Multimodal Deep Learning

Purvanshi Mehta
purvanshi.mehta11@gmail.com

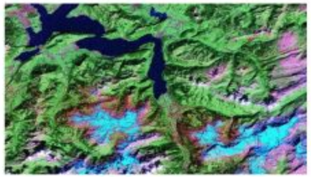
Table of contents

- Introduction
- Challenges
  - Representation
  - Translation
  - Alignment
  - Fusion
  - Co-Learning
- Interpretability in Multimodal Deep Learning

Aim of the presentation

- Identify challenges particular to Multimodal Learning
- Popular research topics in the field
- Brief of the problem I have been working on - Interpretability in Multimodal Deep Learning

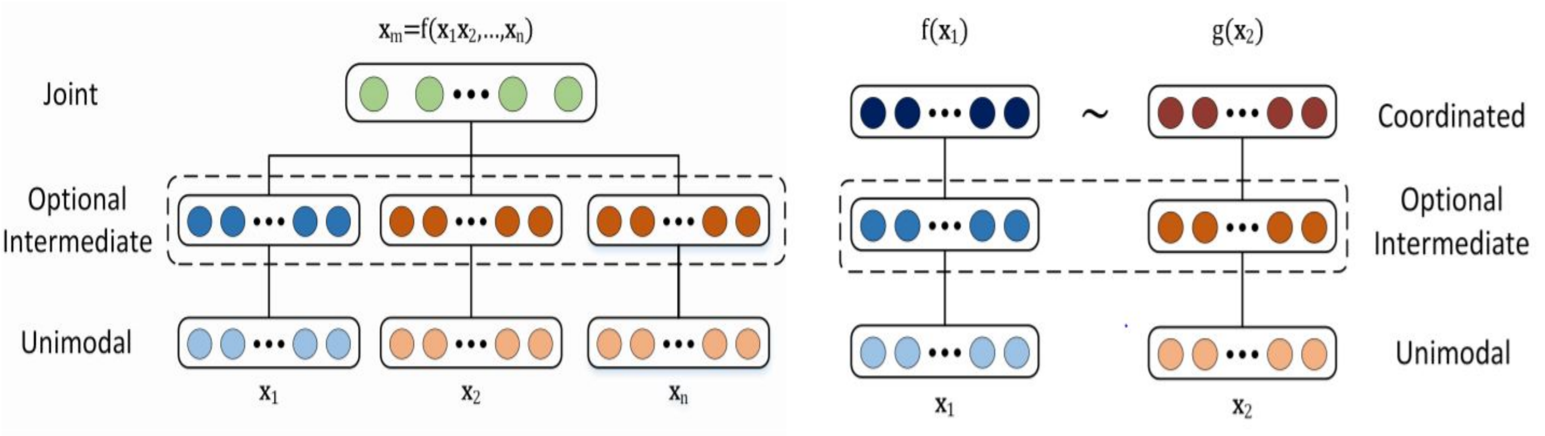
Multimodal Learning - Heterogeneous Information Sources



Challenge - 1) Representation

- How to combine the information from multiple sources?
- How to deal with different levels of noise?
- How to deal with missing data?

