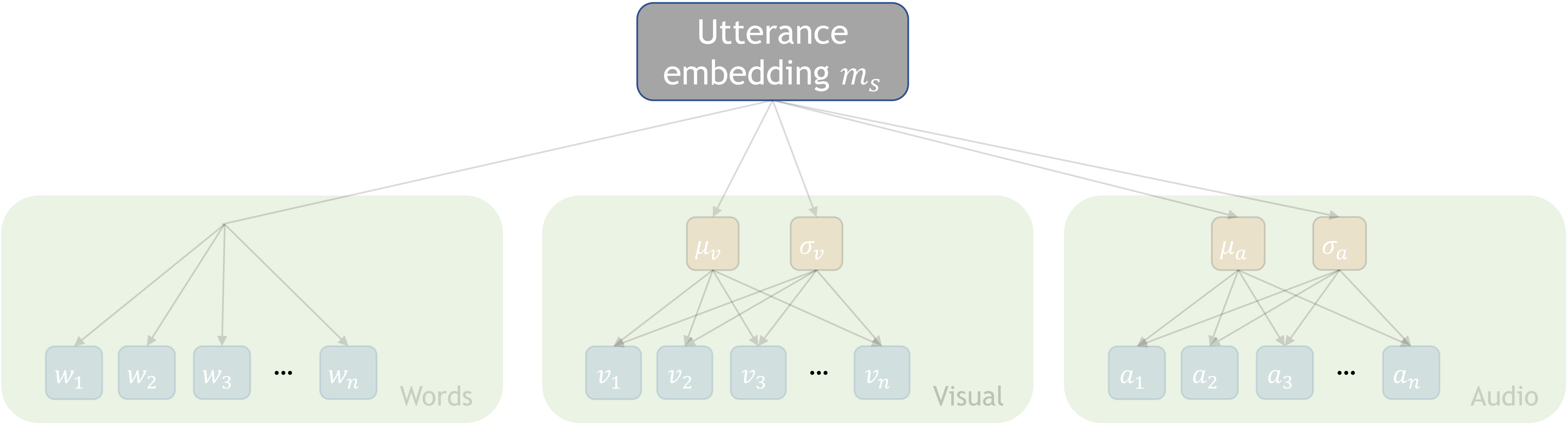


How do we optimize the model?

Coordinate ascent-style

Two steps each iteration:

1. Fix transformation parameters, solve for m_s

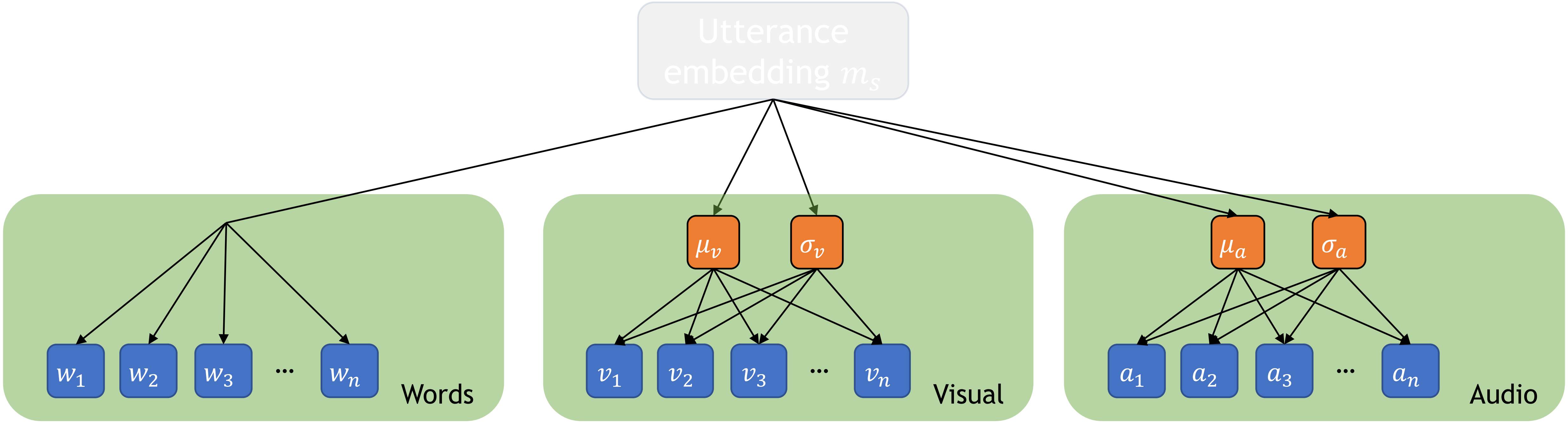


How do we optimize the model?

