MACHINE LEARNING DEPARTMENT

Language Technologies Institute

Carnegie Mellon University

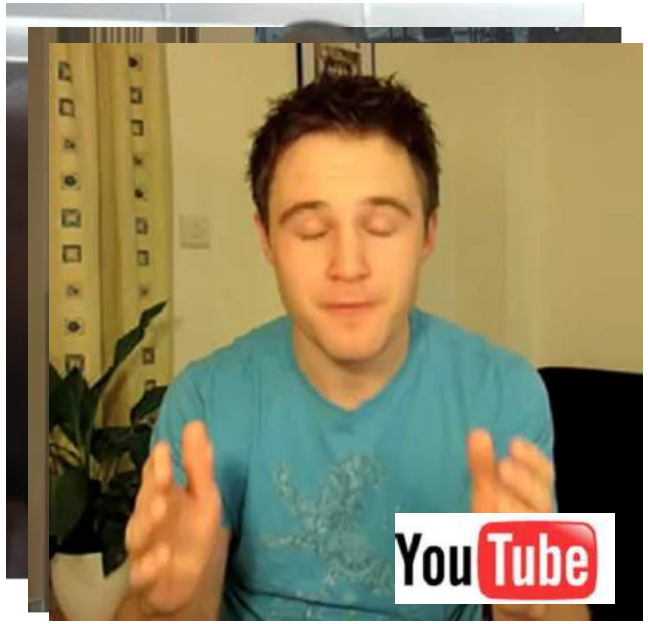
Learning Robust Joint Representations for Multimodal Sentiment Analysis

Presenter: Paul Pu Liang

Hai Pham*, Paul Pu Liang*, Thomas Manzini, Louis-Philippe Morency, Barnabás Póczos

Progress of Artificial Intelligence

Multimedia Content



Intelligent
Personal Assistants

Robots and
Virtual Agents

Multimodal Language Modalities

Language

- Lexicon
- Syntax
- Pragmatics

Visual

- Gestures
- Body language
- Eye contact
- Facial expressions

Acoustic

- Prosody
- Vocal expressions

Multimodal Language Modalities

Language

- Lexicon
- Syntax
- Pragmatics

Visual

- Gestures
- Body language
- Eye contact
- Facial expressions

Acoustic

- Prosody
- Vocal expressions

Sentiment

- Positive
- Negative

Emotion

- Anger
- Disgust
- Fear
- Happiness
- Sadness
- Surprise

Personality

- Confidence
- Persuasion
- Passion

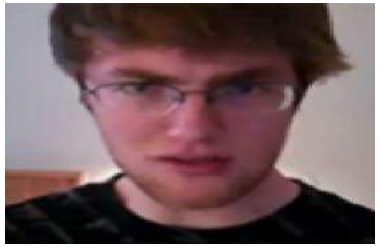
Challenge 1: Intra-modal Interactions

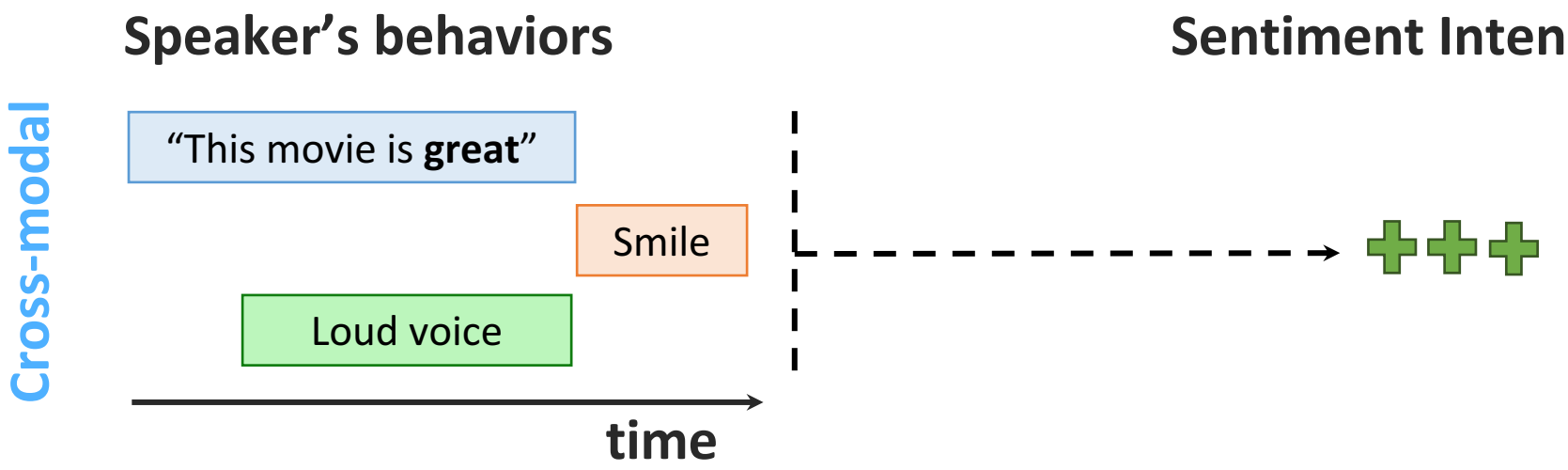
a) Temporal sequences

Challenge 2: Cross-modal Interactions

a) Multiple co-occurring interactions

b) Different weighted combinations





Learning Joint Representations: 2 modalities

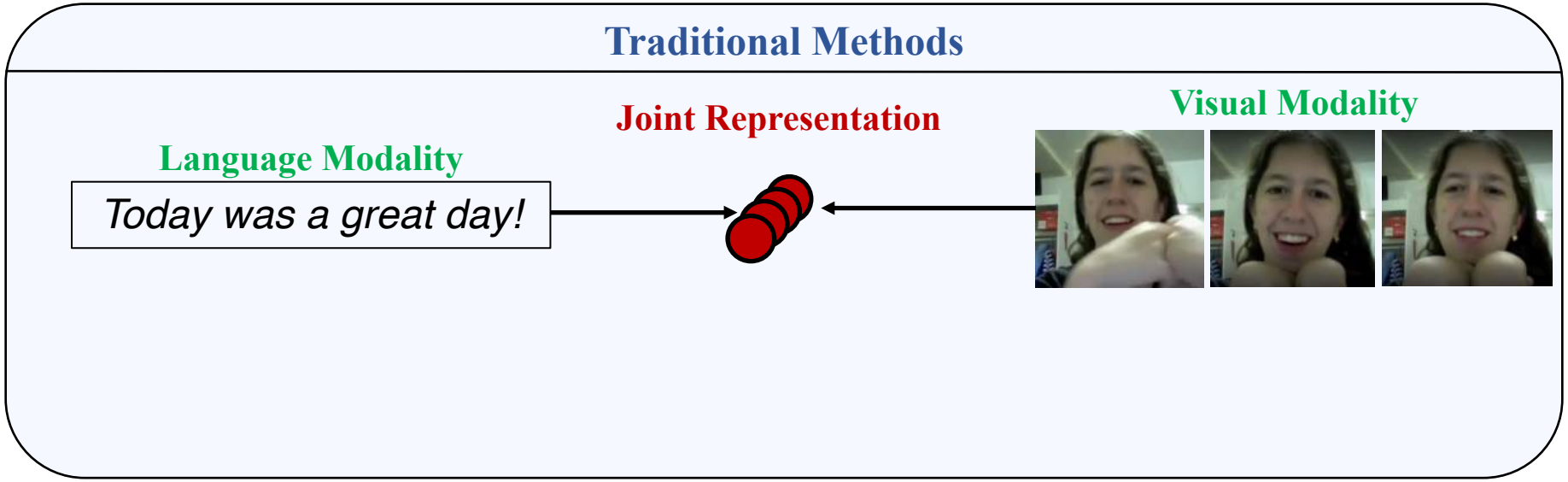
Traditional Methods

Language Modality

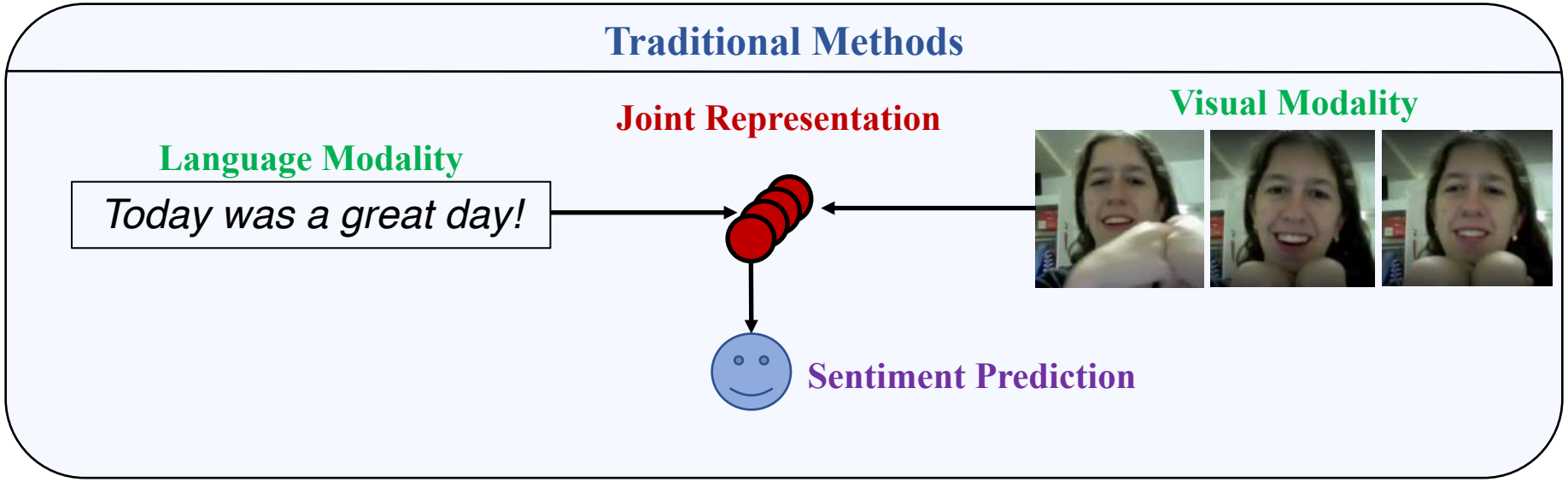
Today was a great day!