1) Representation - Ways of learning

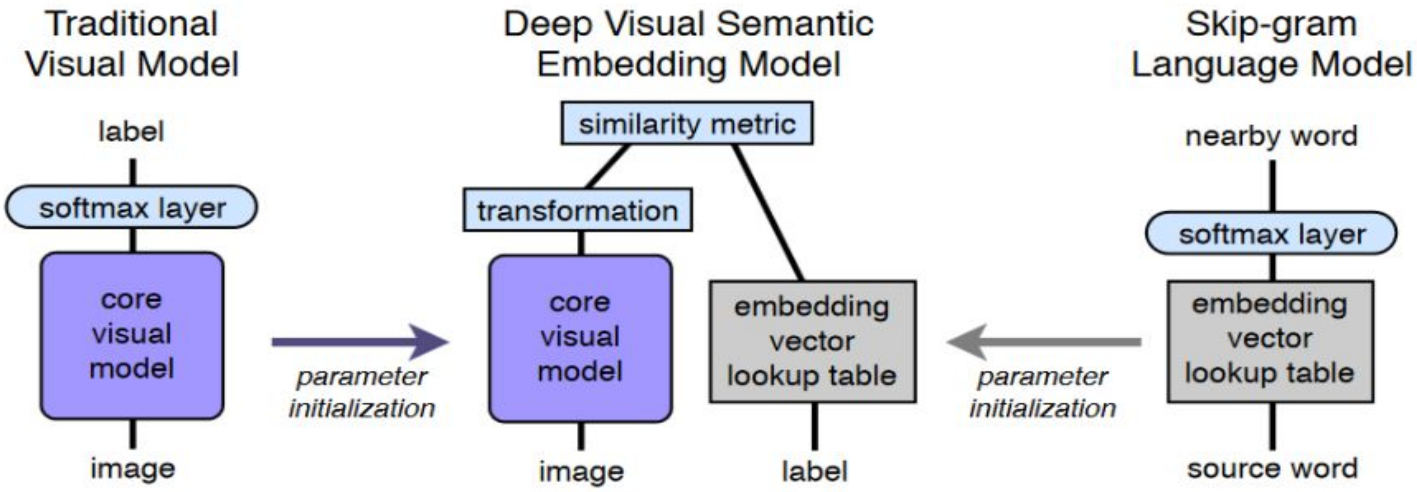


(a) Joint representation

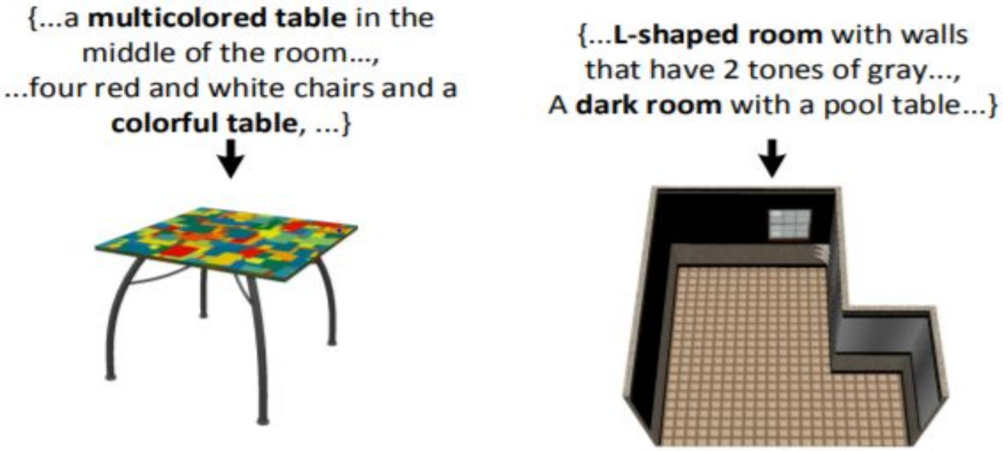
(b) Coordinated representations

Coordinated Representation

DeViSE — a deep visual-semantic embedding - Similarity model



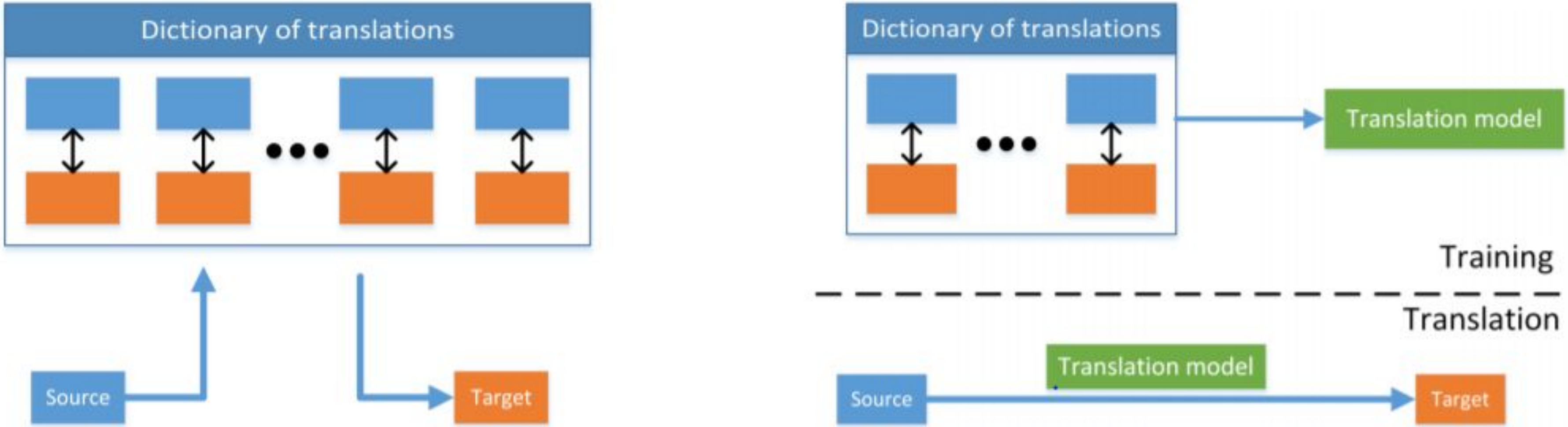
Challenge - 2) Translation



{...a **multicolored table** in the middle of the room..., ...four red and white chairs and a **colorful table**, ...}