Two steps each iteration:

1. Fix transformation parameters, solve for m_s
2. Fix m_s, update transformation parameters by gradient descent

Coordinate ascent-style



Datasets

CMU-MOSI (Zadeh et al. 2016)

- Multimodal Sentiment Analysis dataset
- 2199 English opinion segments (monologues) from online videos

Language I thought it was fun

Visual

Acoustic (elongation) (emphasis)

Datasets

POM (Park et al., 2014)

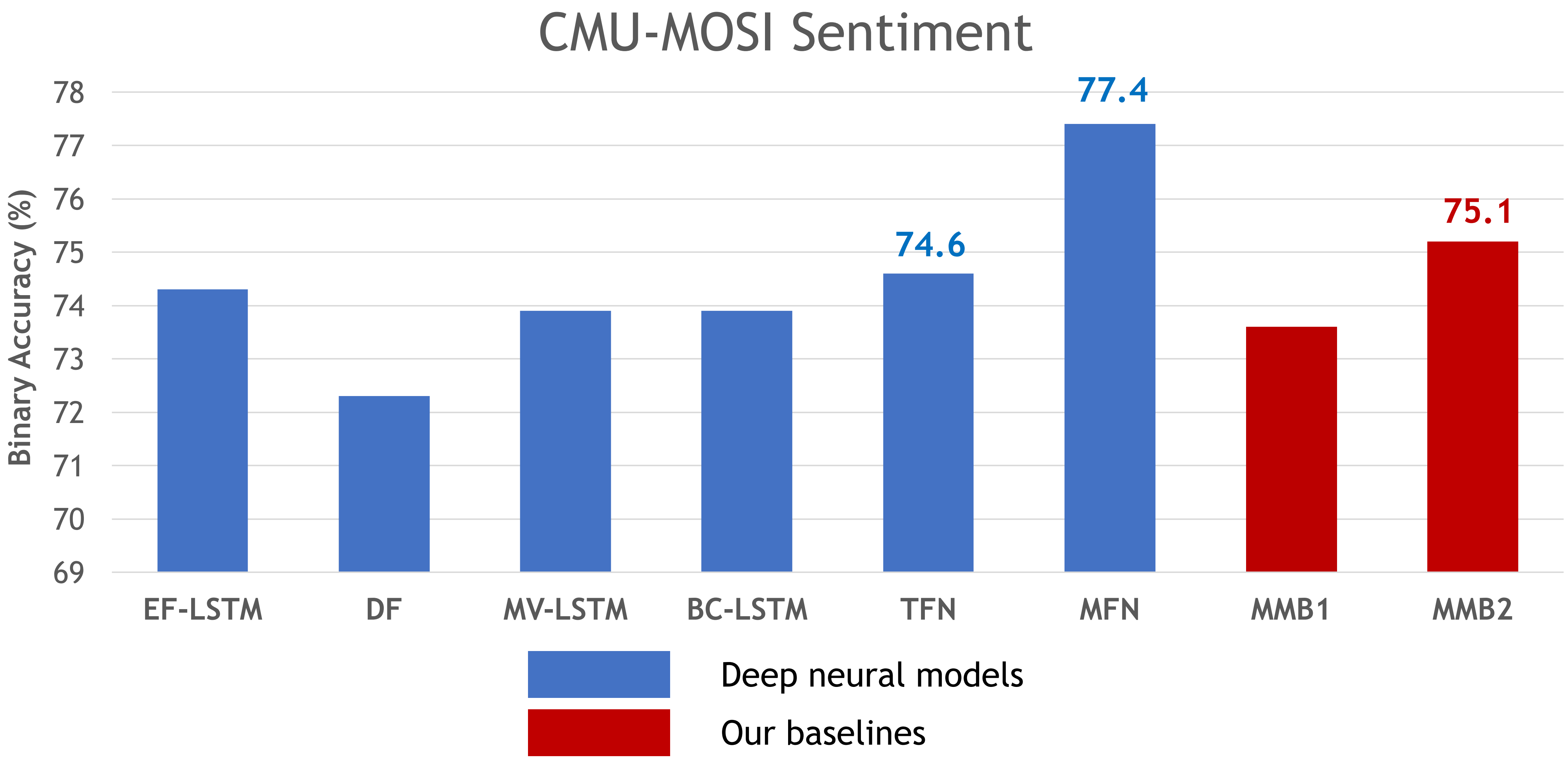
- Multimodal Speaker Traits Recognition
- 903 English videos annotated for speaker traits such as confidence, dominance, vividness, relaxed, nervousness, humor etc.