Visual Modality

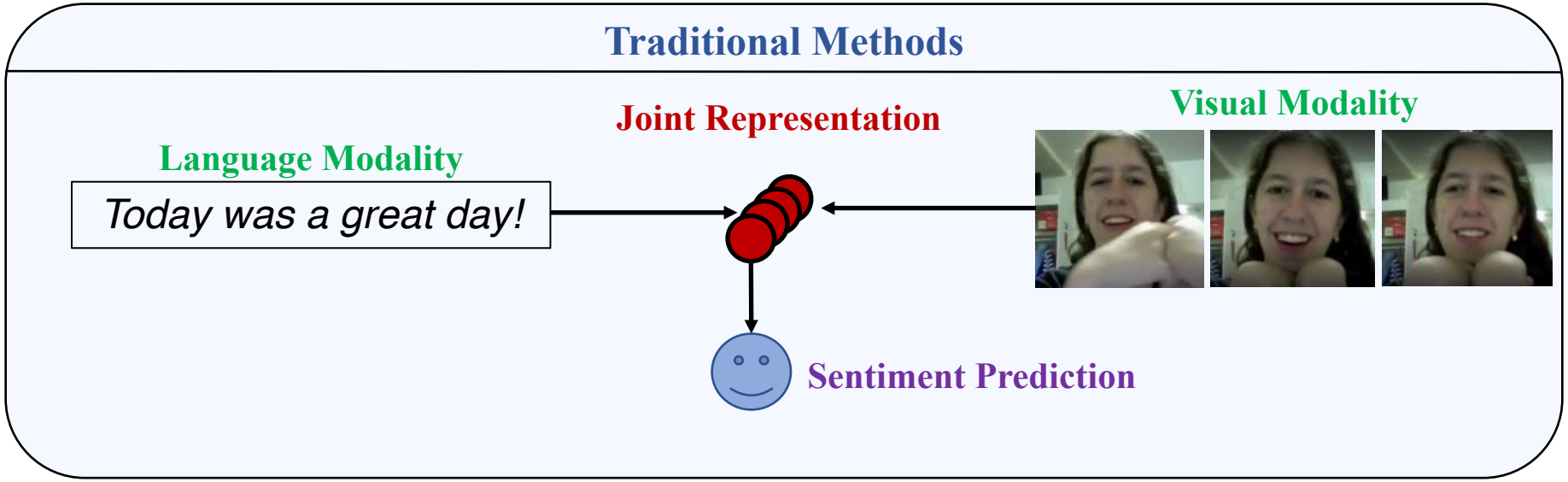
Learning Joint Representations: 2 modalities



Learning Joint Representations: 2 modalities

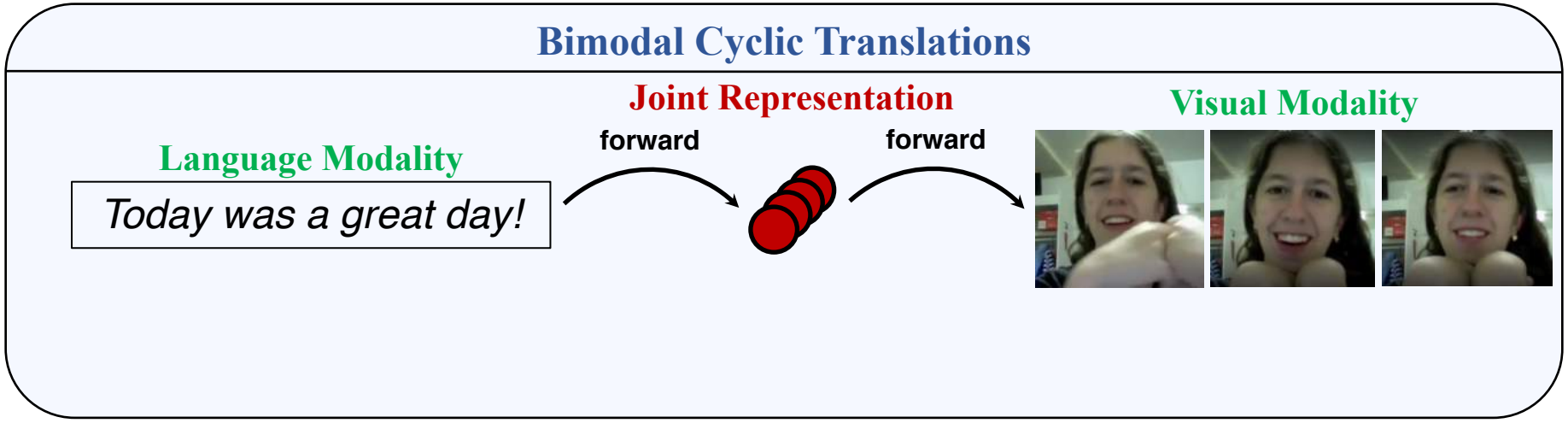


Learning Joint Representations: 2 modalities

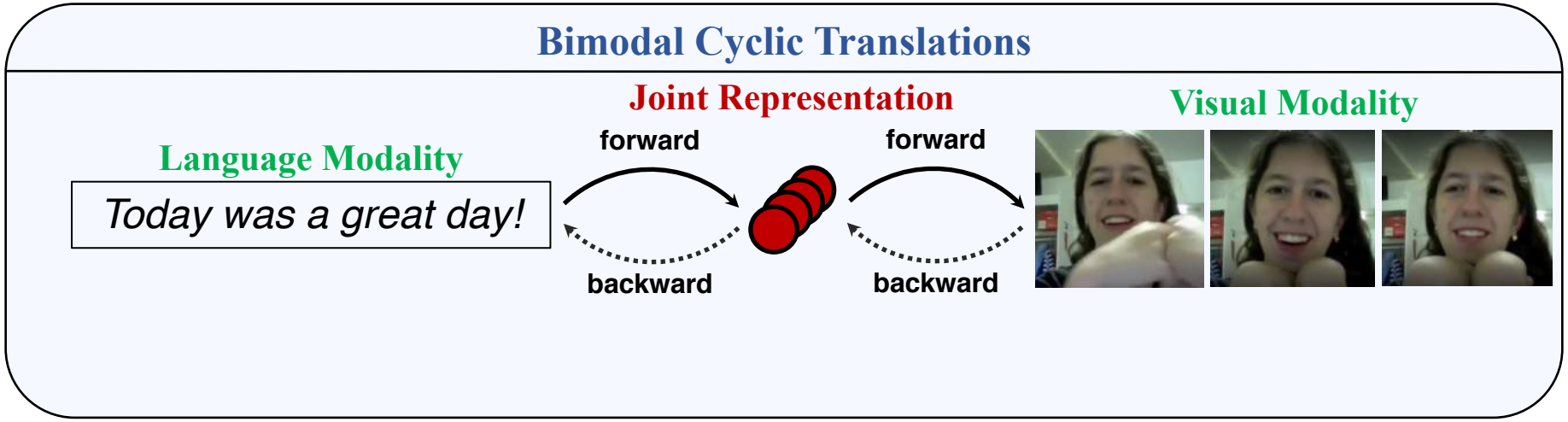


Both modalities required at test time!
Sensitive to missing/noisy visual modality.

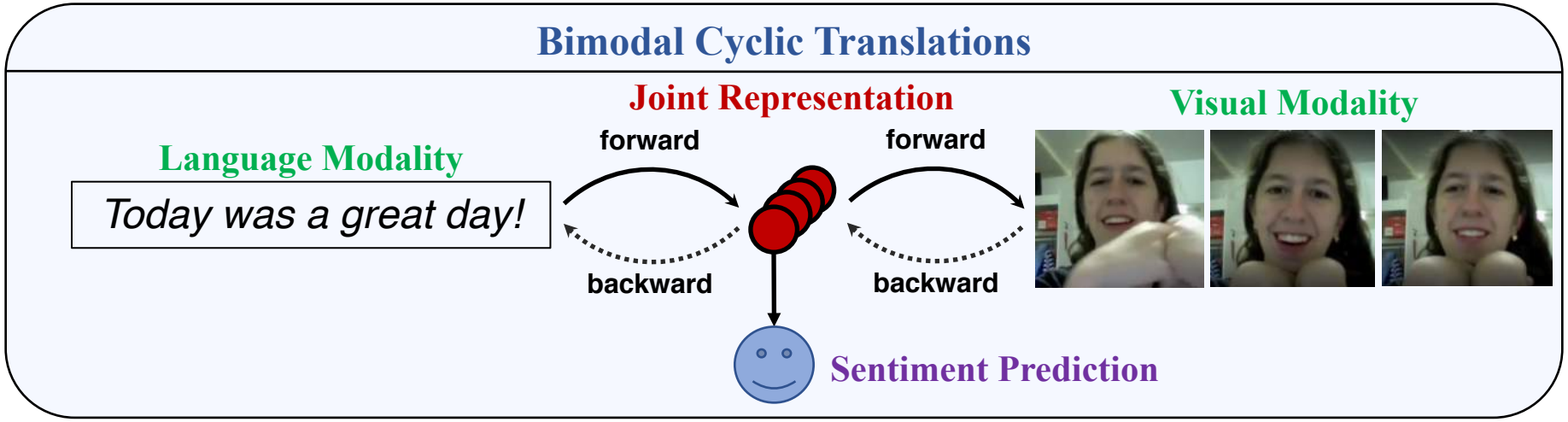
Learning Robust Joint Representations: 2 modalities



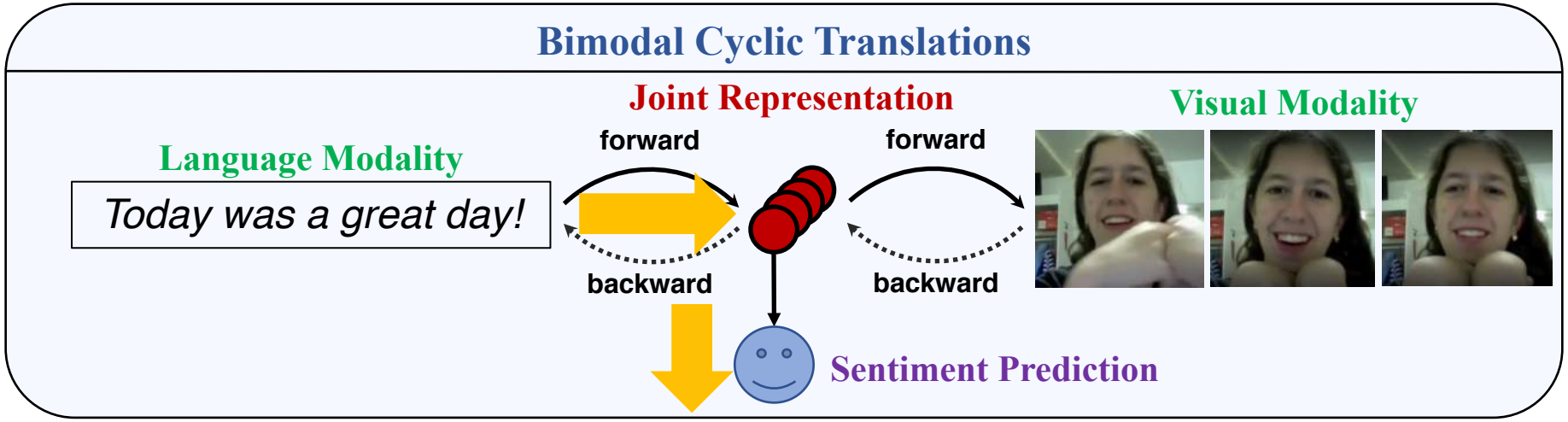
Learning Robust Joint Representations: 2 modalities



Learning Robust Joint Representations: 2 modalities



Learning Robust Joint Representations: 2 modalities



Only language modality required at test time!

Learning Robust Joint Representations: 3 modalities

Trimodal Cyclic Translations

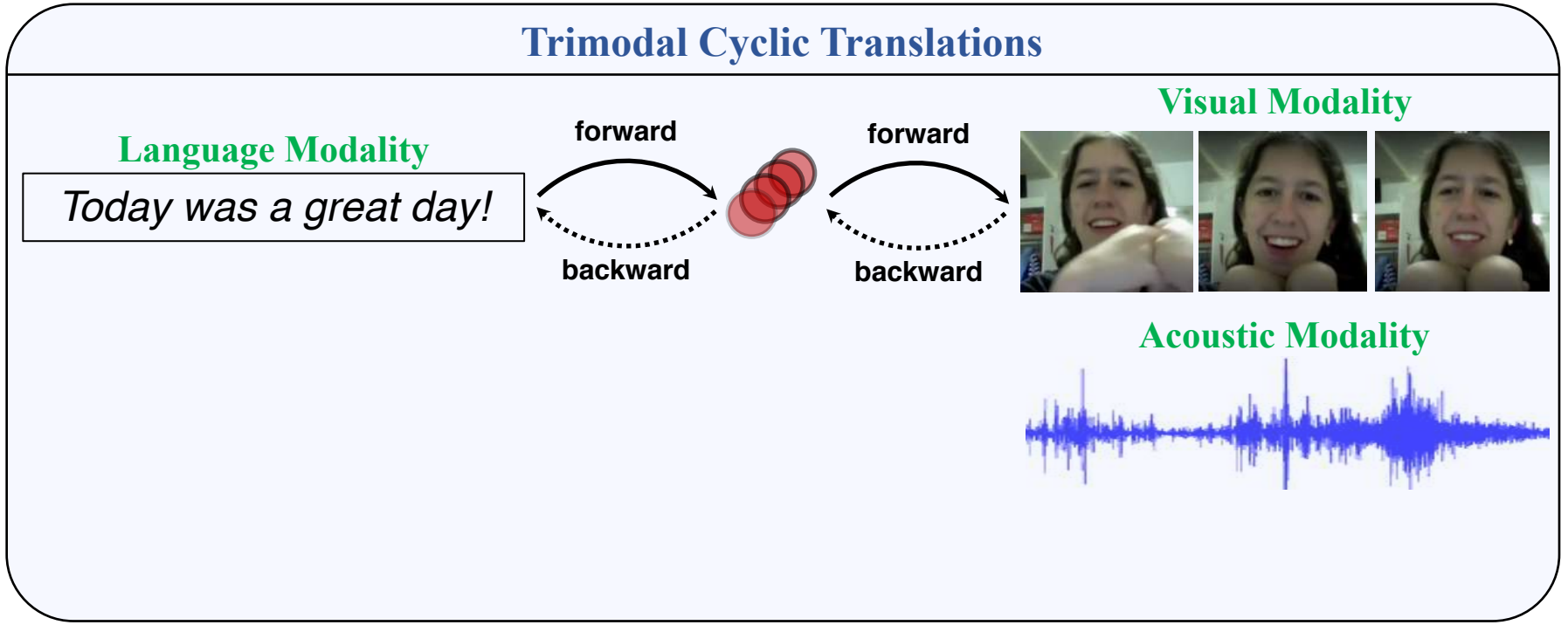
Language Modality

Today was a great day!