{...**L-shaped room** with walls that have 2 tones of gray..., A **dark room** with a pool table...}

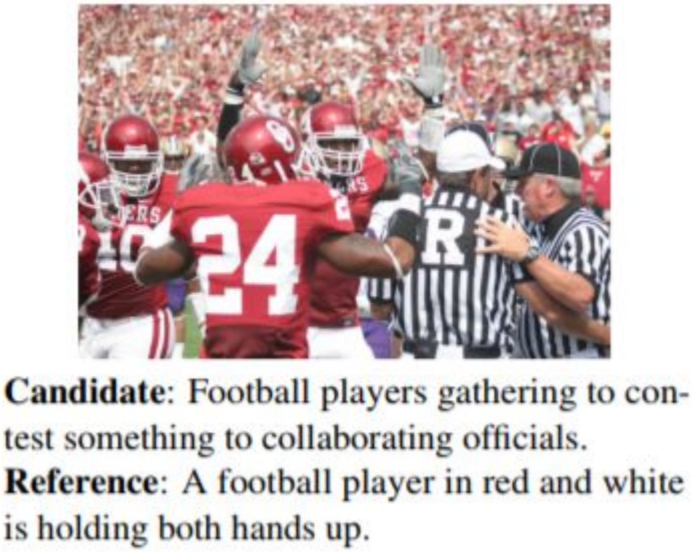
Translation



(a) Example-based

(b) Generative

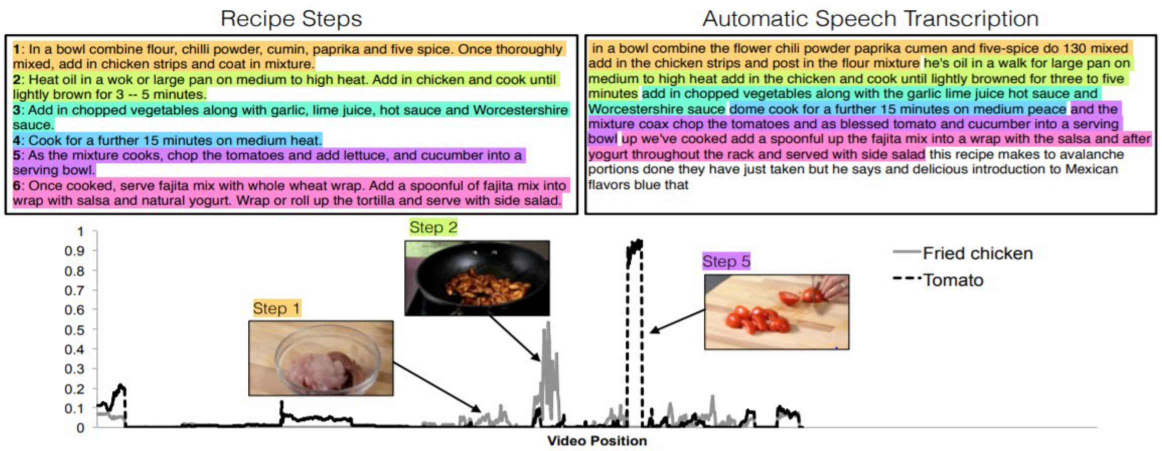
How to evaluate translations



Candidate: Football players gathering to contest something to collaborating officials.
Reference: A football player in red and white is holding both hands up.

Challenge - 3) Alignment

1) Identify the direct relations between (sub)elements from two or more different modalities.



