Seq2Seq2Sentiment: Multimodal Sequence to Sequence Models for Sentiment Analysis

Hai Pham*[1], Thomas Manzini*[1], Paul Pu Liang[2], Barnabás Póczos[2]

[1] *Language Technologies Institute,* [2] *Machine Learning Department*
School of Computer Science, Carnegie Mellon University

*Authors Contributed Equally

Introduction

- Research Problem
  - Does representing multiple modalities jointly improve sentiment prediction for the CMU-MOSI dataset?

Introduction

- Research Problem
    - Does representing multiple modalities jointly improve sentiment prediction for the CMU-MOSI dataset?

- Dataset
    - Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos (MOSI)
        - Modalities: Video, Audio, Text Transcripts
        - 89 speakers, 93 videos split into 2199 labeled opinion segments
        - Labels for Sentiment: {-1, 1} or {-3, -2, -1, 0, 1, 2, 3}

Baseline & Metrics

- LSTM Multi Modal Baseline
  - Concatenates all modalities together and predicts sentiment from these modalities.
  - 75% Accuracy (Chen et. al.)

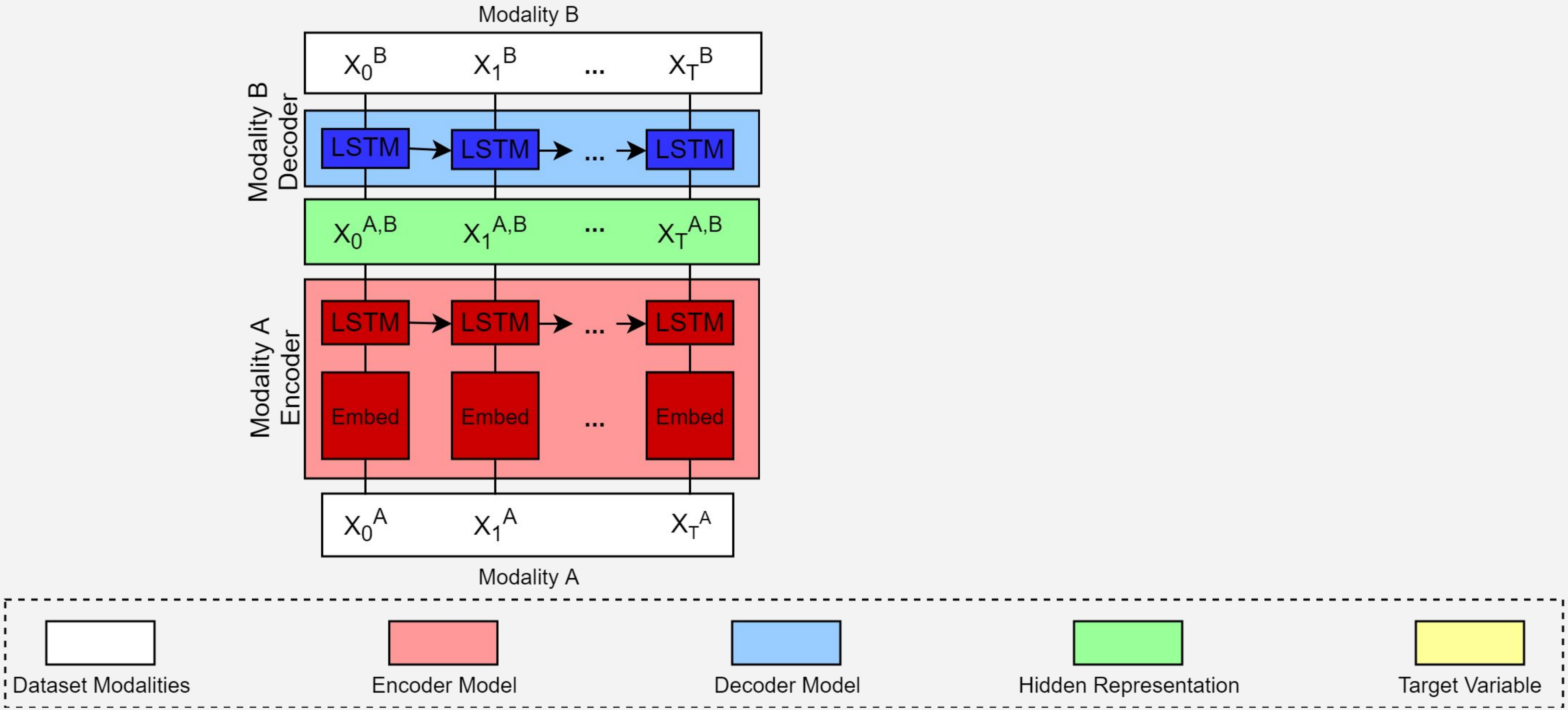
- Metrics
  - For both {-1, 1} or {-3, -2, -1, 0, 1, 2, 3} cases
    - Precision/Recall (Test)
    - F1 Score (Test)