Compared Models

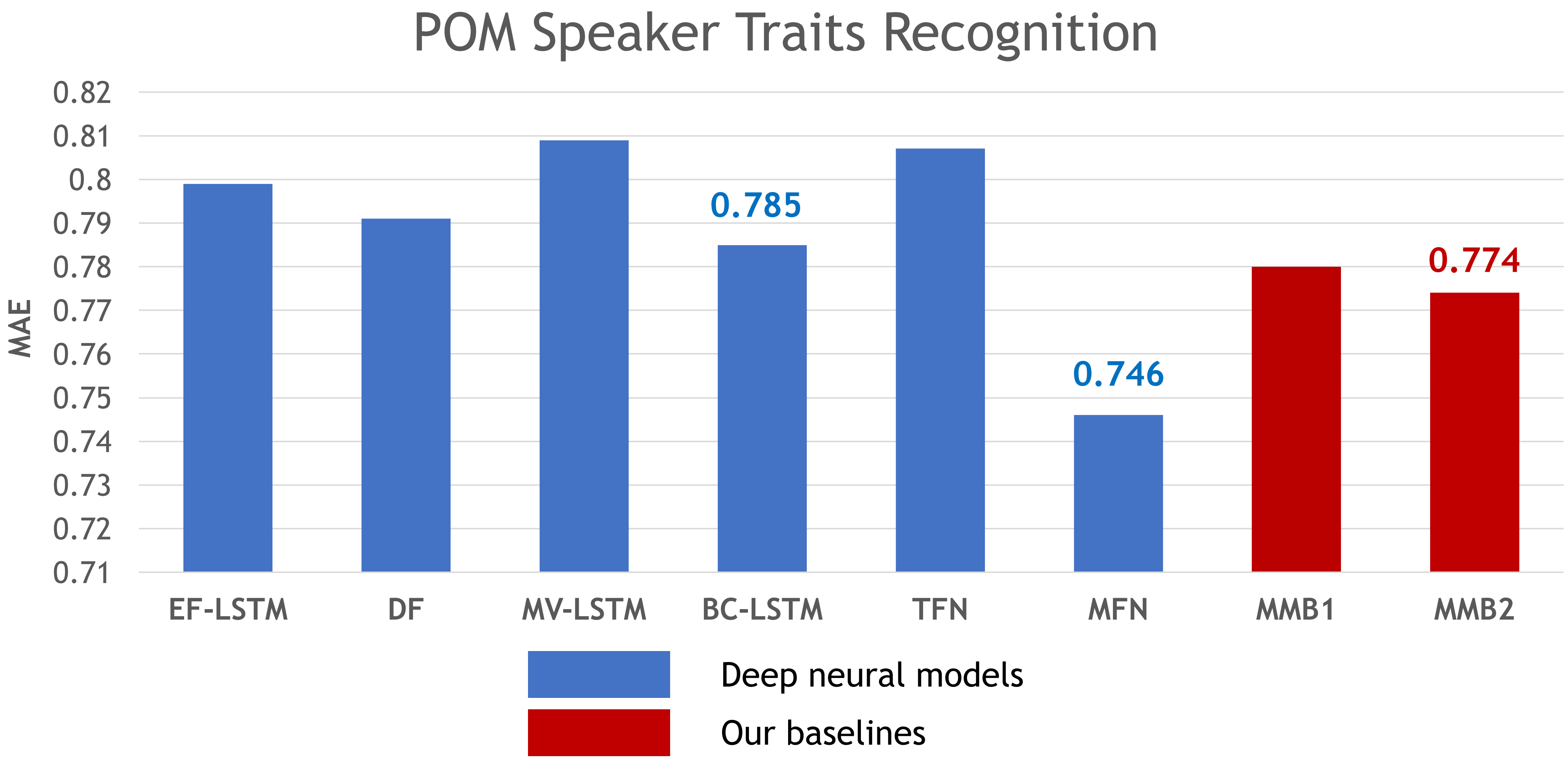
Deep neural models

- Early Fusion: EF-LSTM
- DF (Nojavanasghari et al., 2016)
- Multi-view Learning: MV-LSTM (Rajagopalan et al., 2016)
- Contextual LSTM: BC-LSTM (Poria et al., 2017)
- Tensor Fusion: TFN (Zadeh et al., 2017)
- Memory Fusion: MFN (Zadeh et al., 2018)