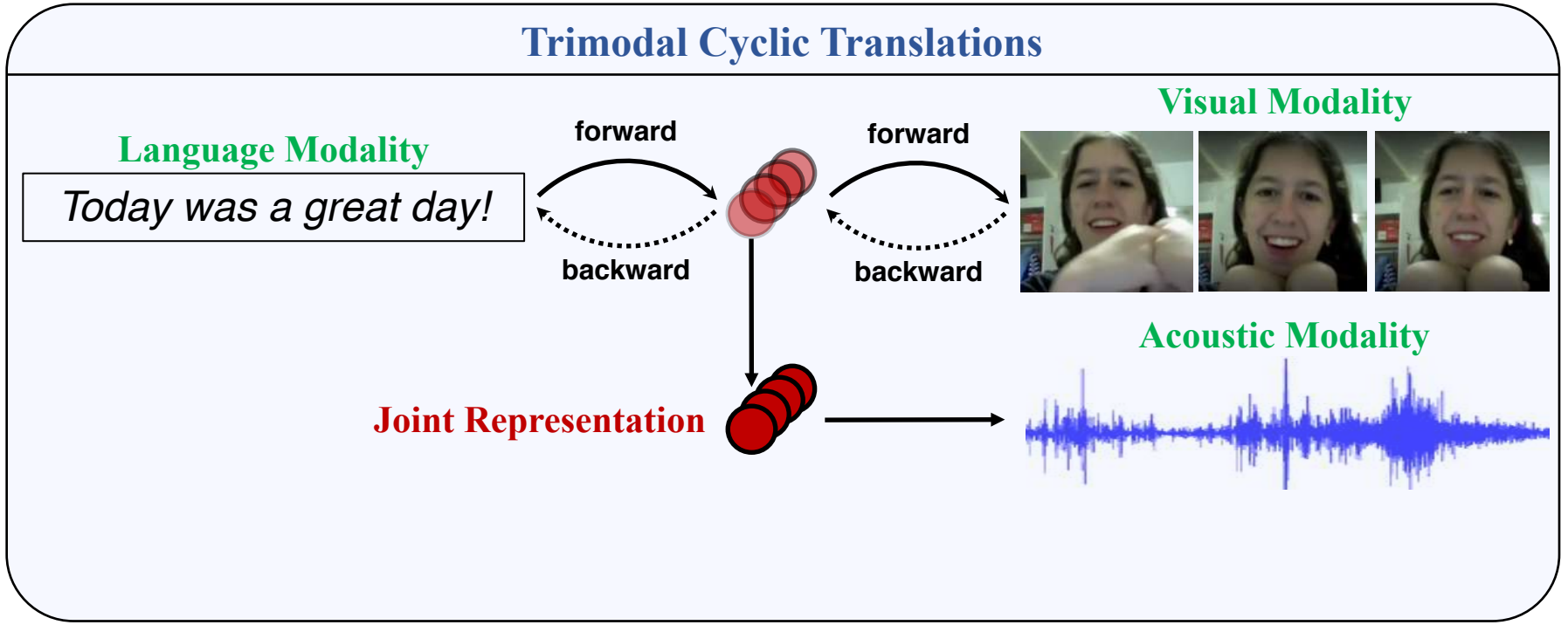
Visual Modality

Acoustic Modality

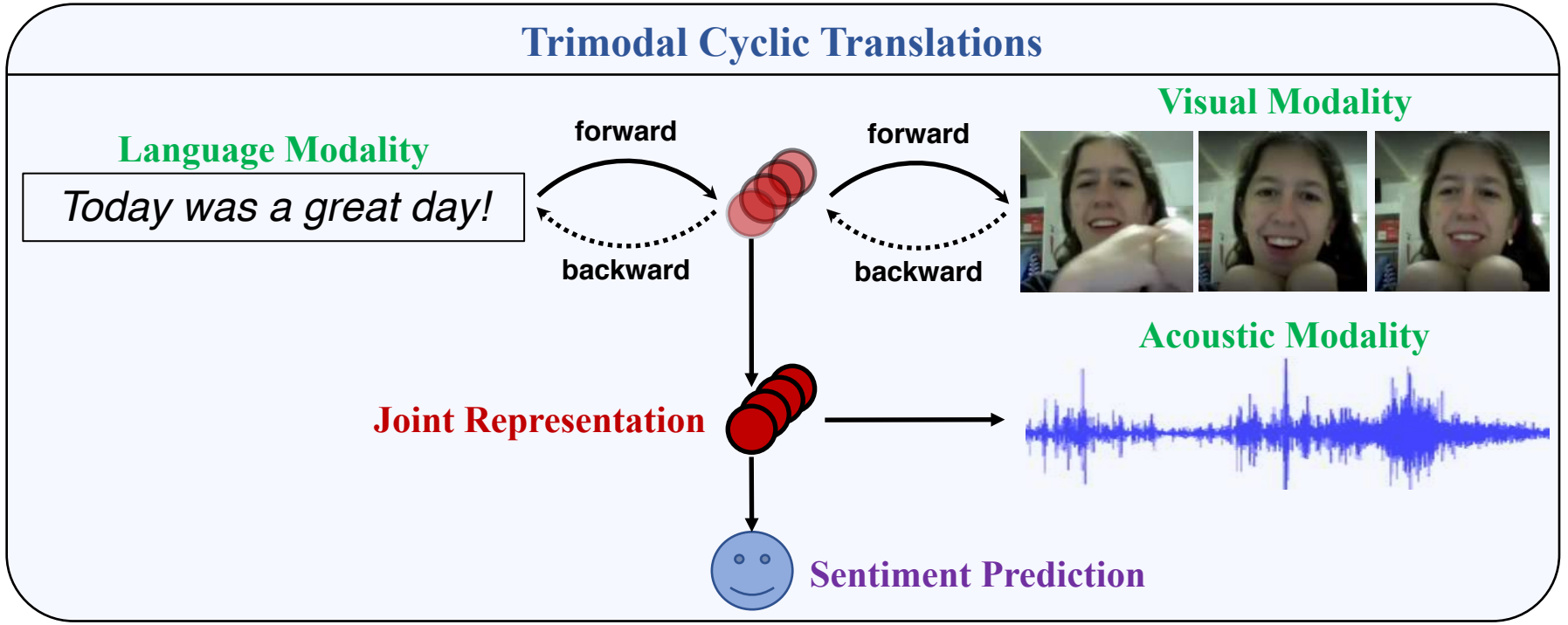
Learning Robust Joint Representations: 3 modalities



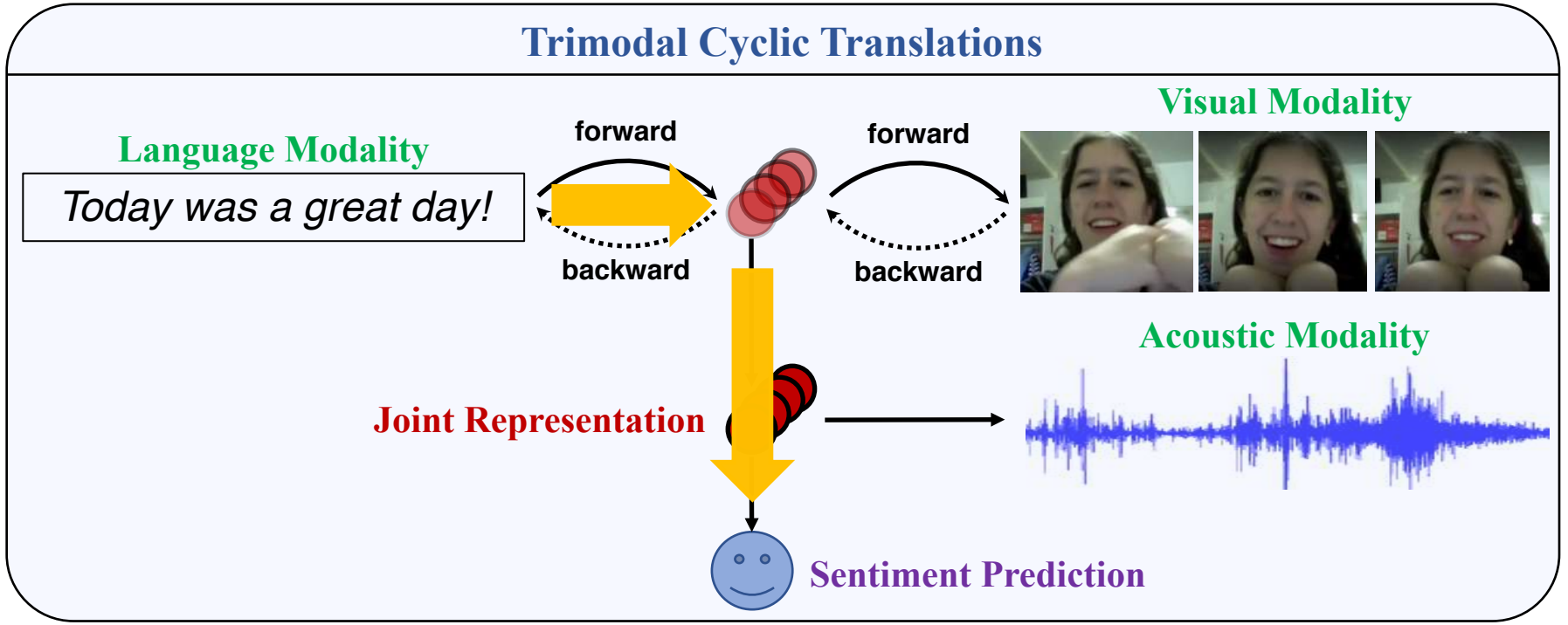
Learning Robust Joint Representations: 3 modalities