Related Work

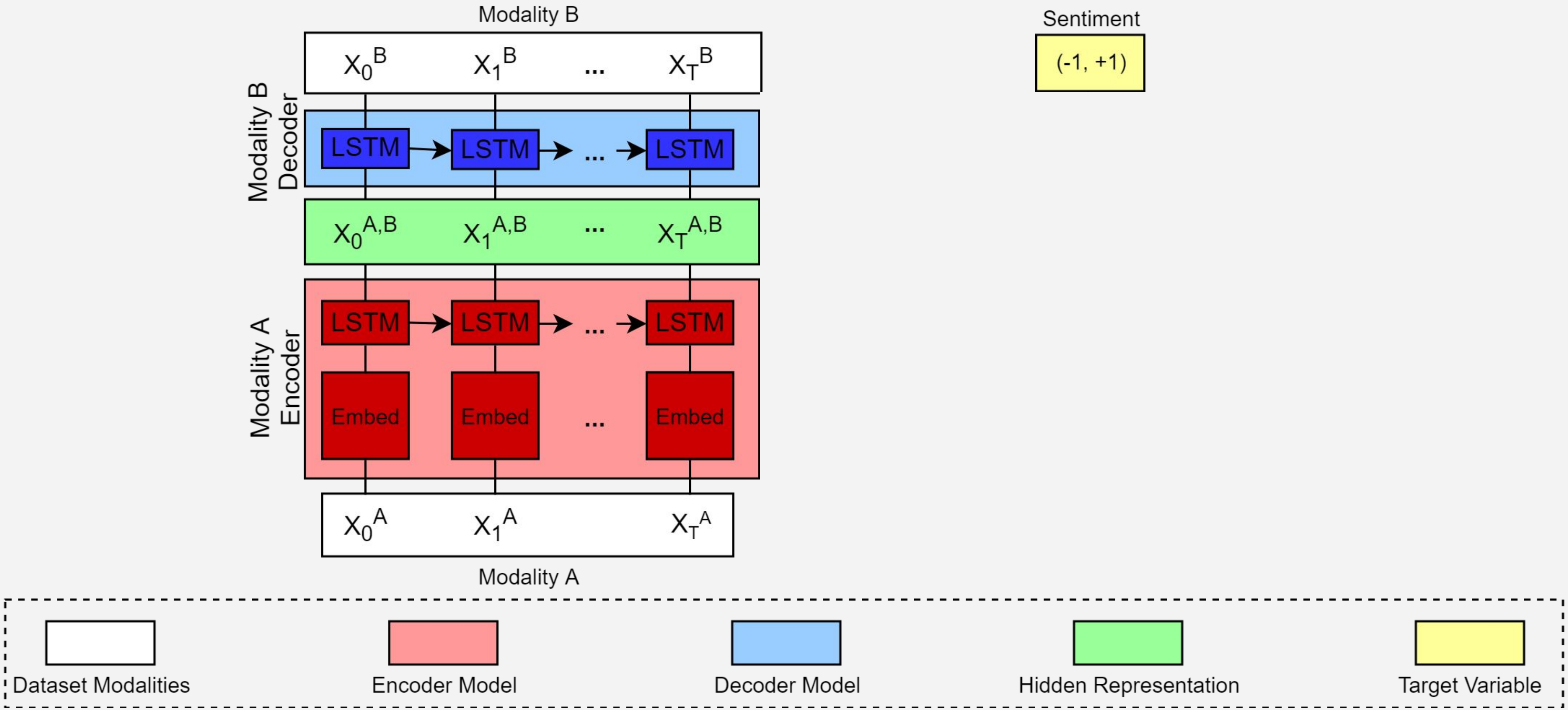
- Word-level Temporal Methods
  - Gu et al. ACL 2018, Chen et al. ICMI 2017
- Context-Dependent Methods
  - Poria et al. ACL 2017
- Memory-based Methods
  - Zadeh et al. AAAI 2018
- Tensor-based Methods
  - Liu et al. ACL 2018, Zadeh et al. EMNLP 2017
- Conditional Approaches
  - Mirza et al. 2014, Kingma et al. 2014, & Pandey et al. 2017
- Attention-based Methods
  - Bahdanau et al. 2014, Luong et al. 2015