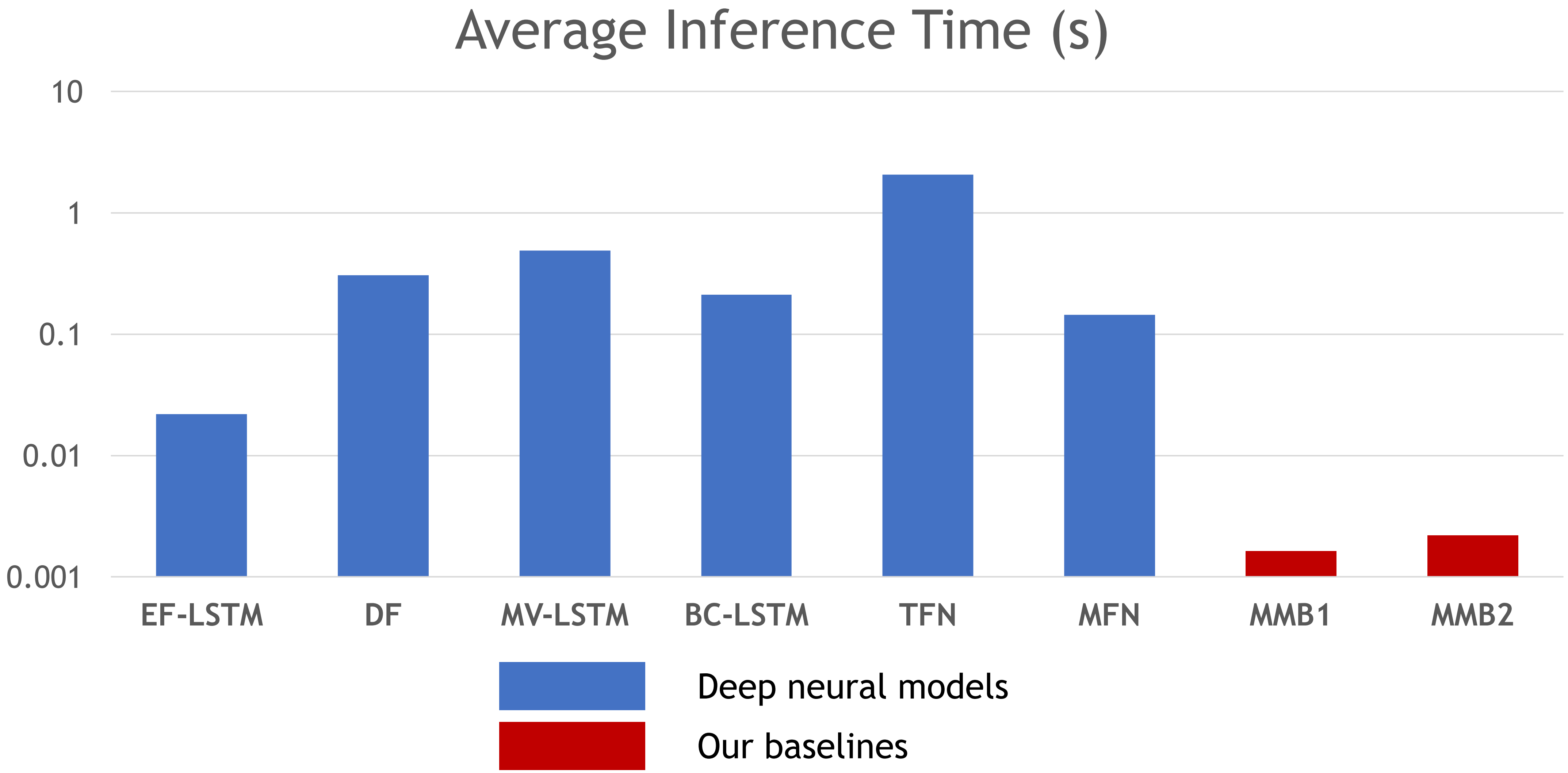
Experiments



Experiments



Speed Comparisons



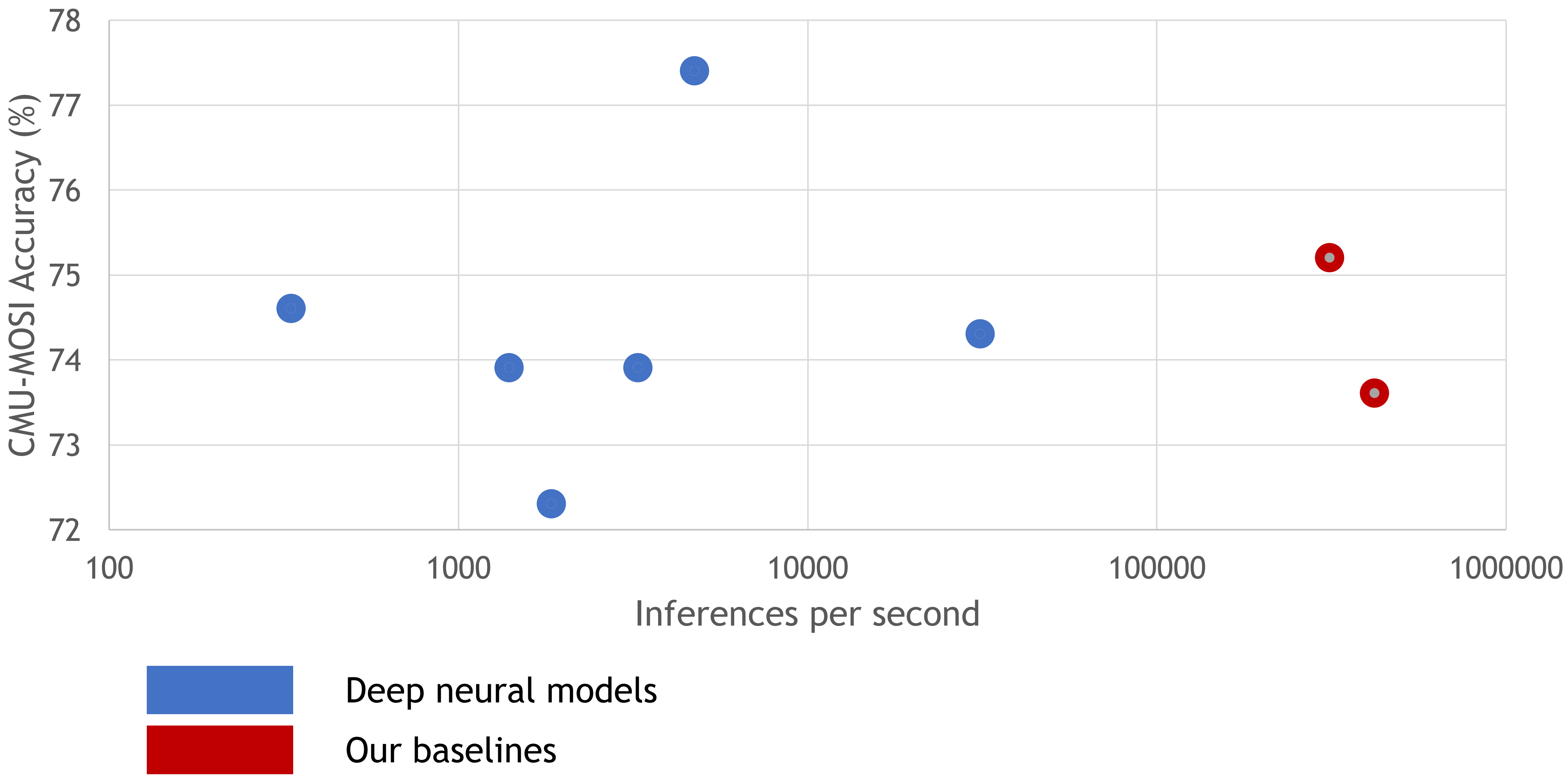
Conclusion

- Proposed two simple but strong baselines for learning embeddings of multimodal utterances
- Try strong baselines before working on complicated models!

Github: yaochie/multimodal-baselines

The End!