Recipe Steps

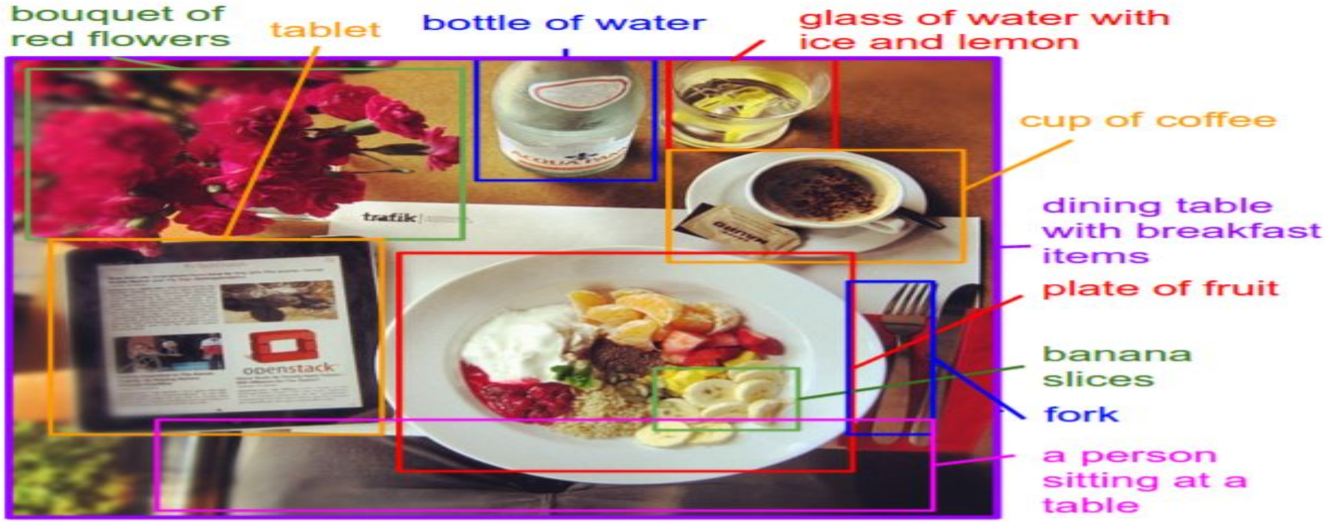
1: In a bowl combine flour, chilli powder, cumin, paprika and five spice. Once thoroughly mixed, add in chicken strips and coat in mixture.
2: Heat oil in a wok or large pan on medium to high heat. Add in chicken and cook until lightly brown for 3 -- 5 minutes.
3: Add in chopped vegetables along with garlic, lime juice, hot sauce and Worcestershire sauce.
4: Cook for a further 15 minutes on medium heat.
5: As the mixture cooks, chop the tomatoes and add lettuce, and cucumber into a serving bowl.
6: Once cooked, serve fajita mix with whole wheat wrap. Add a spoonful of fajita mix into wrap with salsa and natural yogurt. Wrap or roll up the tortilla and serve with side salad.

Automatic Speech Transcription

in a bowl combine the flower chili powder paprika cumen and five-spice do 130 mixed add in the chicken strips and post in the flour mixture he's oil in a walk for large pan on medium to high heat add in the chicken and cook until lightly browned for three to five minutes add in chopped vegetables along with the garlic lime juice hot sauce and Worcestershire sauce dome cook for a further 15 minutes on medium peace and the mixture coax chop the tomatoes and as blessed tomato and cucumber into a serving bowl up we've cooked add a spoonful up the fajita mix into a wrap with the salsa and after yogurt throughout the rack and served with side salad this recipe makes to avalanche portions done they have just taken but he says and delicious introduction to Mexican flavors blue that

Alignment

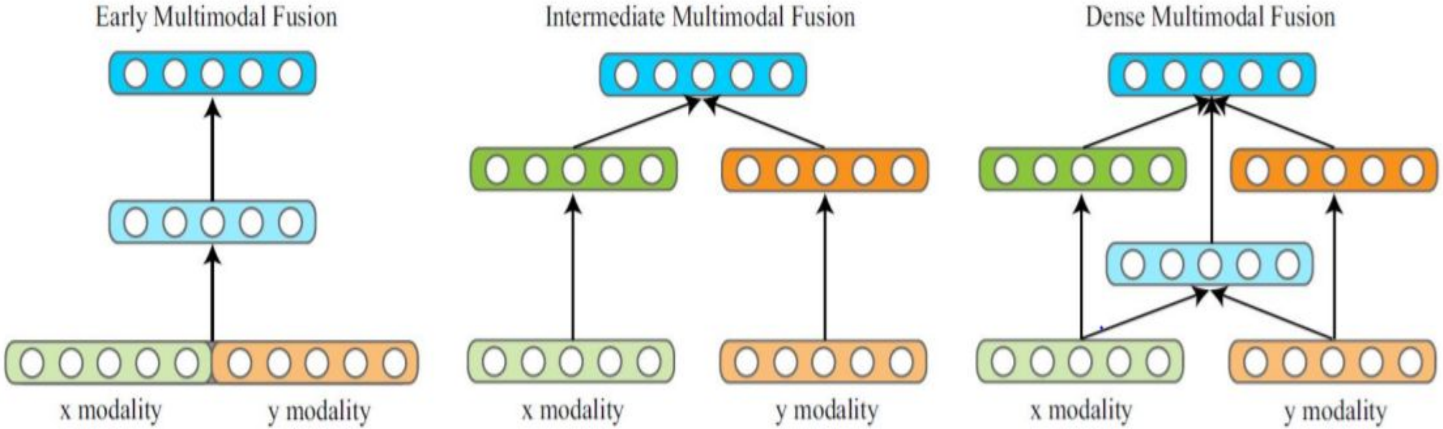
Given an image and a caption we want to find the areas of the image corresponding to the caption's words or phrases


bouquet of red flowers
tablet
bottle of water
glass of water with ice and lemon
cup of coffee
dining table with breakfast items
plate of fruit
banana slices
fork
a person sitting at a table