See our paper for an exhaustive review of related work

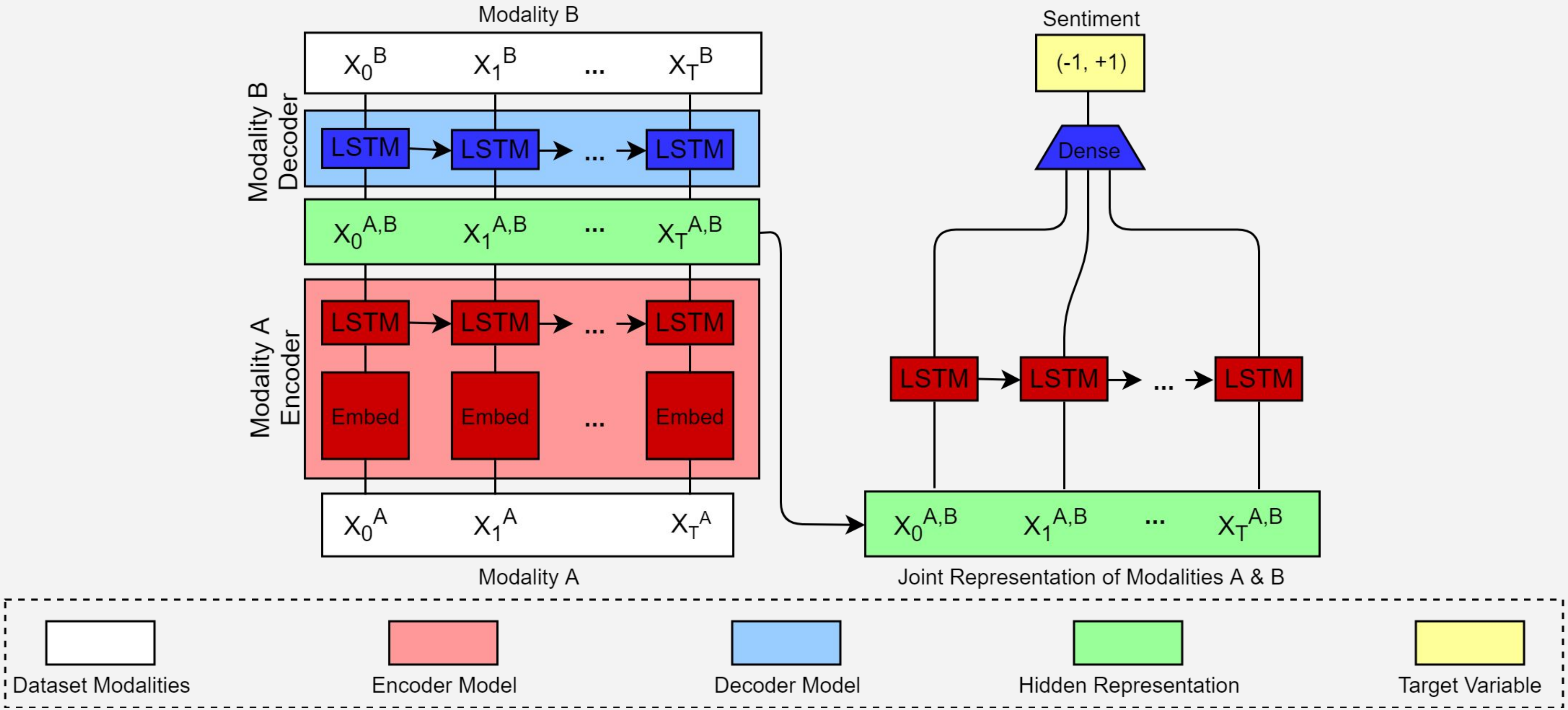
Seq2Seq Modality Translation



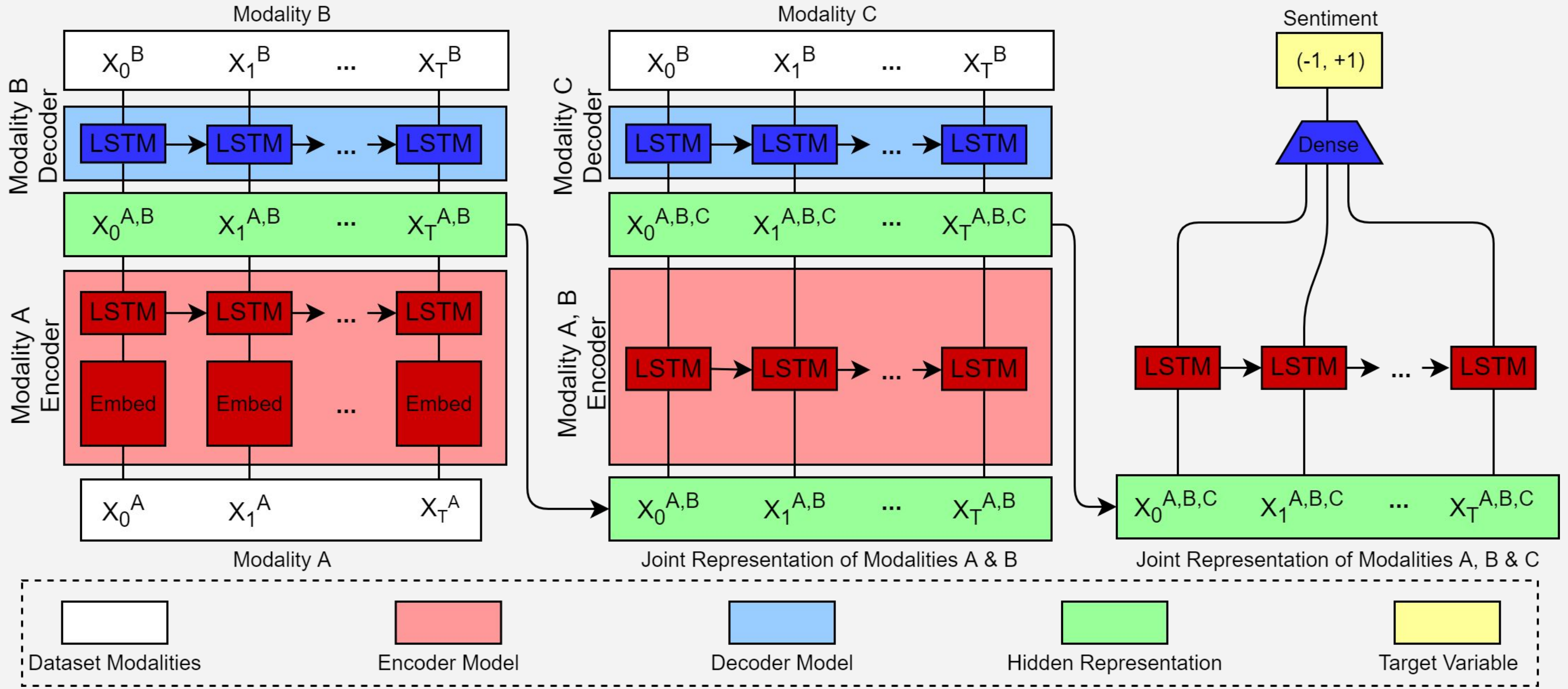
Seq2Seq Modality Translation



Seq2Seq Modality Translation