Learning Robust Joint Representations: 3 modalities

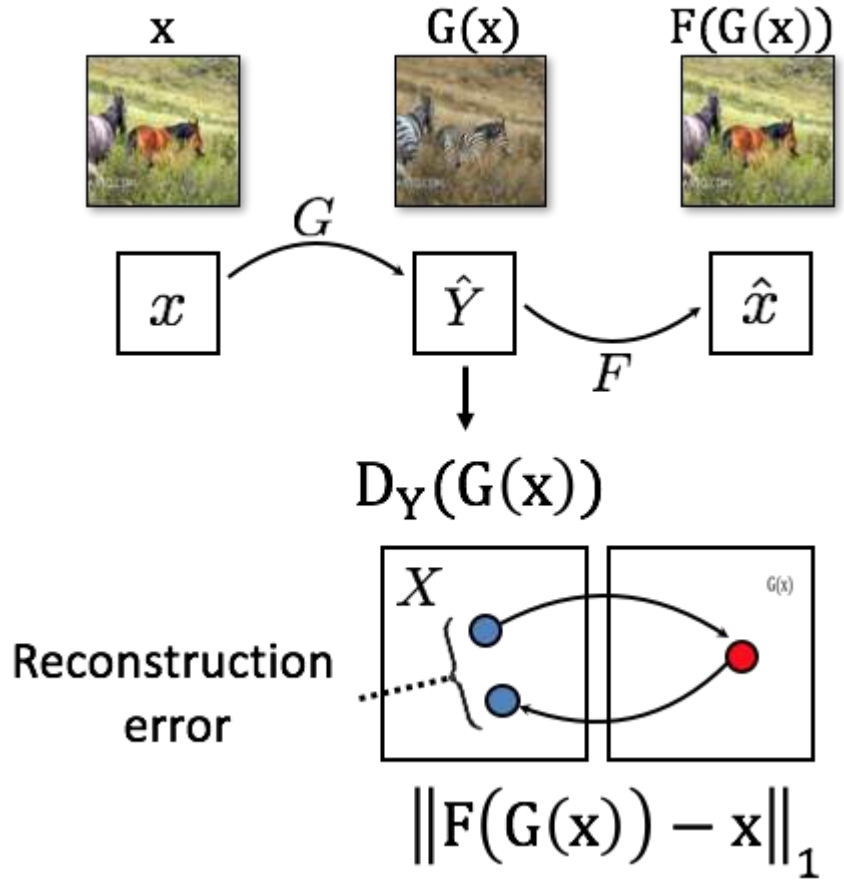


Learning Robust Joint Representations: 3 modalities



Only language modality required at test time!

Cyclic Translations



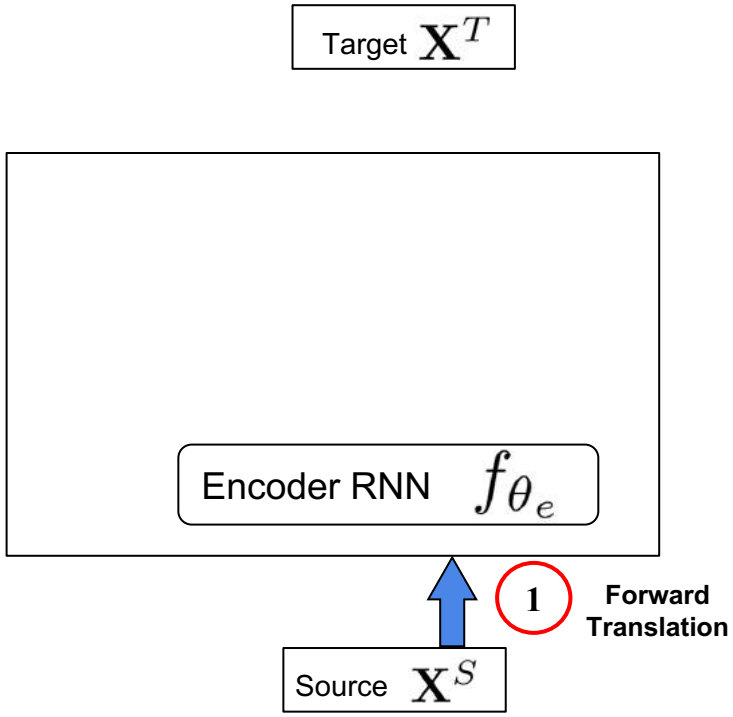
[Zhu*, Park*, Isola, and Efros, ICCV 2017]