Email:
pliang@cs.cmu.edu
yaochonl@cs.cmu.edu



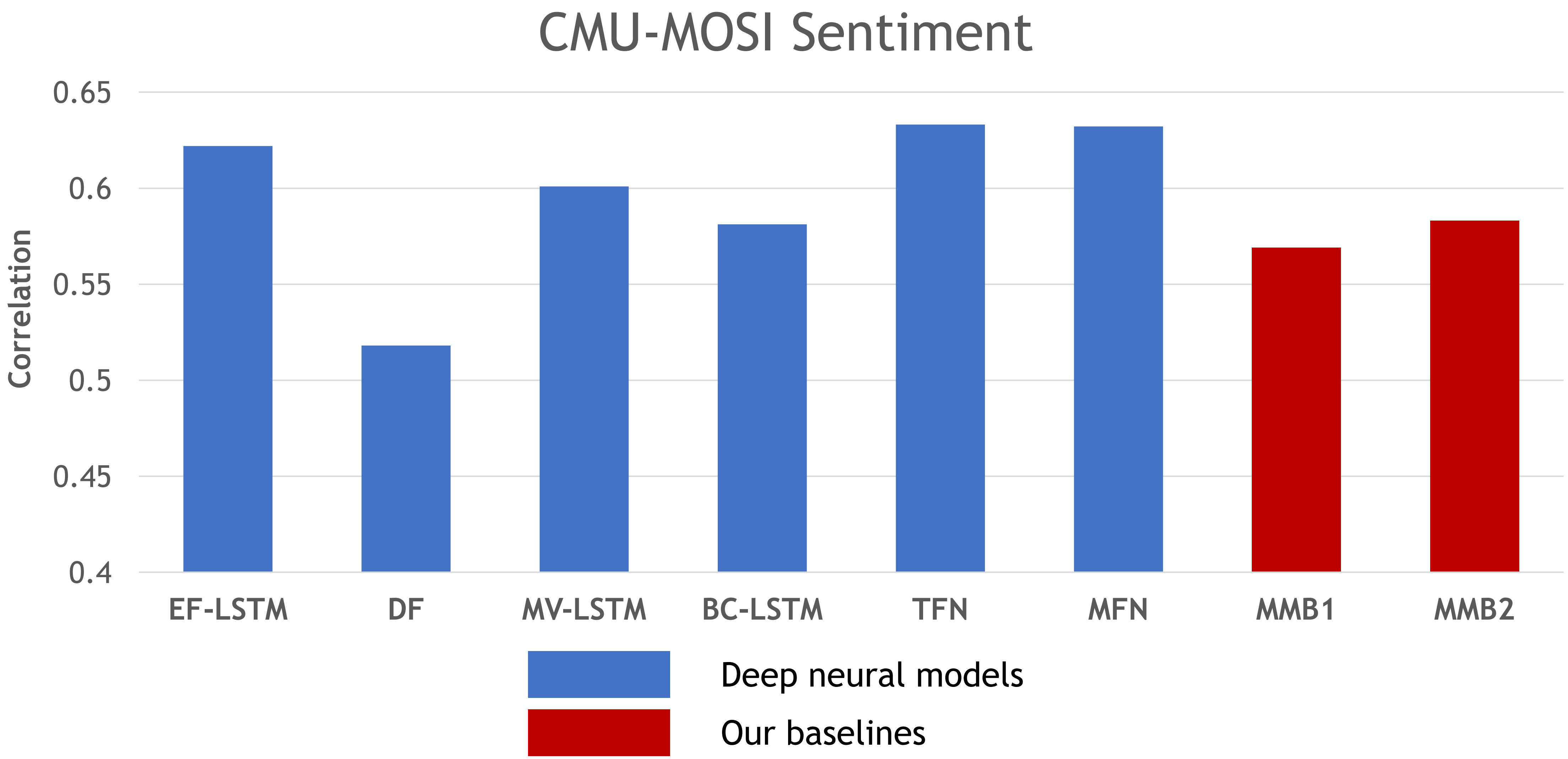
Additional Results

Dataset	CMU-MOSI	
Task	**Sentiment**	
Metric	A(2)	F1
Majority	50.2	50.1
RF	56.4	56.3
THMM	50.7	45.4
EF-HCRF$^{(\star)}$	65.3	65.4
MV-HCRF$^{(\star)}$	65.6	65.7
SVM-MD	71.6	72.3
C-MKL	72.3	72.0
DF	72.3	72.1
SAL-CNN	73.0	72.6
EF-LSTM$^{(\star)}$	74.3	74.3
MV-LSTM	73.9	74.0
BC-LSTM	73.9	73.9
TFN	74.6	74.5
MFN	**77.4**	**77.3**
MMB1	73.6	73.4
MMB2	**75.2**	**75.1**