Alignement problems

- Few datasets with explicitly annotated alignments
- It is difficult to design similarity metrics between modalities
- Existence of multiple possible alignments and not all elements in one modality have correspondences in another

Challenge - 4) Fusion



Fusion

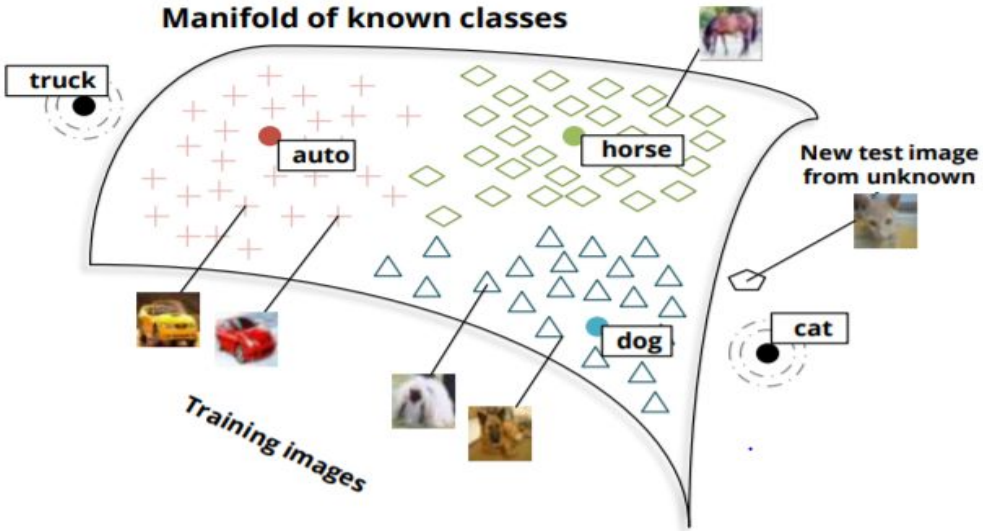
- Signals not temporarily aligned
- Lack of interpretability of where is the prediction coming from
  - (This is what I am working on)

Challenge - 5) Co-Learning

- Aiding the modeling of a (resource poor) modality by exploiting knowledge from another (resource rich) modality.
- When one modality has lack of annotated data, noisy inputs and unreliable labels.

Colearning - Zero Shot learning

Using text embeddings to classify unseen classes of images



Interpretability in Multimodal Deep Learning

Problem statement -

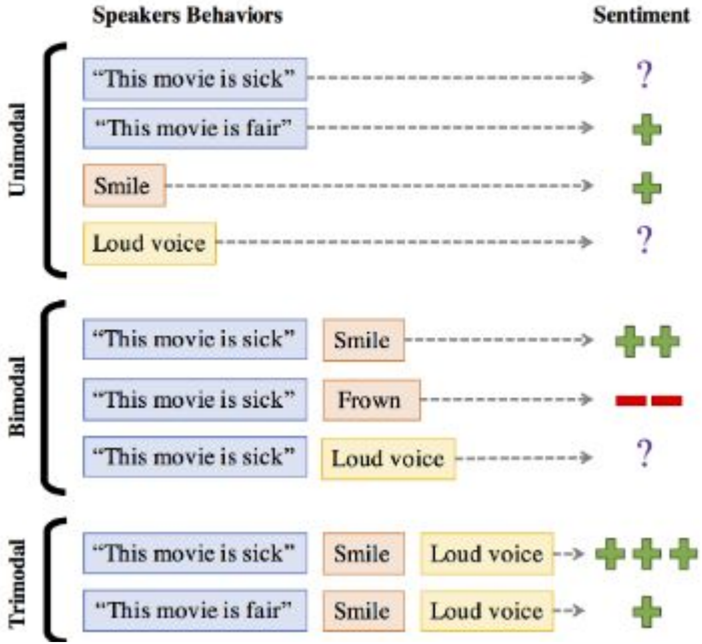
Not every modality has equal contribution to the prediction