Multimodal Cyclic Translation Network

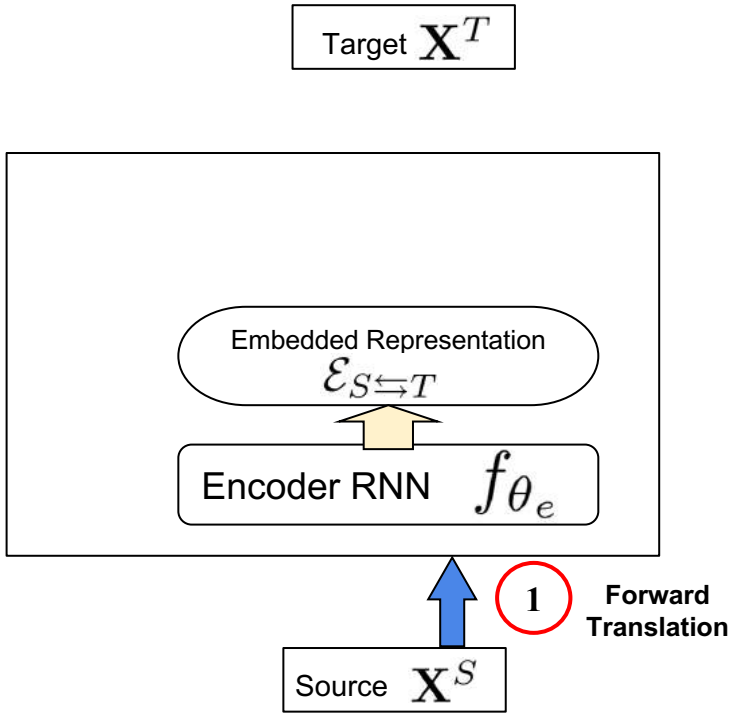
Target $\mathbf{X}^T$

Source $\mathbf{X}^S$

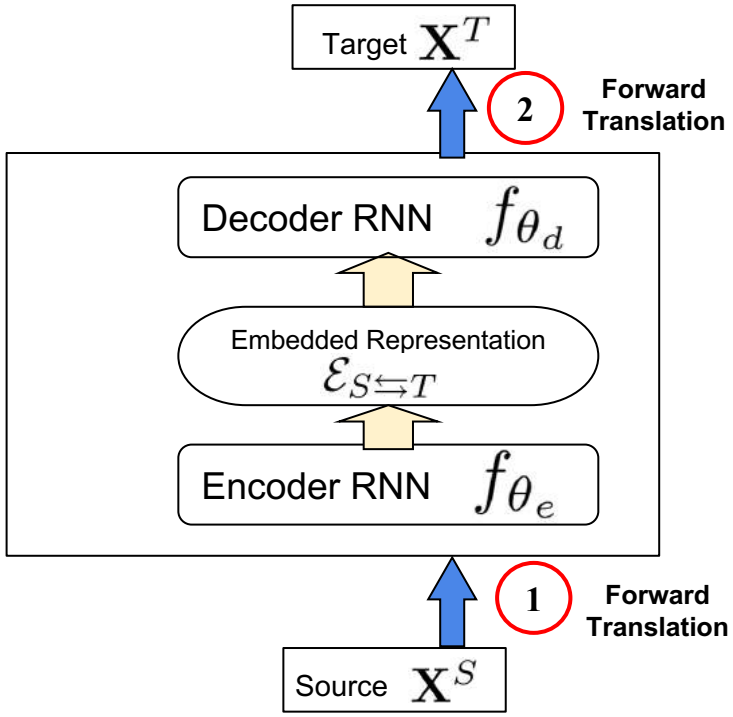
Multimodal Cyclic Translation Network



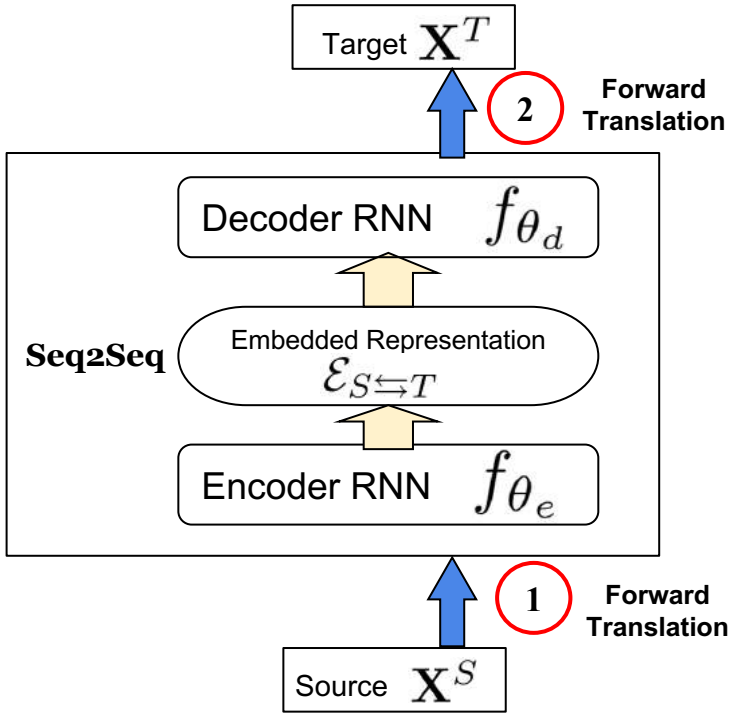
Multimodal Cyclic Translation Network



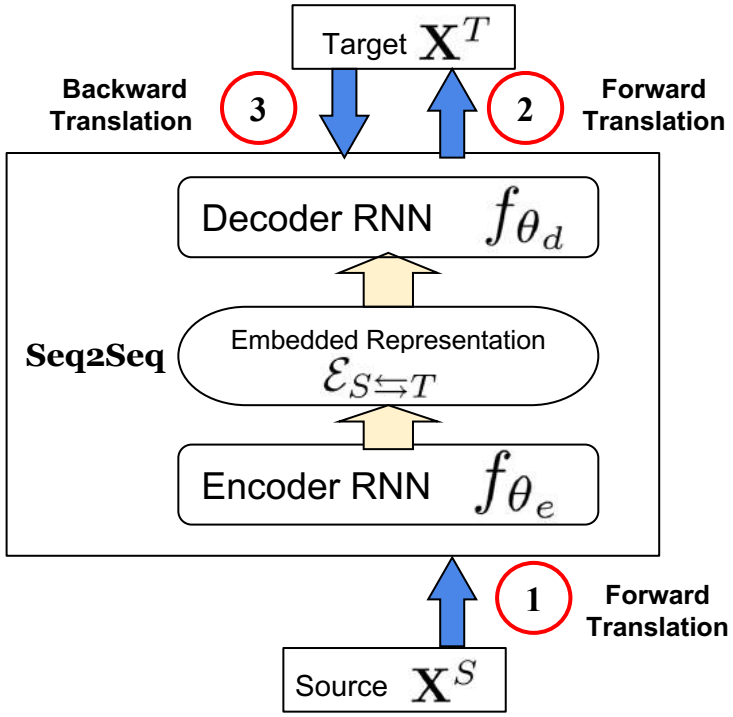
Multimodal Cyclic Translation Network



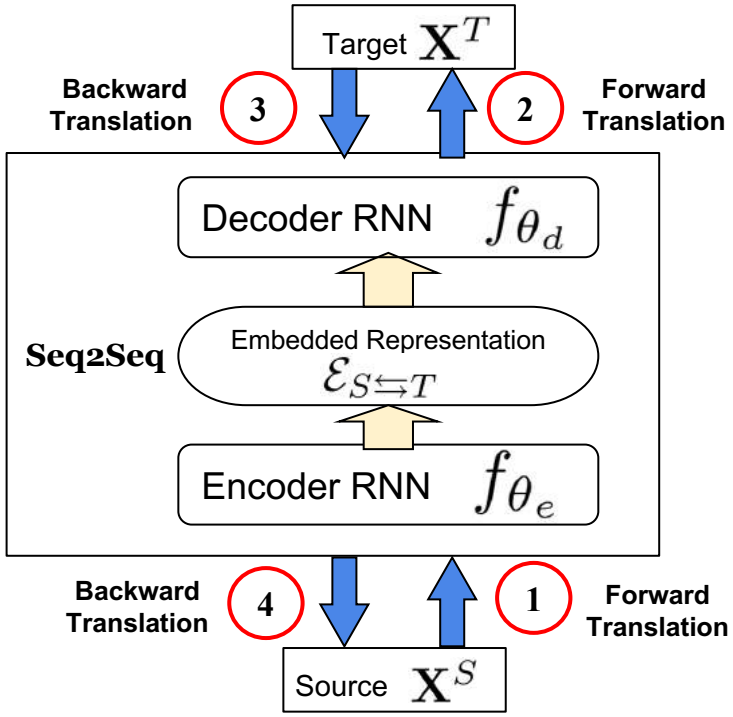
Multimodal Cyclic Translation Network



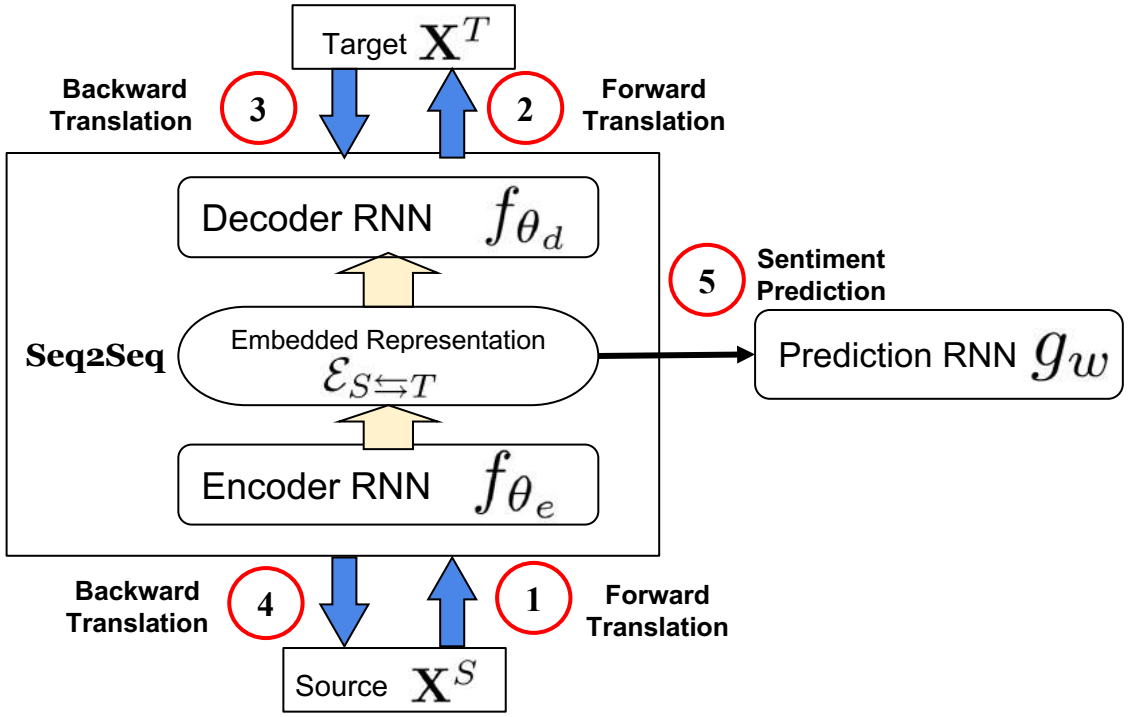
Multimodal Cyclic Translation Network



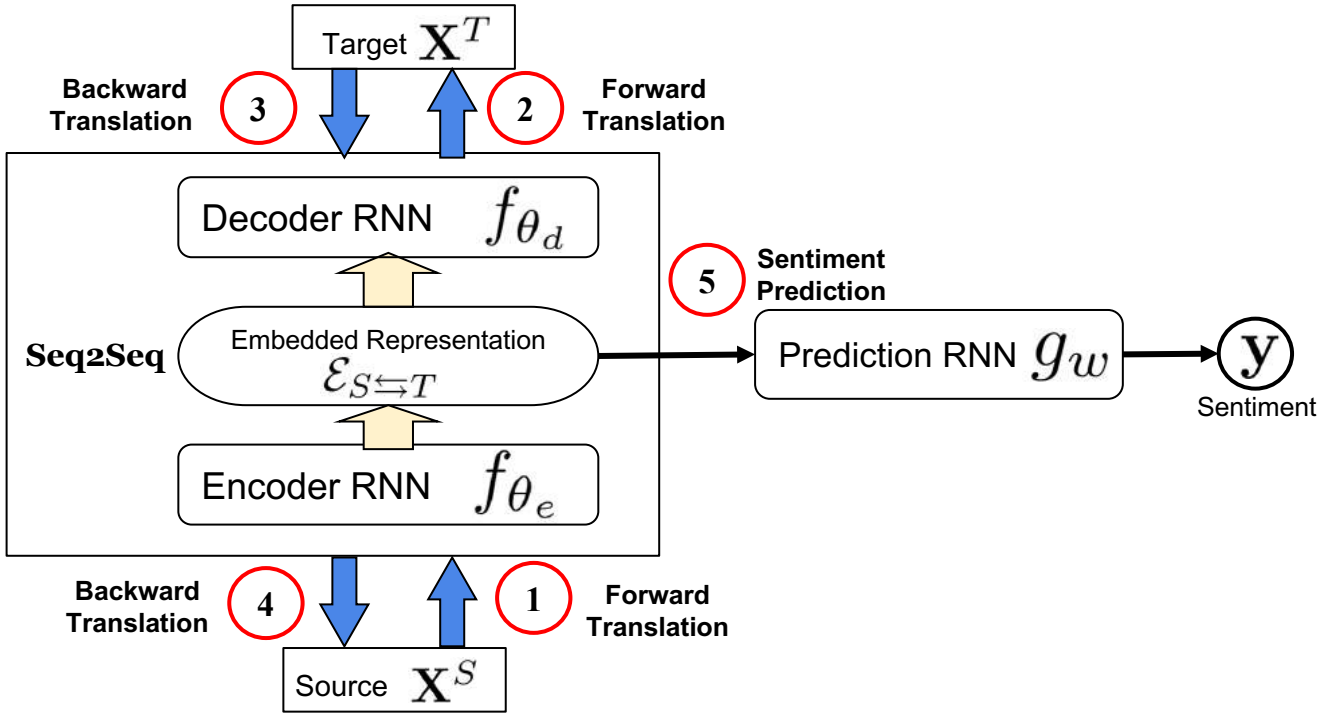
Multimodal Cyclic Translation Network



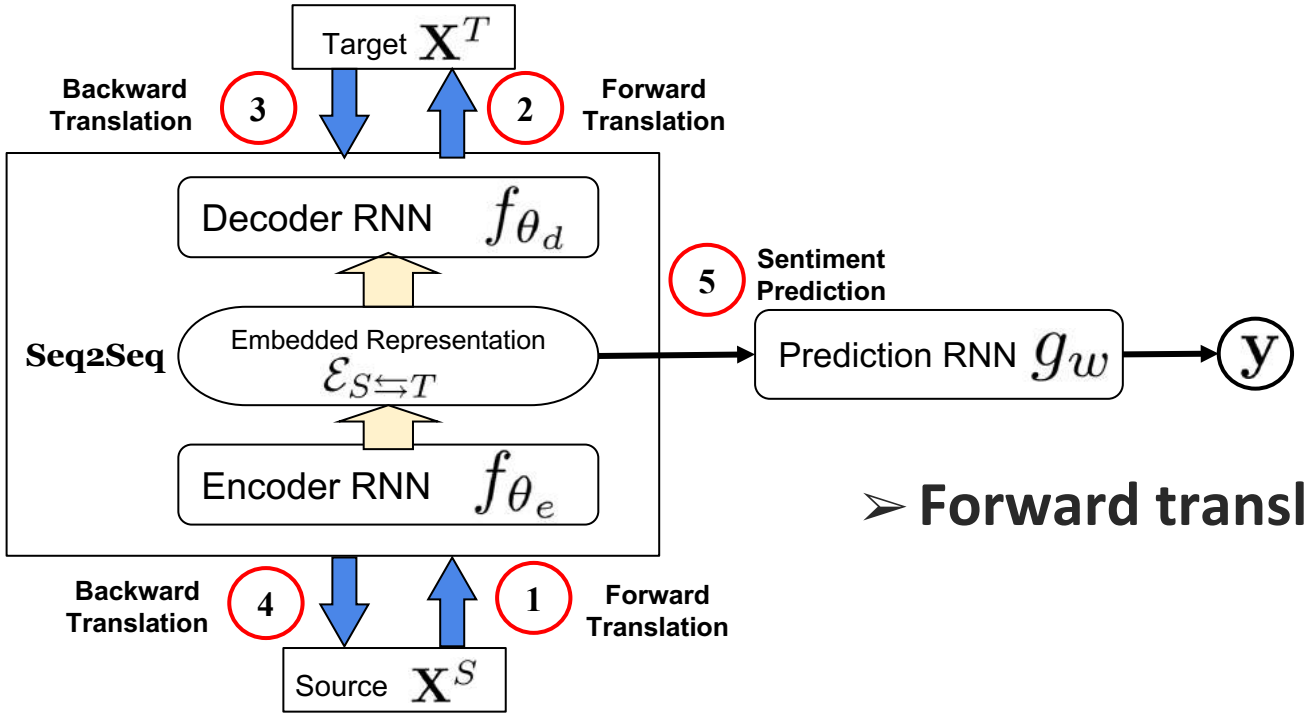
Multimodal Cyclic Translation Network



Multimodal Cyclic Translation Network

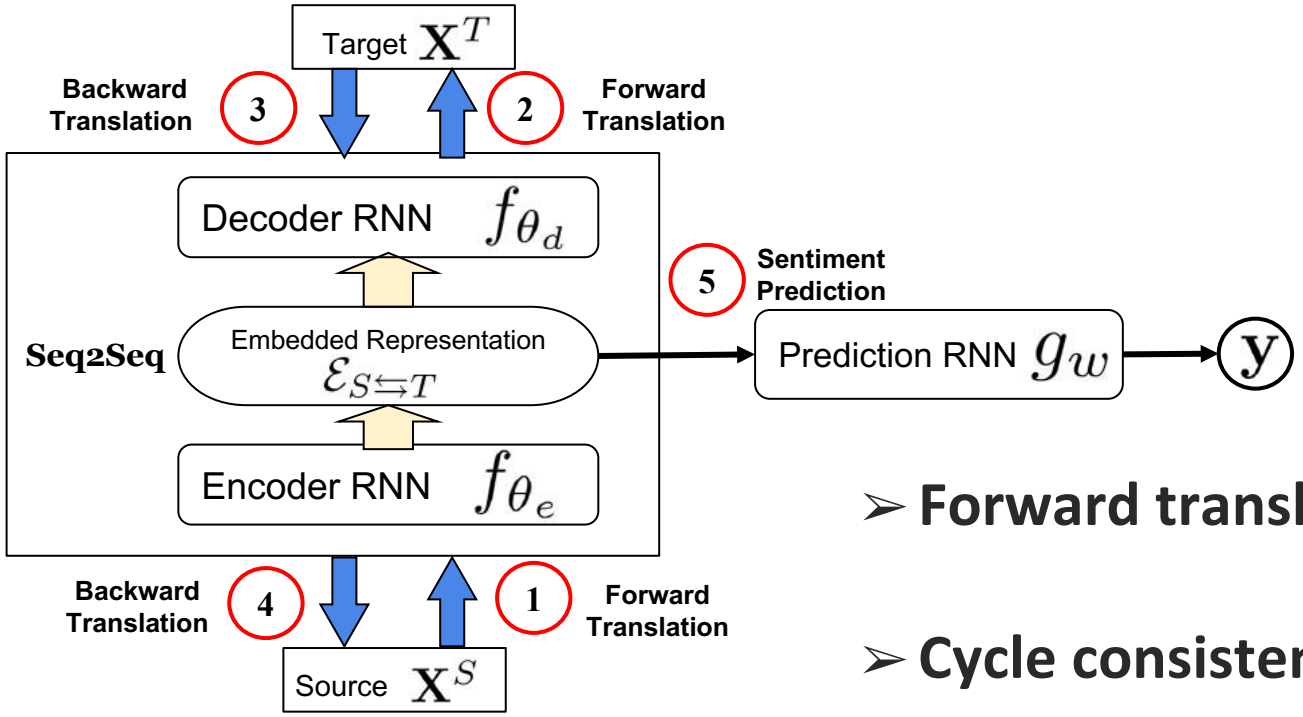


Coupled Translation-Prediction Objective



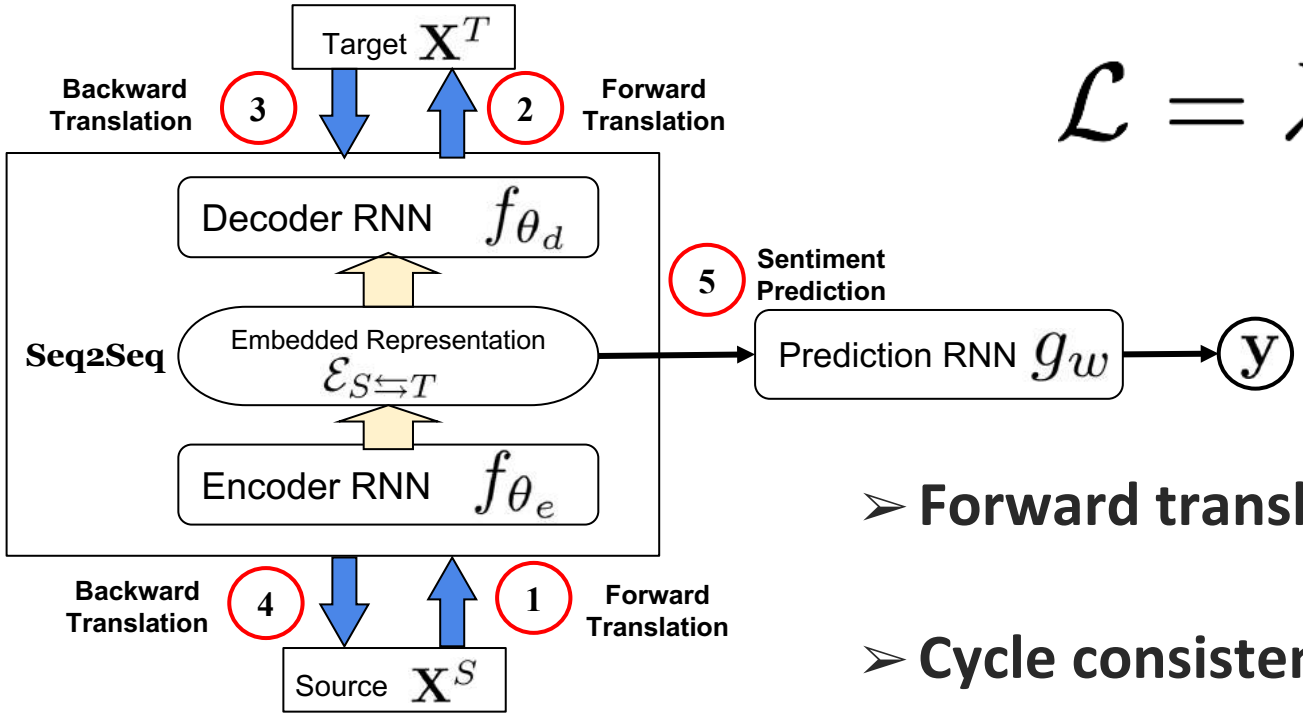