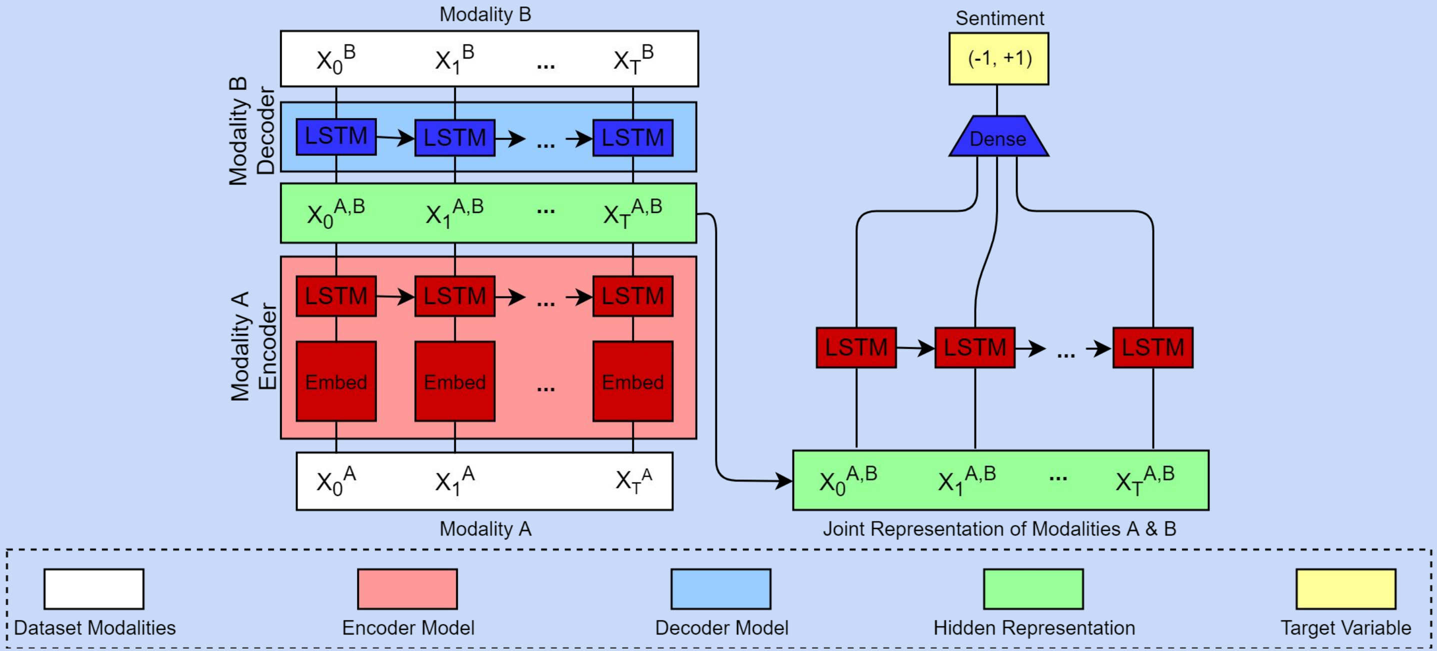
Hierarchical Seq2Seq Modality Translation



Experiments

Denoted as
A → B

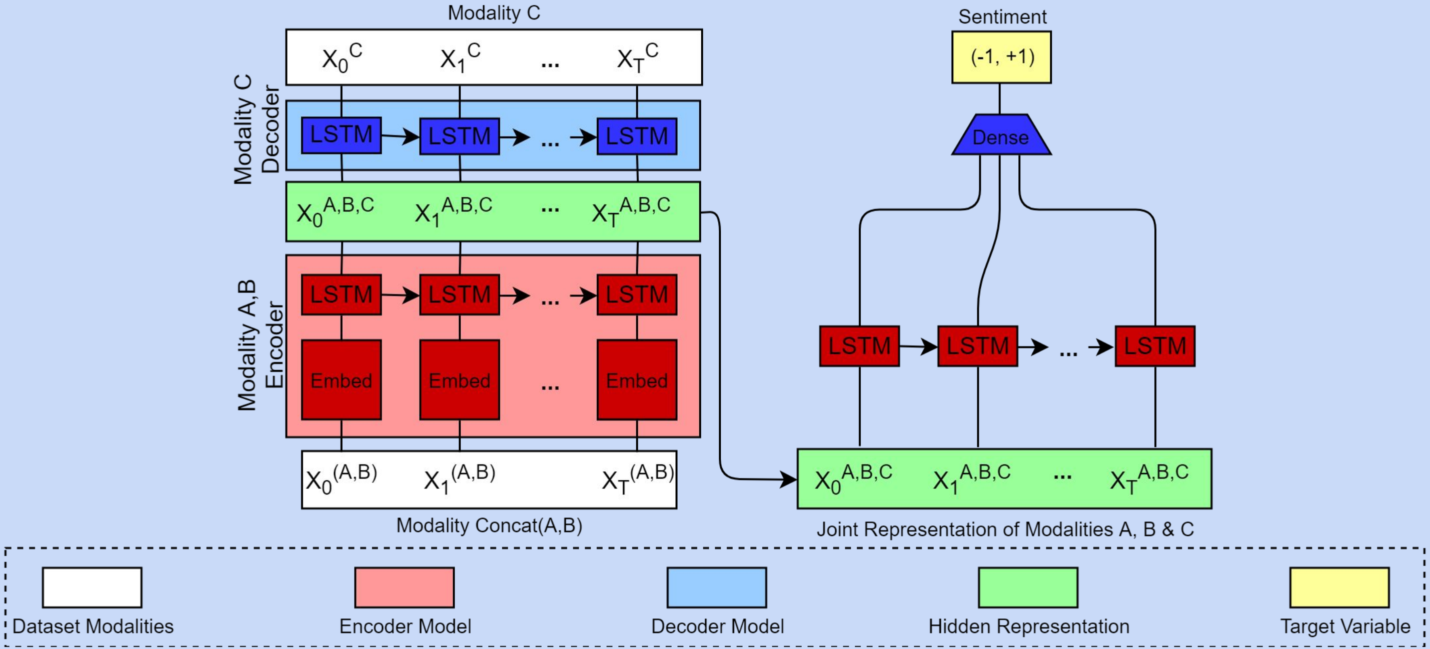
- We explored several different network architectures for generating representations