Interpretability in Multimodal Deep Learning

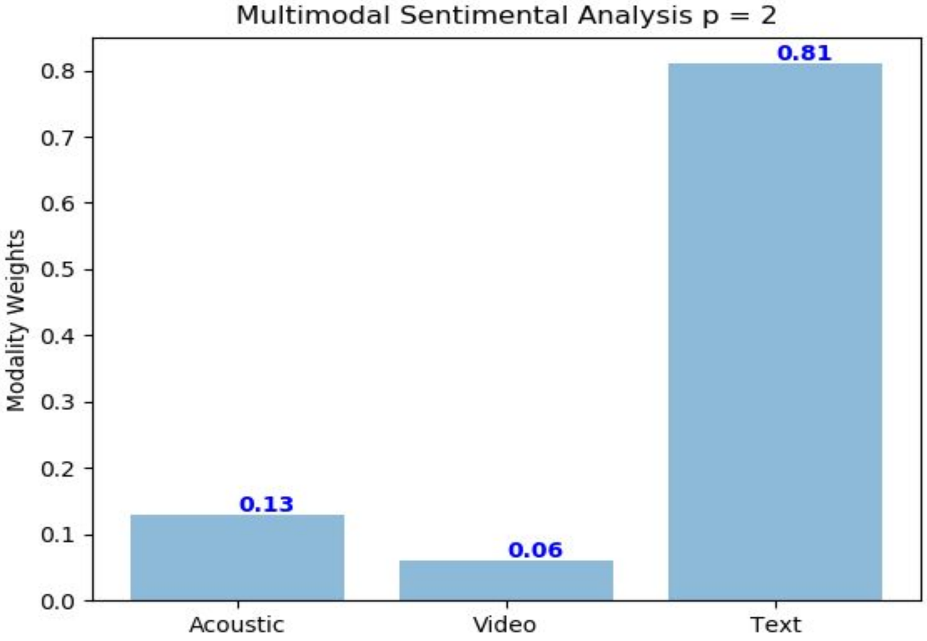
Solution -

We give different weights to different modalities

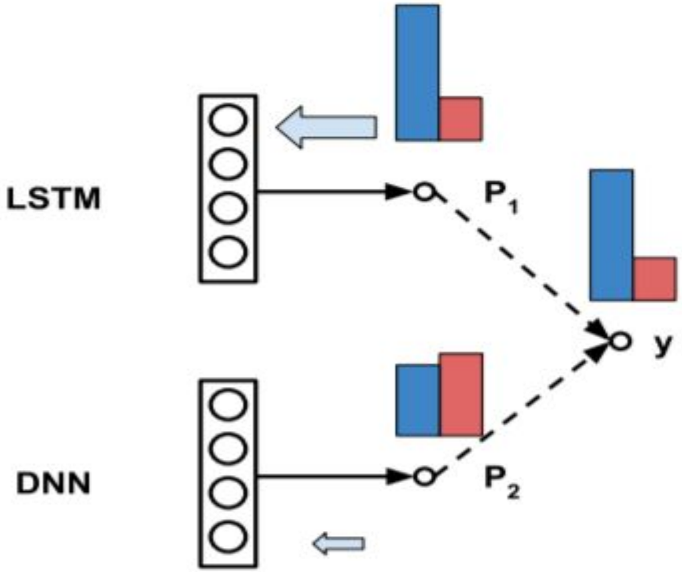
Real data experiments - Multimodal Sentiment Analysis(MOSI Dataset)


Speakers Behaviors
Sentiment
Unimodal
"This movie is sick" ?
"This movie is fair" +
Smile +
Loud voice ?
Bimodal
"This movie is sick" Smile ++
"This movie is sick" Frown −−
"This movie is sick" Loud voice ?
Trimodal
"This movie is sick" Smile Loud voice +++
"This movie is fair" Smile Loud voice +