➢ **Forward translation loss** $\mathcal{L}_t = \mathbb{E}[\ell_{\mathbf{X}^T}(\hat{\mathbf{X}}^T, \mathbf{X}^T)]$

Coupled Translation-Prediction Objective



- **Forward translation loss** $\mathcal{L}_t = \mathbb{E}[\ell_{\mathbf{X}^T}(\hat{\mathbf{X}}^T, \mathbf{X}^T)]$
- **Cycle consistent loss** $\mathcal{L}_c = \mathbb{E}[\ell_{\mathbf{X}^S}(\hat{\mathbf{X}}^S, \mathbf{X}^S)]$

Coupled Translation-Prediction Objective



$$\mathcal{L} = \lambda_t \mathcal{L}_t + \lambda_c \mathcal{L}_c + \mathcal{L}_p$$

- **Forward translation loss** $\mathcal{L}_t = \mathbb{E}[\ell_{\mathbf{X}^T}(\hat{\mathbf{X}}^T, \mathbf{X}^T)]$
- **Cycle consistent loss** $\mathcal{L}_c = \mathbb{E}[\ell_{\mathbf{X}^S}(\hat{\mathbf{X}}^S, \mathbf{X}^S)]$
- **Prediction loss** $\mathcal{L}_p = \mathbb{E}[\ell_{\mathbf{y}}(\hat{\mathbf{y}}, \mathbf{y})]$

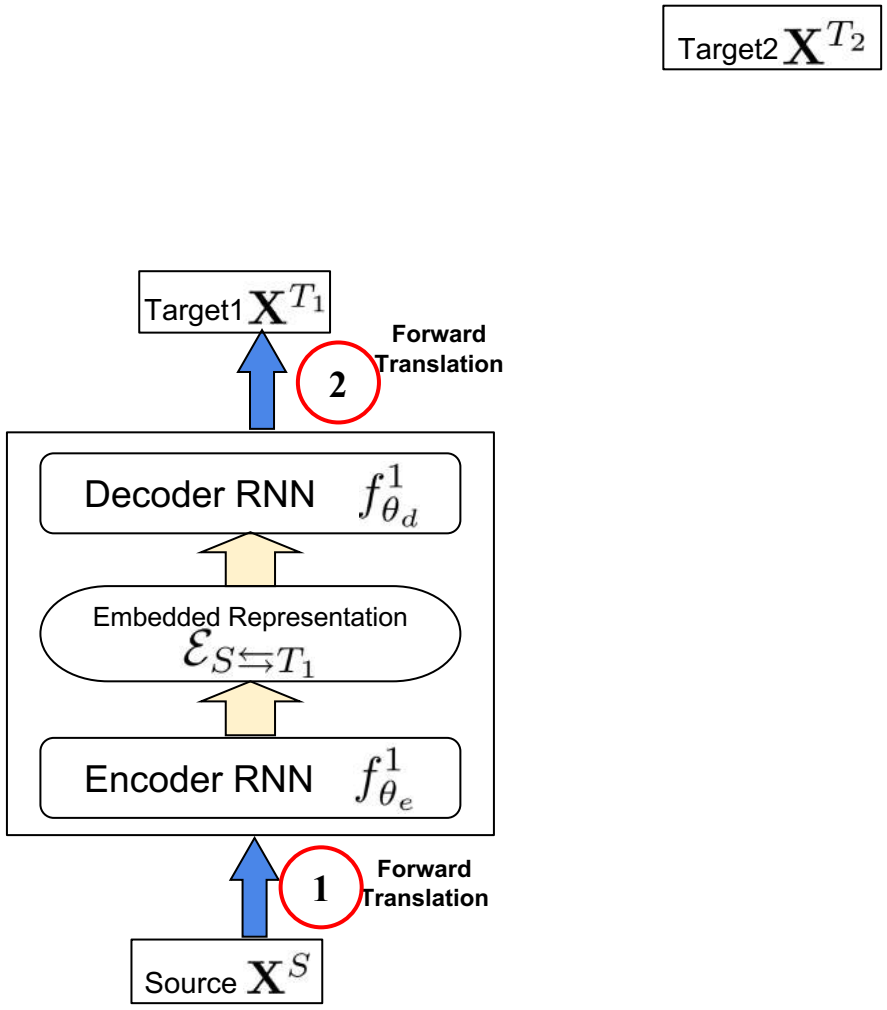
Hierarchical Multimodal Cyclic Translation Network