Experiments

Denoted as
Concat(A,B) → C

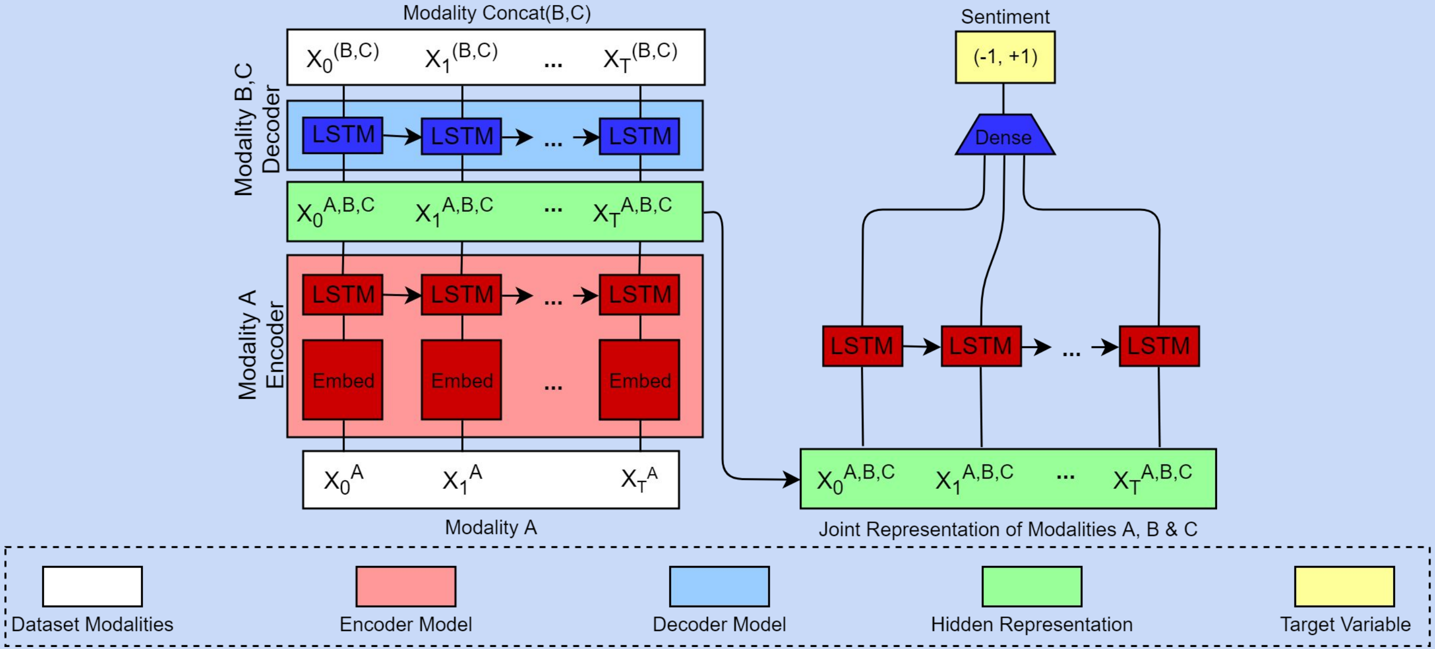
- We explored several different network architectures for generating representations



Experiments

Denoted as
A → Concat(B,C)