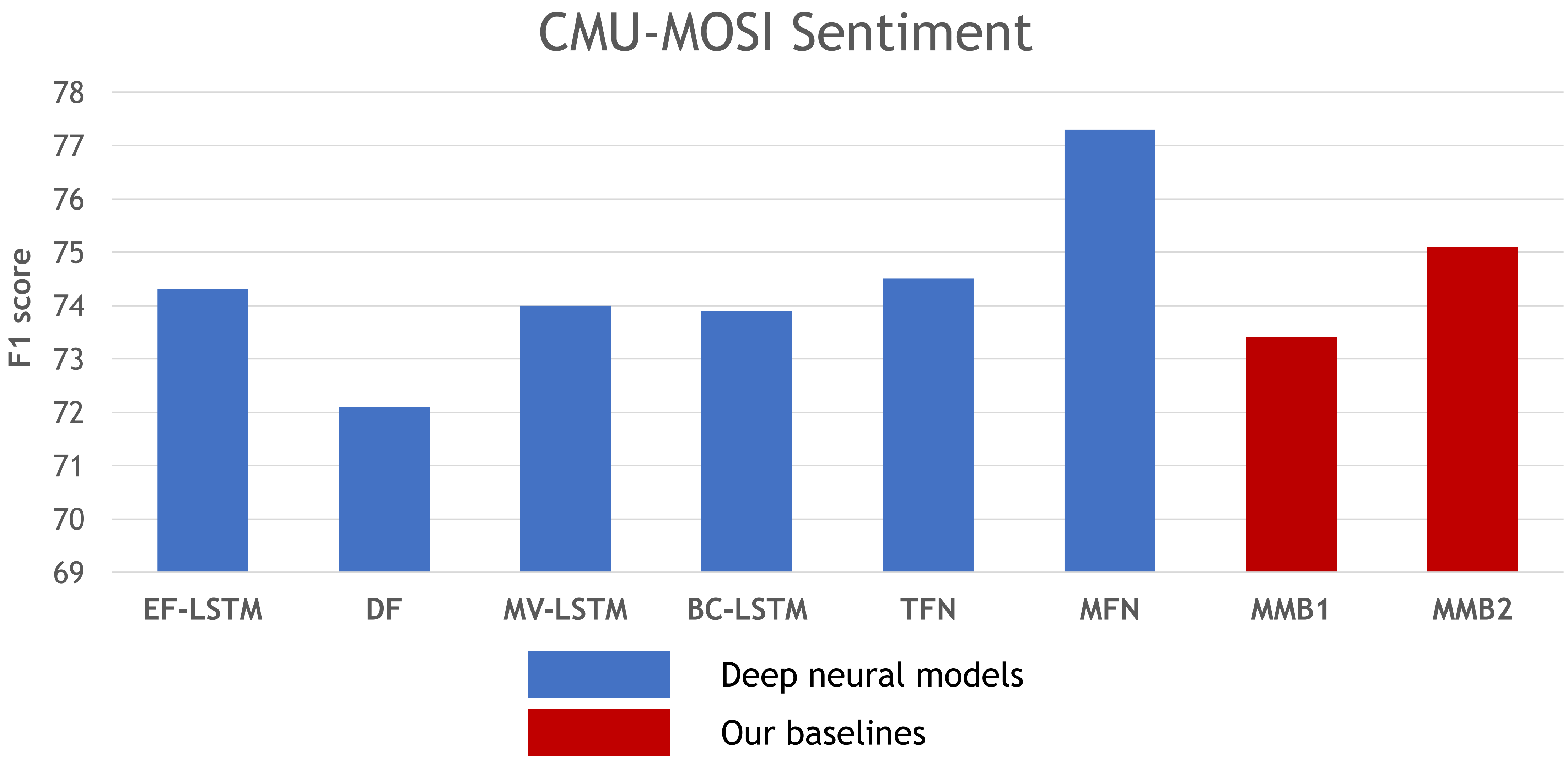
Dataset	**POM Personality Trait Recognition, measured in** MAE										
Task	Con	Voi	Dom	Viv	Res	Tru	Rel	Out	Tho	Ner	Hum
Majority	1.483	1.089	1.167	1.158	1.166	0.743	0.753	0.872	0.939	1.181	1.774
SVM	1.071	0.938	0.865	1.043	0.877	0.536	0.594	0.702	0.728	0.714	0.801
DF	1.033	0.899	0.870	0.997	0.884	0.534	0.591	0.698	0.732	0.695	0.768
EF-LSTM$^{(\star)}$	1.035	0.911	0.880	0.981	0.872	0.556	0.594	0.700	0.712	0.706	0.762
MV-LSTM	1.029	0.971	0.944	0.976	0.877	0.523	0.625	0.703	0.792	**0.687**	0.770
BC-LSTM	1.016	0.914	**0.859**	**0.905**	0.888	0.564	0.630	0.708	**0.680**	0.705	0.767
TFN	1.049	0.927	0.864	1.000	0.900	0.572	0.621	0.706	0.743	0.727	0.770
MFN	**0.952**	**0.882**	**0.835**	**0.908**	**0.821**	**0.521**	**0.566**	**0.679**	**0.665**	**0.654**	**0.727**
MMB2	**1.015**	**0.878**	0.885	0.967	**0.857**	**0.522**	**0.578**	**0.685**	0.705	0.692	**0.726**

Dataset	POM Personality Trait Recognition, measured in r										
Task	Con	Voi	Dom	Viv	Res	Tru	Rel	Out	Tho	Ner	Hum
Majority	-0.041	-0.104	-0.031	-0.044	0.006	-0.077	-0.024	-0.085	-0.130	0.097	-0.069
SVM	0.063	-0.004	0.141	0.076	0.134	0.168	0.104	0.066	0.134	0.068	0.147
DF	0.240	0.017	0.139	0.173	0.118	0.143	0.019	0.093	0.041	0.136	0.259
EF-LSTM$^{(\star)}$	0.221	0.042	0.151	0.239	0.268	0.069	0.092	0.215	0.252	0.159	0.272
MV-LSTM	0.358	0.131	0.146	0.347	0.323	**0.237**	0.119	0.238	0.284	0.258	0.317
BC-LSTM	**0.359**	0.081	0.234	0.417	0.450	0.109	0.075	0.078	0.363	0.184	0.319
TFN	0.089	0.030	0.020	0.204	-0.051	-0.064	0.114	0.060	0.048	-0.002	0.213
MFN	**0.395**	**0.193**	**0.313**	**0.431**	**0.333**	**0.296**	**0.255**	**0.259**	**0.381**	**0.318**	**0.386**
MMB2	0.350	**0.220**	**0.333**	**0.434**	**0.332**	0.176	**0.224**	**0.318**	**0.394**	**0.296**	**0.366**