Target2 $\mathbf{X}^{T_2}$

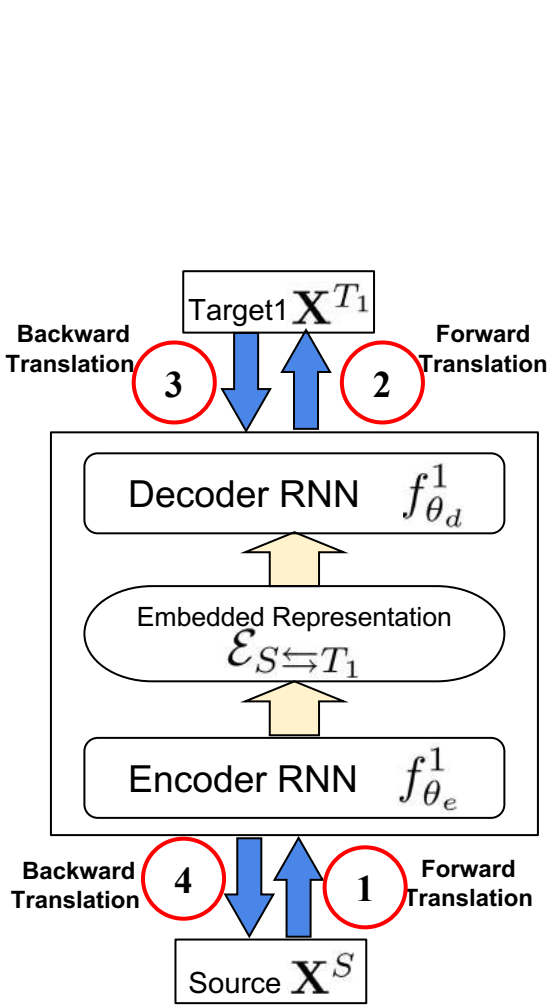
Target1

Source $\mathbf{X}^{S}$

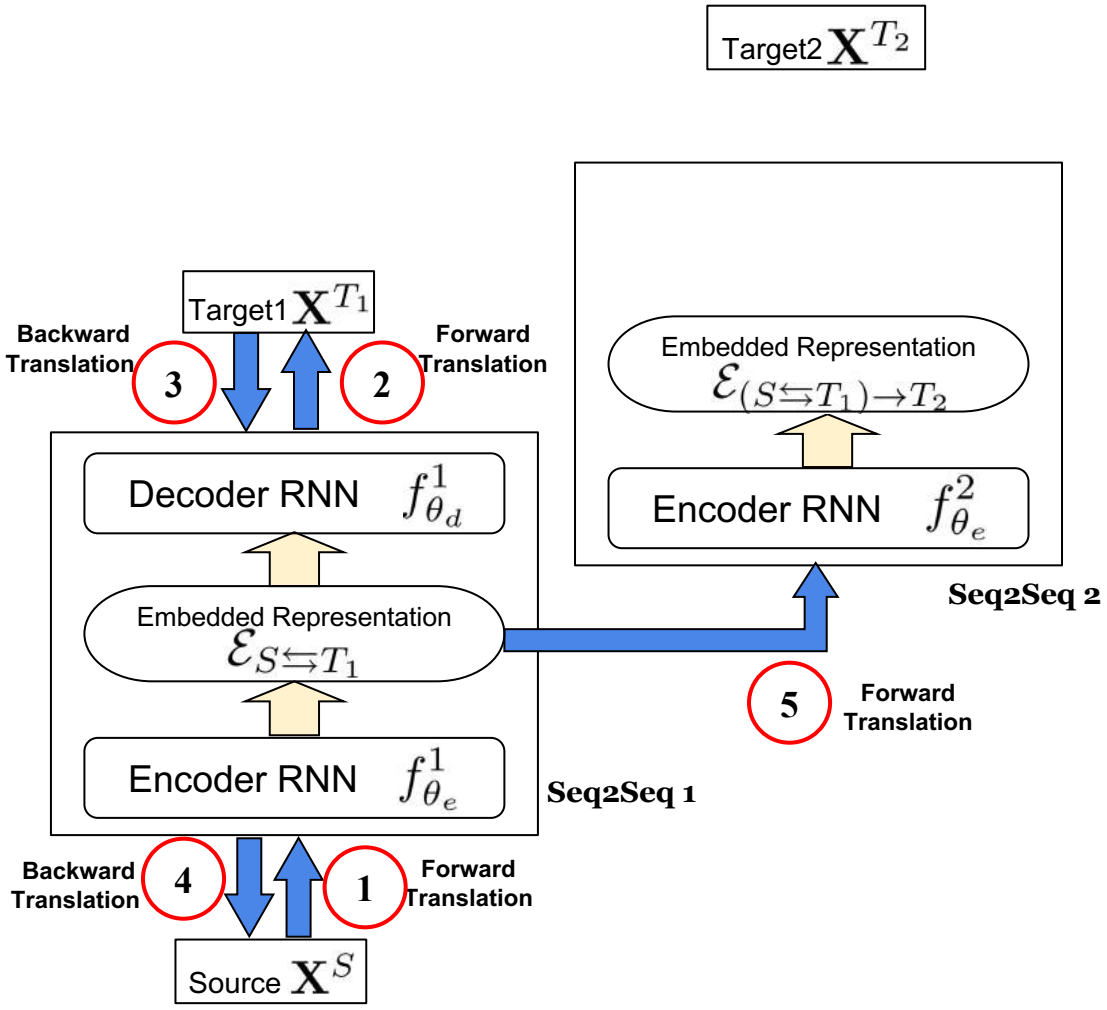
Hierarchical Multimodal Cyclic Translation Network



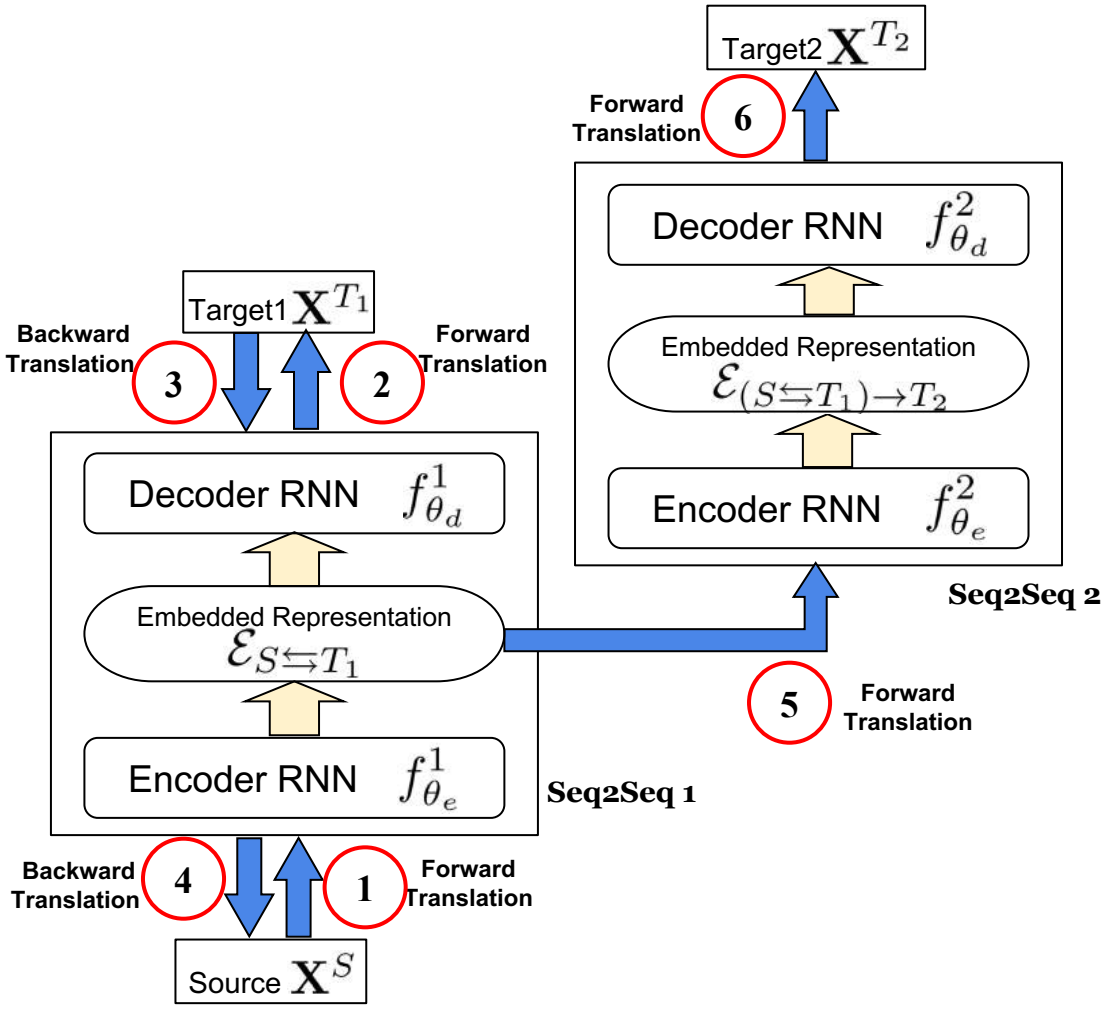
Hierarchical Multimodal Cyclic Translation Network



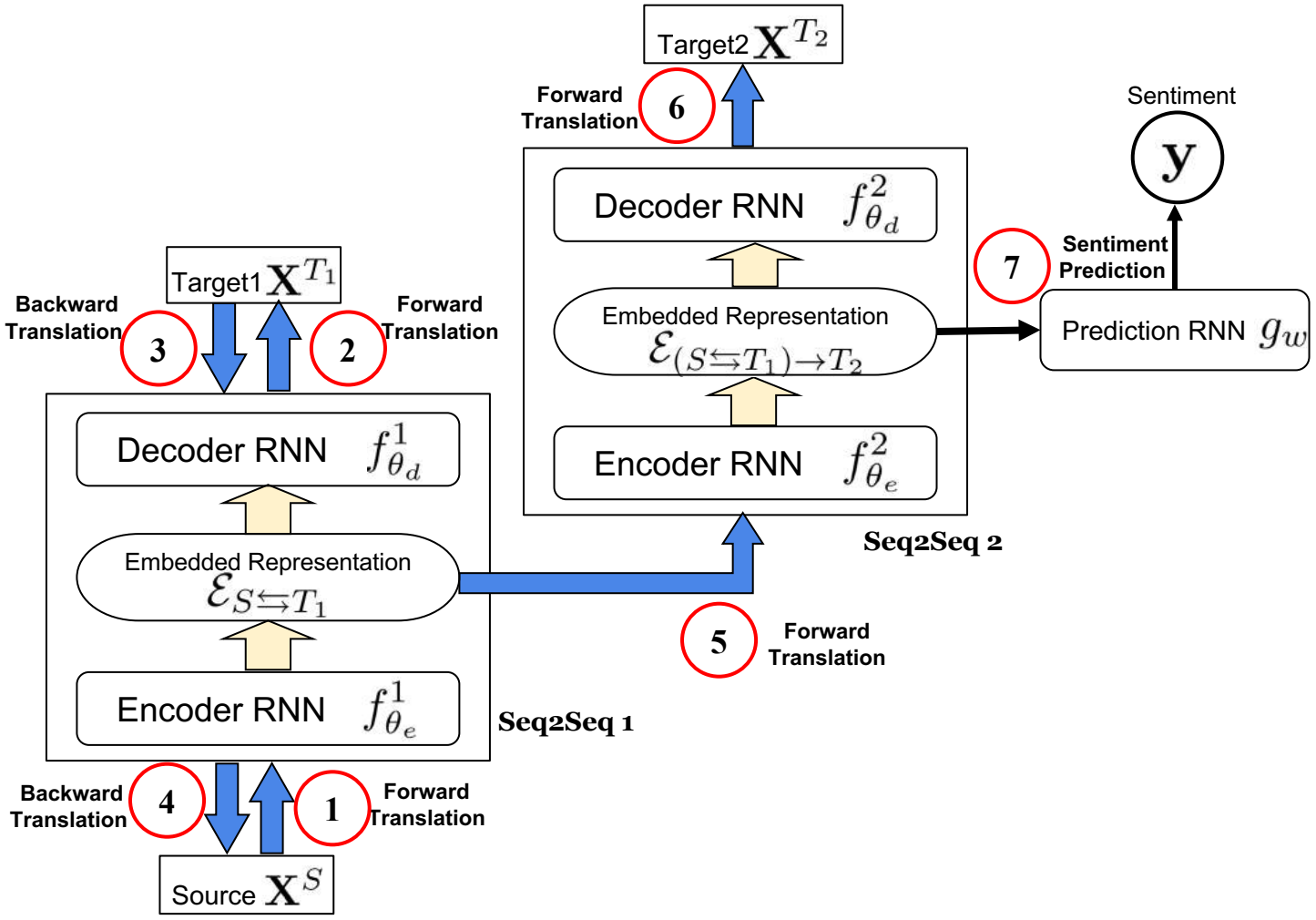
Hierarchical Multimodal Cyclic Translation Network



Hierarchical Multimodal Cyclic Translation Network



Hierarchical Multimodal Cyclic Translation Network



Baseline Models

1. Non-temporal models: SVM (Cortes and Vapnik, 1995), DF (Nojavanasghari et al., 2016)
2. Early fusion: EF-LSTM (Hochreiter and Schmidhuber, 1997), EF-RHN (Zilly et al., 2016)
3. Late fusion: LMF (Liu et al., 2018), TFN (Zadeh et al., 2017), BC-LSTM (Poria et al., 2017)
4. Multi-view learning: MV-LSTM (Rajagopalan et al., 2016)
5. Memory-based models: MARN, MFN (Zadeh et al., 2018)
6. Multi-stage model: RMFN (Liang et al., 2018)

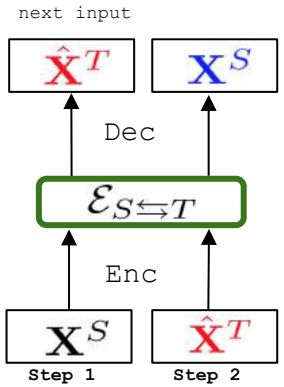
State-of-the-art Results: CMU-MOSI

Dataset		CMU-MOSI			
Model	Test Inputs	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
RF	$\{\ell, v, a\}$	56.4	56.3	-	-
SVM	$\{\ell, v, a\}$	71.6	72.3	1.100	0.559
THMM	$\{\ell, v, a\}$	50.7	45.4	-	-
EF-HCRF	$\{\ell, v, a\}$	65.3	65.4	-	-
MV-HCRF	$\{\ell, v, a\}$	65.6	65.7	-	-
DF	$\{\ell, v, a\}$	74.2	74.2	1.143	0.518
EF-LSTM	$\{\ell, v, a\}$	74.3	74.3	1.023	0.622
MV-LSTM	$\{\ell, v, a\}$	73.9	74.0	1.019	0.601
BC-LSTM	$\{\ell, v, a\}$	75.2	75.3	1.079	0.614
TFN	$\{\ell, v, a\}$	74.6	74.5	1.040	0.587
GME-LSTM(A)	$\{\ell, v, a\}$	76.5	73.4	0.955	-
MARN	$\{\ell, v, a\}$	77.1	77.0	0.968	0.625
MFN	$\{\ell, v, a\}$	77.4	77.3	0.965	0.632
LMF	$\{\ell, v, a\}$	76.4	75.7	0.912	0.668
RMFN	$\{\ell, v, a\}$	78.4	78.0	0.922	**0.681**
MCTN	$\{\ell\}$	**79.3**	**79.1**	**0.909**	0.676