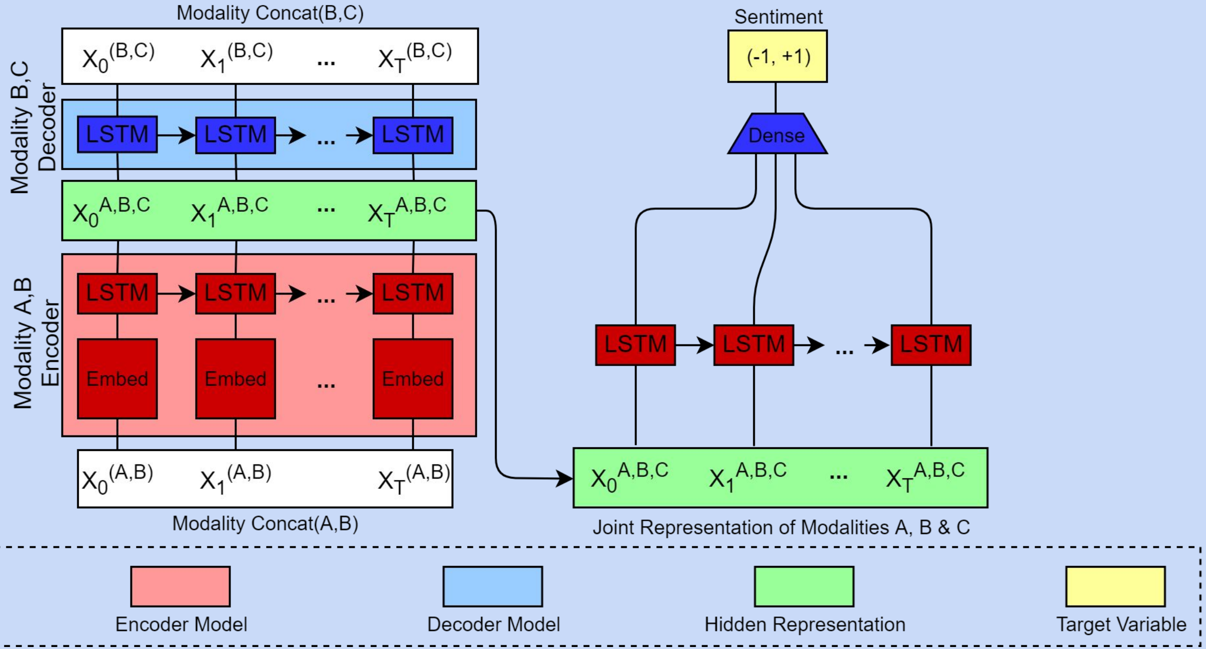
- We explored several different network architectures for generating representations



Experiments

Denoted as
Concat(A,B) → Concat(B,C)

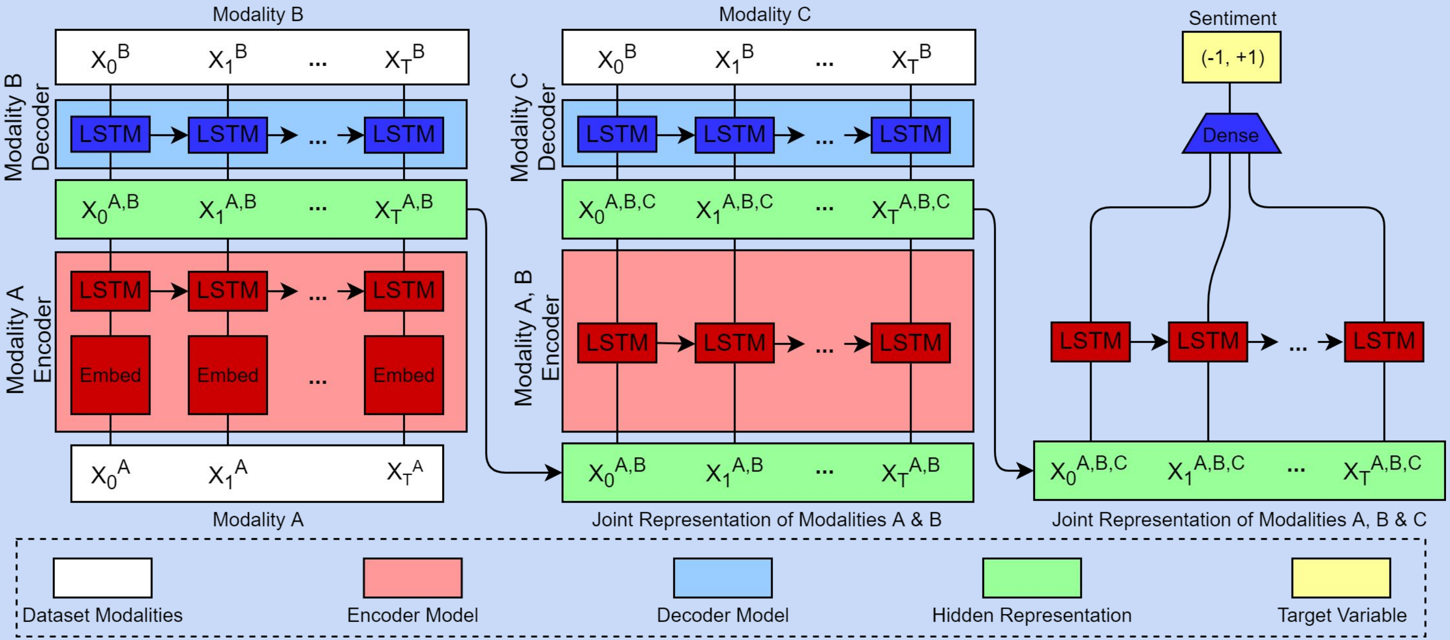
- We explored several different network architectures for generating representations



Denoted as
Embed(A, B) → C

Experiments

- We explored several different network architectures for generating representations