MUltimodal Sentiment Analysis

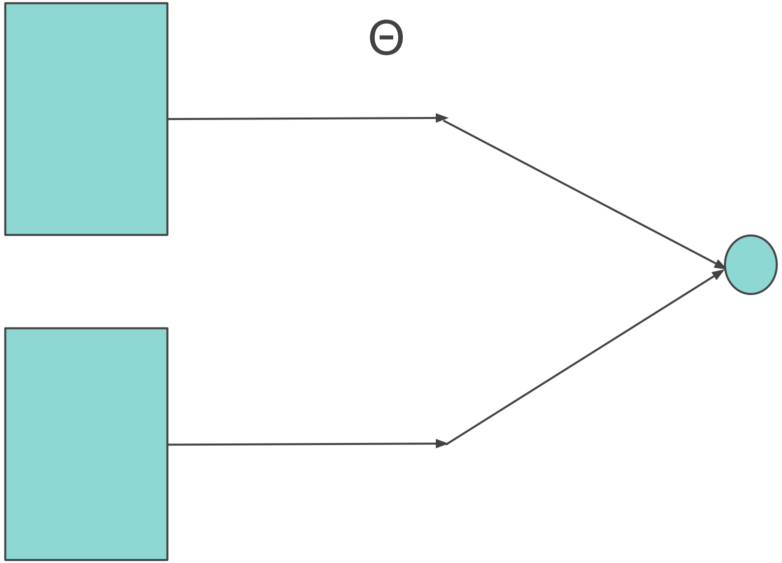


P1 and P2 contribution of each modality



$$N_p = \sum_{m=1}^{M} p_m < f^{m}_{w_1^m, w_2^m .. w_{L-1}^m}(x), W_L^m > +b$$

Θ



Modified loss function with ***β*** weight given to each modality

$$\min_{\substack{w_1, w_2, \ldots, w_L, \beta \\ \beta \in \mathbb{R}^M, \beta \geq 0, \|\beta\|_p \leq 1}} \left(\sum_{i=1}^{n} \ell\Big(\sum_{m=1}^{M} \sqrt{\beta_m} \langle w_L^m, f^m(x_i^m) \rangle + b, y_i \Big) + \Lambda \sum_{l=1}^{L} \sum_{m=1}^{M} \|w_l^m\|_2^2 \right)$$

Modified loss function with ***β*** weight given to each modality

$$\min_{\substack{w_1,w_2,\dots,w_L,\beta \\ \beta\in\mathbb{R}^M,\beta\geq 0,\|\beta\|_p\leq 1}} \left(\sum_{i=1}^{n} \ell\Big(\sum_{m=1}^{M} \sqrt{\beta_m}\langle w_L^m, f^m(x_i^m)\rangle + b, y_i\Big) + \Lambda \sum_{l=1}^{L}\sum_{m=1}^{M} \|w_l^m\|_2^2 \right)$$

$$w_L^m \leftarrow \sqrt{\beta_m}\, w_L^m,$$

Let’s make ***β*** trainable

$$\min_{\substack{w_1,w_2,\dots,w_L,\beta \\ \beta\in\mathbb{R}^M,\beta\geq 0,\|\beta\|_p\leq 1}} \sum_{i=1}^{n} \ell\left(\sum_{m=1}^{M} \langle w_L^m, f^m_{w_1^m,w_2^m,\dots,w_{L-1}^m}(x_i^m)\rangle + b, y_i \right) + \Lambda \sum_{l=1}^{L-1}\sum_{m=1}^{M} \|w_l^m\|_2^2 + \Lambda \sum_{m=1}^{M} \frac{\|w_L^m\|_2^2}{\beta_m}.$$

Modified loss function with β weight given to each modality

$$\min_{\substack{w_1, w_2, \ldots, w_L, \beta \\ \in \mathbb{R}^M, \beta \geq 0, \|\beta\|_p \leq 1}} \sum_{i=1}^{n} \ell \left(\sum_{m=1}^{M} \langle w_L^m, f^m_{w_1^m, w_2^m, \ldots, w_{L-1}^m}(x_i^m) \rangle + b, y_i \right)$$

$$+\Lambda \sum_{l=1}^{L-1} \sum_{m=1}^{M} \|w_l^m\|_2^2 + \Lambda \sum_{m=1}^{M} \frac{\|w_L^m\|_2^2}{\beta_m}.$$

Using the Lemma

$$\min_{\beta > 0, \|\beta\|_p^p \leq 1} \sum_{m=1}^{M} \frac{A_m}{\beta_m} = (\sum_{m=1}^{M} A_m^{\frac{p}{p+1}})^{\frac{p+1}{p}}$$

$$\min_{w_1, w_2, \ldots, w_L} \sum_{i=1}^{n} \ell \left(\sum_{m=1}^{M} \langle w_L^m, f^m(x_i^m) \rangle + b, y_i \right)$$

$$+\Lambda \sum_{l=1}^{L-1} \sum_{m=1}^{M} \|w_l^m\|_2^2 + \Lambda \left(\sum_{m=1}^{M} \|w_L^m\|_2^q \right)^{\frac{2}{q}},$$

Final Optimization problem

$$\min_{\substack{w_1, w_2, \dots, w_L, \beta \\ \beta \in \mathbb{R}^M, \beta \geq 0, \|\beta\|_p \leq 1}} \left(\sum_{i=1}^{n} \ell\Big(\sum_{m=1}^{M} \sqrt{\beta_m} \langle w_L^m, f^m(x_i^m) \rangle + b, y_i \Big) + \Lambda \sum_{l=1}^{L} \sum_{m=1}^{M} \|w_l^m\|_2^2 \right)$$

$$\min_{w_1, w_2, \dots, w_L} \sum_{i=1}^{n} \ell \left(\sum_{m=1}^{M} \langle w_L^m, f^m(x_i^m) \rangle + b, y_i \right)$$
$$+ \Lambda \sum_{l=1}^{L-1} \sum_{m=1}^{M} \|w_l^m\|_2^2 + \Lambda \left(\sum_{m=1}^{M} \|w_L^m\|_2^q \right)^{\frac{2}{q}},$$

$$\beta_m = \frac{\|w_l^m\|_2^{\frac{2}{p+1}}}{\left(\sum_{\tilde{m}=1}^{M} \|w_l^{\tilde{m}}\|_2^{\frac{2p}{p+1}} \right)^{\frac{1}{p}}}$$