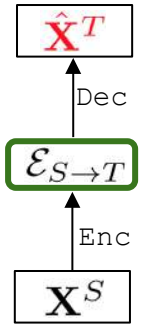
State-of-the-art Results: ICT-MMMO and YouTube

Dataset		ICT-MMMO		YouTube	
Model	Test Inputs	Acc(↑)	F1(↑)	Acc(↑)	F1(↑)
RF	$\{\ell, v, a\}$	70.0	69.8	33.3	32.3
SVM	$\{\ell, v, a\}$	68.8	68.7	42.4	37.9
THMM	$\{\ell, v, a\}$	53.8	53.0	42.4	27.9
EF-HCRF	$\{\ell, v, a\}$	73.8	73.1	45.8	45.0
MV-HCRF	$\{\ell, v, a\}$	68.8	67.1	44.1	44.0
DF	$\{\ell, v, a\}$	65.0	58.7	45.8	32.0
EF-LSTM	$\{\ell, v, a\}$	72.5	70.9	44.1	43.6
MV-LSTM	$\{\ell, v, a\}$	72.5	72.3	45.8	43.3
BC-LSTM	$\{\ell, v, a\}$	70.0	70.1	45.0	45.1
TFN	$\{\ell, v, a\}$	72.5	72.6	45.0	41.0
MARN	$\{\ell, v, a\}$	71.3	70.2	48.3	44.9
MFN	$\{\ell, v, a\}$	73.8	73.1	**51.7**	51.6
MCTN	$\{\ell\}$	**81.3**	**80.8**	**51.7**	**52.4**

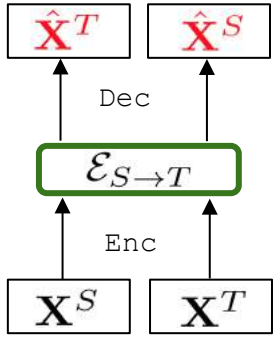
Bimodal Variations



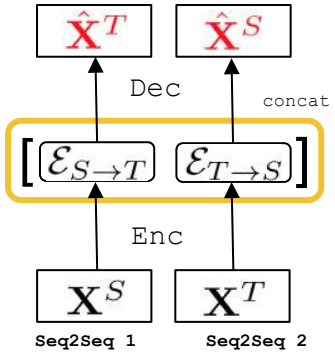
MCTN Bi



Simple Bi



No-Cycle Bi

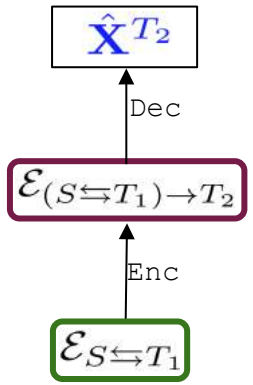


Double Bi

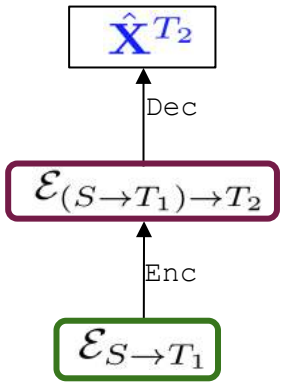
Bimodal Variations Results

Dataset		CMU-MOSI			
Model	Translation	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
MCTN Bi (Fig. 4a)	$V \leftrightarrows A$	53.1	53.2	1.420	0.034
	$T \leftrightarrows A$	76.4	76.4	0.977	**0.636**
	$T \leftrightarrows V$	**76.8**	**76.8**	1.034	0.592
Simple Bi (Fig. 4b)	$V \rightarrow A$	55.4	55.5	1.422	0.119
	$T \rightarrow A$	74.2	74.2	0.988	0.616
	$T \rightarrow V$	75.7	75.6	1.002	0.617
No cycle Bi (Fig. 4c)	$V \rightarrow A,\ A \rightarrow V$	55.4	55.5	1.422	0.119
	$T \rightarrow A,\ A \rightarrow T$	75.5	75.6	**0.971**	0.629
	$T \rightarrow V,\ V \rightarrow T$	75.2	75.3	0.972	0.627
Double Bi (Fig. 4d)	$[V \rightarrow A, A \rightarrow V]$	57.0	57.1	1.502	0.168
	$[T \rightarrow A, A \rightarrow T]$	72.3	72.3	1.035	0.578
	$[T \rightarrow V, V \rightarrow T]$	73.3	73.4	1.020	0.570

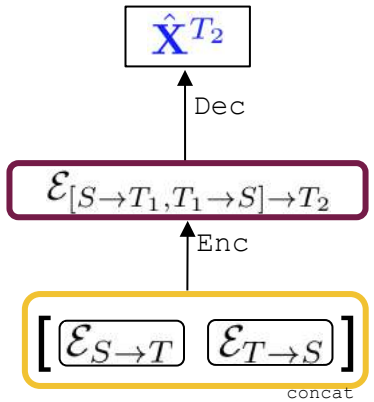
Trimodal Variations



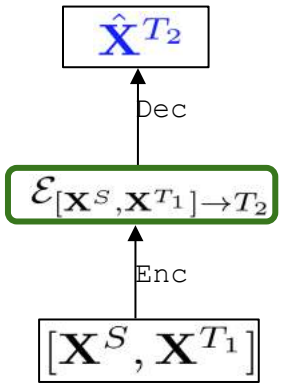
MCTN Tri



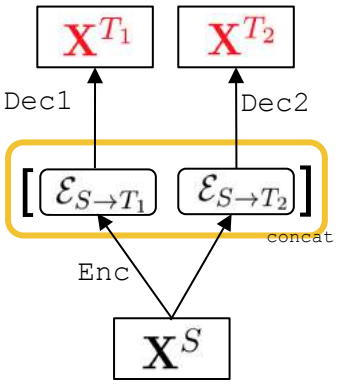
Simple Tri



Double Tri



Concat Tri



Paired Tri

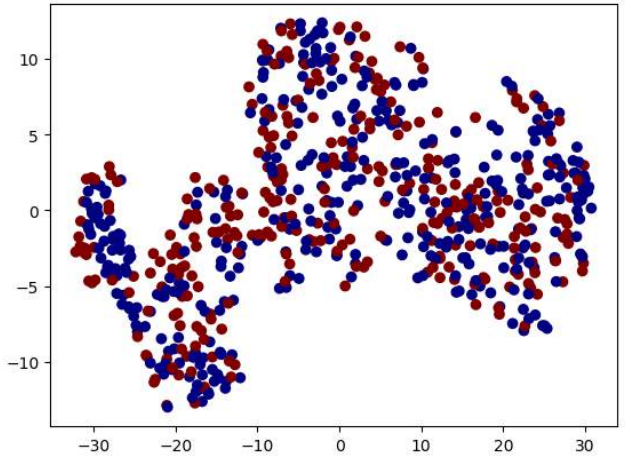
Trimodal Variations Results

Dataset		CMU-MOSI			
Model	Translation	Acc(↑)	F1(↑)	MAE(↓)	Corr(↑)
MCTN Tri (Fig. 4e)	$(V \leftrightarrows A) \rightarrow T$	56.4	56.3	1.455	0.151
	$(T \leftrightarrows A) \rightarrow V$	78.7	78.8	0.960	0.650
	$(T \leftrightarrows V) \rightarrow A$	**79.3**	**79.1**	**0.909**	**0.676**
Simple Tri (Fig. 4f)	$(V \rightarrow T) \rightarrow A$	54.1	52.9	1.408	0.040
	$(V \rightarrow A) \rightarrow T$	52.0	51.9	1.439	0.015
	$(A \rightarrow V) \rightarrow T$	56.6	56.7	1.593	0.067
	$(A \rightarrow T) \rightarrow V$	54.1	54.2	1.577	0.028
	$(T \rightarrow A) \rightarrow V$	74.3	74.4	1.001	0.609
	$(T \rightarrow V) \rightarrow A$	74.3	74.4	0.997	0.596
Double Tri (Fig. 4g)	$[T \rightarrow V, V \rightarrow T] \rightarrow A$	73.3	73.1	1.058	0.578
Concat Tri (Fig. 4h)	$[V, A] \rightarrow T$	55.0	54.6	1.535	0.176
	$[A, T] \rightarrow V$	73.3	73.4	1.060	0.561
	$[T, V] \rightarrow A$	72.3	72.3	1.068	0.576
	$A \rightarrow [T, V]$	55.5	55.6	1.617	0.056
	$T \rightarrow [A, V]$	75.7	75.7	0.958	0.634
	$[T, A] \rightarrow [T, V]$	73.2	73.2	1.008	0.591
	$[T, V] \rightarrow [T, A]$	74.1	74.1	0.999	0.607
Paired Tri (Fig. 4i)	$[T \rightarrow A, T \rightarrow V]$	73.8	73.8	1.022	0.611

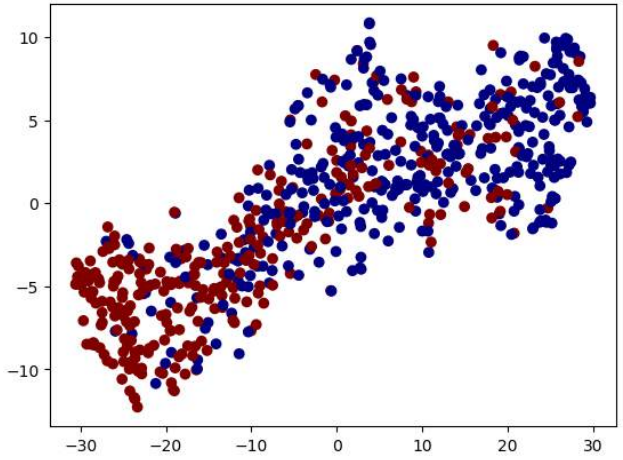
Adding More Modalities

Dataset		CMU-MOSI			
Model	Translation	Acc	F1	MAE	Corr
MCTN Bi (Fig. 4a)	$V \leftrightarrows A$	53.1	53.2	1.420	0.034
	$T \leftrightarrows A$	76.4	76.4	0.977	0.636
	$T \leftrightarrows V$	76.8	76.8	1.034	0.592
MCTN Tri (Fig. 4e)	$(V \leftrightarrows A) \rightarrow T$	56.4	56.3	1.455	0.151
	$(T \leftrightarrows A) \rightarrow V$	78.7	78.8	0.960	0.650
	$(T \leftrightarrows V) \rightarrow A$	**79.3**	**79.1**	**0.909**	**0.676**

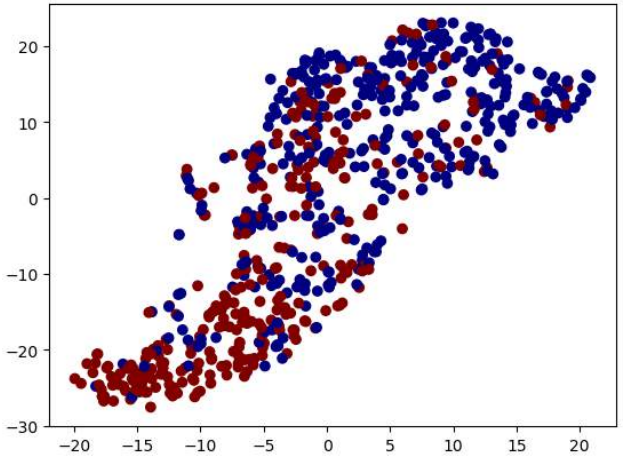
Adding More Modalities



Bimodal MCTN
without
cyclic translation



Bimodal MCTN
with
cyclic translation



Trimodal MCTN
with
cyclic translation

Thank you for your attention!

Email: htpham@cs.cmu.edu
Twitter: @hai_t_pham

Email: pliang@cs.cmu.edu
Twitter: @pliang279