Experiments

- Unimodal Baseline

Method	Feature	BINARY (−1, +1) Prec	Recall	F1	7-CLASS (−3, ..., +3) Prec	Recall	F1
UniModal-Baseline	Text (T)	**0.77**	**0.76**	**0.76**	**0.32**	**0.35**	**0.33**
	Audio (A)	0.56	0.56	0.56	0.12	0.19	0.14
	Video (V)	0.57	0.47	0.48	0.12	0.19	0.12

T = Text Modality, A = Audio Modality, V = Visual (facial) modality

Results (Bi-Modal)

- Bimodal Baseline & Experimental Results

Method	Feature	BINARY (−1, +1) Prec	Recall	F1	7-CLASS (−3, ..., +3) Prec	Recall	F1
BiModal-Baseline	concat(T + V)	**0.78**	**0.67**	0.55	0.01	0.16	0.05
	concat(T + A)	0.44	0.66	0.53	0.02	0.15	0.04
	concat(A + V)	0.55	0.47	0.48	0.13	0.16	0.11
BiModal-Seq2Seq	T → V	0.67	**0.67**	**0.67**	0.26	0.22	**0.22**
	T → A	0.66	0.64	0.65	**0.28**	0.24	0.18
	A → T	0.55	0.60	0.56	0.17	**0.34**	0.11
	A → V	0.55	0.55	0.54	0.16	0.18	0.16
	V → T	0.58	0.58	0.58	0.05	0.16	0.08
	V → A	0.58	0.62	0.58	0.12	0.17	0.01

T = Text Modality, A = Audio Modality, V = Visual (facial) modality

Results (Bi-Modal)

- Bimodal Baseline & Experimental Results

Method	Feature	BINARY (−1, +1) Prec	Recall	F1	7-CLASS (−3, ..., +3) Prec	Recall	F1
BiModal-Baseline	concat(T + V)	**0.78**	**0.67**	0.55	0.01	0.16	0.05
	concat(T + A)	0.44	0.66	0.53	0.02	0.15	0.04
	concat(A + V)	0.55	0.47	0.48	0.13	0.16	0.11
BiModal-Seq2Seq	T → V	0.67	**0.67**	**0.67**	0.26	0.22	**0.22**
	T → A	0.66	0.64	0.65	**0.28**	0.24	0.18
	A → T	0.55	0.60	0.56	0.17	**0.34**	0.11
	A → V	0.55	0.55	0.54	0.16	0.18	0.16
	V → T	0.58	0.58	0.58	0.05	0.16	0.08
	V → A	0.58	0.62	0.58	0.12	0.17	0.01

T = Text Modality, A = Audio Modality, V = Visual (facial) modality

10 Point
Boost

Results (Bi-Modal)

- Bimodal Baseline & Experimental Results

Method	Feature	BINARY (−1, +1) Prec	Recall	F1	7-CLASS (−3, ..., +3) Prec	Recall	F1
BiModal-Baseline	concat(T + V)	**0.78**	**0.67**	0.55	0.01	0.16	0.05
	concat(T + A)	0.44	0.66	0.53	0.02	0.15	0.04
	concat(A + V)	0.55	0.47	0.48	0.13	0.16	0.11
BiModal-Seq2Seq	T → V	0.67	**0.67**	**0.67**	0.26	0.22	**0.22**
	T → A	0.66	0.64	0.65	**0.28**	0.24	0.18
	A → T	0.55	0.60	0.56	0.17	**0.34**	0.11
	A → V	0.55	0.55	0.54	0.16	0.18	0.16
	V → T	0.58	0.58	0.58	0.05	0.16	0.08
	V → A	0.58	0.62	0.58	0.12	0.17	0.01

T = Text Modality, A = Audio Modality, V = Visual (facial) modality

12 Point Boost

Results (Tri-Modal)

- Trimodal Baseline & Experimental Results

Method	Feature	BINARY (−1, +1) Prec	Recall	F1	7-CLASS (−3, ..., +3) Prec	Recall	F1
TriModal-Baseline	`concat`(T + V + A)	**0.75**	**0.75**	**0.75**	0.24	**0.27**	**0.24**
TriModal-Seq2Seq	`embed`(T, V) → A	0.56	0.60	0.57	0.10	0.16	0.09
	`embed`(T, A) → V	0.60	0.55	0.56	0.26	0.15	0.07
	`embed`(A, V) → T	0.66	0.53	0.44	0.16	0.04	0.09
	`embed`(A, T) → V	0.59	0.51	0.52	0.13	0.15	0.09
	`embed`(V, T) → A	0.59	0.60	0.59	0.11	0.17	0.10
	`embed`(V, A) → T	0.57	0.61	0.58	0.11	0.17	0.09
	`concat`(T, V) → A	0.67	0.66	0.65	0.22	0.17	0.18
	`concat`(A, T) → V	0.54	0.55	0.63	0.19	0.15	0.21
	`concat`(V, A) → T	0.59	0.59	0.58	0.16	0.12	0.12
	T → `concat`(A, V)	0.70	0.65	0.66	0.23	0.22	0.18
	A → `concat`(T, V)	0.55	0.53	0.54	0.18	0.20	0.18
	`concat`(T, A) → `concat`(T, V)	0.62	0.60	0.61	0.23	0.24	0.22
	`concat`(T, V) → `concat`(T, A)	0.68	0.70	0.67	**0.31**	0.24	0.19