Tensor fusion

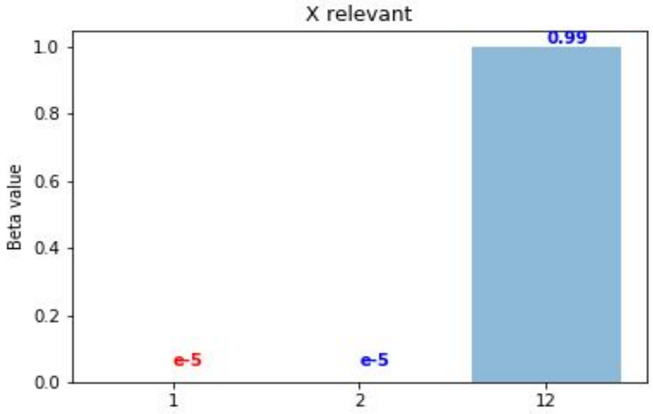
To allow interaction between modalities - take tensor product between modalities.

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \otimes \begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} = \begin{bmatrix} a\begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} & b\begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} \\ c\begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} & d\begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} aa^* & ac^* & ba^* & bc^* \\ ab^* & ad^* & bb^* & bd^* \\ ca^* & cc^* & da^* & dc^* \\ cb^* & cd^* & db^* & dd^* \end{bmatrix}$$

Problem with interpretability and tensorfusion - Degree Inflation

Beta weights came to be higher for higher dimension weights - Always high for tensor fusion.



Degree Inflation - reason

When one modality learns constant features.

The information from the first modality can be expressed in the tensorfusion.

Relevant source-with information

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \otimes \begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} = \begin{bmatrix} a\begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} & b\begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} \\ c\begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} & d\begin{bmatrix} a^* & c^* \\ b^* & d^* \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} aa^* & ac^* & ba^* & bc^* \\ ab^* & ad^* & bb^* & bd^* \\ ca^* & cc^* & da^* & dc^* \\ cb^* & cd^* & db^* & dd^* \end{bmatrix}$$

Irrelevant
source-constant

Iterative batch normalisation

We derived a batch norm which does not allow the lower dimension information to be represented in higher dimension(TensorFusion).

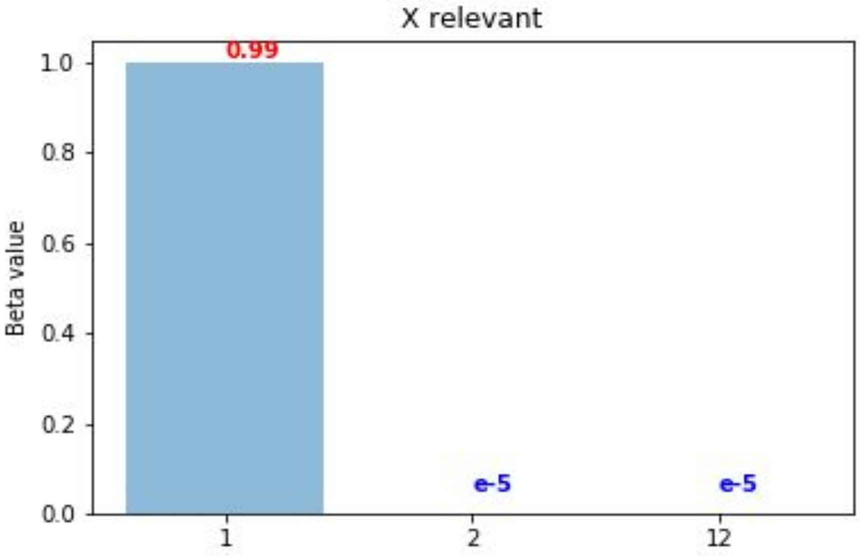
This method helps to remove noise from the data and give better accuracies.

Synthetic data experiment

- Relevant data - Sequences of length 100 (composed of 4 letters - ATGC) with a signature which defines the label

- Irrelevant data - Random 100 length sequences

Example - 1 as relevant source



Thank You