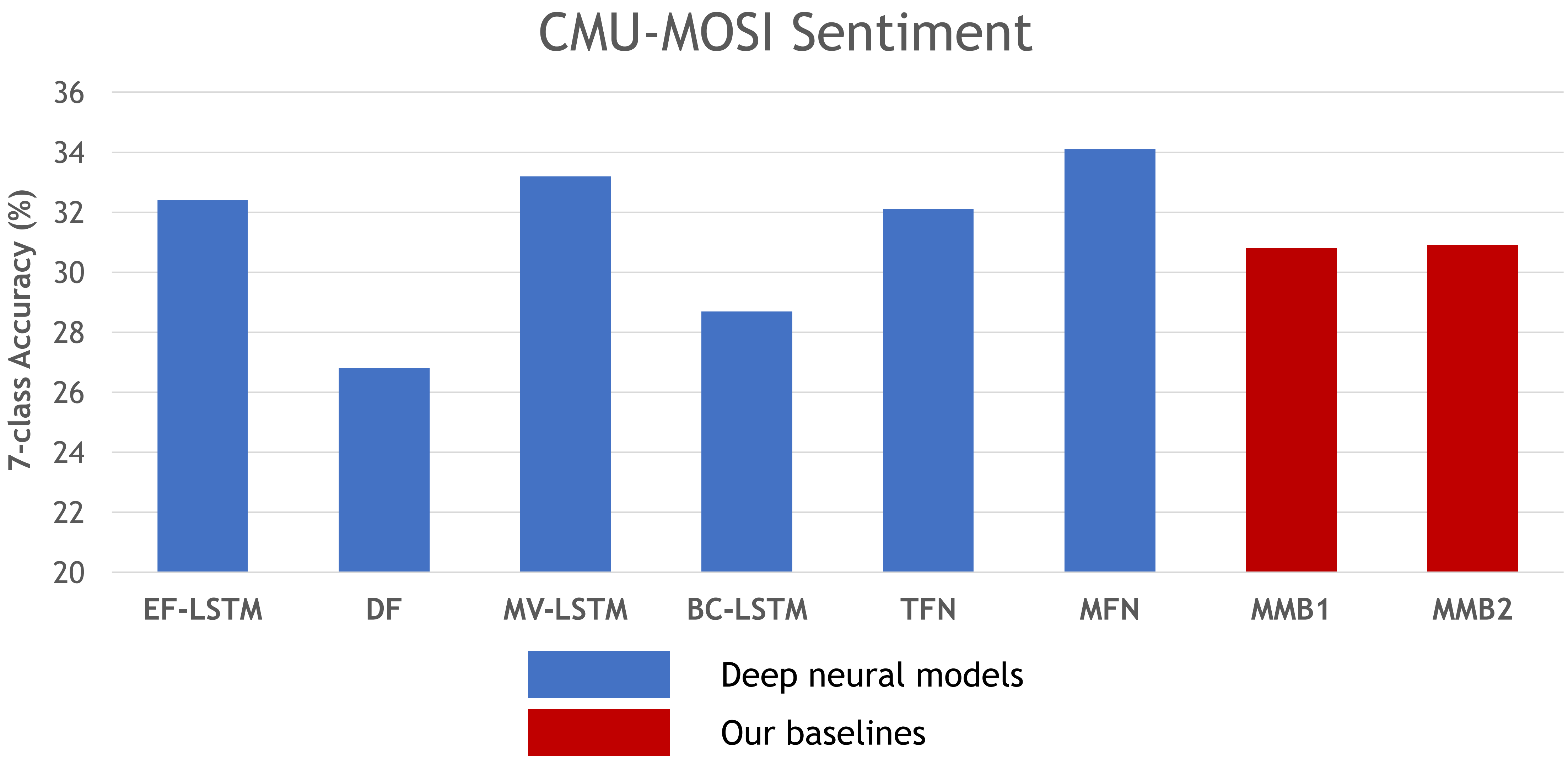
Experiments



Experiments



Experiments



Experiments