T = Text Modality, A = Audio Modality, V = Visual (facial) modality

Results - Takeaways

- We clearly outperform the baselines in the **bi-modal** domain
  - In the 7-class paradigm we often outperform by a large margin
  - For datasets without transcripts this approach may result in significant gains
- Slightly outperform baseline in tri-modal multiclass setting
- Significantly longer training times than the baseline alone

Future Work

- Our method is unsupervised, we will pre-train seq2seq model with external dataset
- Use variational seq2seq to refine the training
- Further explore end-to-end training
- Explore additional methods for encoding sequences with 2 modalities
  - Multi-view LSTM

Acknowledgements

- Amir Zadeh
- Volkan Cirik
- Louis-Philippe Morency
- Somya Wadhwa
- Minghai Chen
- Hieu Pham
- Workshop Reviewers
- *AWS*

Thank you!

Appendix

Problem Formulation

- Input: $X = (X_1, X_2, \ldots, X_n)$ where $X_i = ({X_i}^{text}, {X_i}^{audio}, {X_i}^{video})$
- Output : $Y = (Y_1, Y_2, ..., Y_n)$, $Y_i \in \mathbb{R}$
- Align based on word-level

$$ {X_i}^{text} = ({w_i}^{(1)}, {w_i}^{(2)}, ..., {w_i}^{(T_i)}) $$

$$ {X_i}^{audio} = ({a_i}^{(1)}, {a_i}^{(2)}, ..., {a_i}^{(T_i)}) $$

$$ {X_i}^{video} = ({v_i}^{(1)}, {v_i}^{(2)}, ..., {v_i}^{(T_i)}) $$

- Goal: Learn the embedding representation

$$ \widetilde{X_i} = f(X_i) = f(({X_i}^{text}, {X_i}^{audio}, {X_i}^{video})) $$

$$ \widetilde{X_i} = f(X_i) = Seq2Seq\_Encoder(X_i) $$

Problem Formulation (cont'd)

- Transformed input: $\widetilde{X} = (\widetilde{x}^1, \widetilde{x}^2, ..., \widetilde{x}^T)$ with output $Y = (y^1, y^2, ..., y^T)$
- Using RNN with K hidden layers:

$$h = (h^1, h^2, \ldots, h^K)$$
$$h^k = (h_1^k, h_2^k, ..., h_D^k),\ k \in [1, K]$$

- First layer: $h^1{}_t = H(W_{xh^1}\widetilde{x}_t + W_{h^1h^1}h^1{}_{t-1} + b_{h^1})$
- Layer k : $h^k{}_t = H(W_{h^{t-1}h^t}h_t^{k-1} + W_{h^kh^k}h^k{}_{t-1} + b_{h^k})$
- Using soft attention at last hidden layer K:

$$\alpha = softmax\left(\begin{bmatrix} W_\alpha h_1^K \\ W_\alpha h_2^K \\ \ldots \\ W_\alpha h_T^K \end{bmatrix}\right)$$

Problem Formulation (cont'd)

- Output of last hidden layer: $A = [h_1^K, h_2^K, ..., h_T^K]\alpha = H^K\alpha$
- Final output: $\widetilde{y}_t = W_{Ay}A + b_y$
- Mean Absolute Error Loss: $\mathbb{L}_{MAE}(\widetilde{Y}, Y) = \mathbb{E}[|\widetilde{Y} - Y|]$
- Model is trained with SGD

Seq2Seq Modality Translation

Algorithm 1 Seq2Seq Modality Translation

X, Y, S are 2 modalities and sentiment sequences

Phase 1: Train Seq2Seq

$\mathcal{E}_{XY} \leftarrow Seq2Seq\_RNN\_Encode(X)$

$\widetilde{Y} \leftarrow Seq2Seq\_RNN\_Decode(\mathcal{E}_{XY})$

$loss = cross\_entropy(\widetilde{Y}, Y)$

Backprop to update params

Phase 2: Sentiment Regression

$\mathcal{E}_{XY} \leftarrow Seq2Seq\_RNN\_Encode(X)$ ▷ trained encoder in Seq2Seq model

$R = RNN(\mathcal{E}_{XY})$

$score \leftarrow Regression(R)$

$loss \leftarrow MAE(score, S)$

Backprop to update params

Hierarchical Seq2Seq Modality Translation

Algorithm 2 Hierarchical Seq2Seq Modality Translation: X, Y, Z, S are 3 modalities and sentiment sequences

```
Phase 1: Train Seq2Seq for 2 modalities
    E_XY ← Seq2Seq_RNN_Encode(X)
    Ỹ ← Seq2Seq_RNN_Decode(E_XY)
    loss = cross_entropy(Ỹ, Y)
    Backpropagate to update parameters

Phase 2: Train Seq2Seq for 3 modalities
    E_XYZ ← Seq2Seq_RNN_Encode(E_XY)
    Z̃ ← Seq2Seq_RNN_Decode(E_XYZ)
    loss = cross_entropy(Z̃, Z)
    Backpropagate to update parameters

Phase 3: Sentiment Regression
    E_XYZ ← Seq2Seq_RNN_Encode(E_XY)
    R = RNN(E_XYZ)
    score ← Regression(R)
    loss ← MAE(score, S)
    Backpropagate to update parameters
```