Countering Modal Redundancy and Heterogeneity: A Self-Correcting Multimodal Fusion

Pengkun Wang ¹, Xu Wang ¹, Binwu Wang ¹, Yudong Zhang ¹, Lei Bai ².*, and Yang Wang ¹.*

¹ University of Science and Technology of China (USTC), China
² Shanghai AI Laboratory, China

Reporter: Pengkun Wang
ICDM-2022
Outline

- Background
- Our Method
- Experiment
- Conclusion
Background

What is a Modality?

- A certain type of information or the representation format in which information is stored

  - Tactile, auditory, visual and olfactory data
  - Audio, image, video, text
  - Radar, infrared, accelerometer
  - Different languages
  - Data sets collected under different conditions

Fig. 1 Various Modalities.
Background

What is Multimodal Learning?

- Process and understand **multi-source modal information** by means of machine learning
- Five Challenges

Multimodal Fusion

Joining information from two or more modalities to perform a prediction
Purpose
- Extract unified and compact joint representations by using the complementarity and uniqueness among different modalities
- Apply the learned representations to prop up downstream applications

Related work
- Traditional methods
  - Bayesian based fusion
  - Sparse representation based fusion
- Deep learning based methods
  - Early fusion
  - Late fusion
  - Intermediate fusion
Background

**Purpose**
- Extract unified and compact joint representations by using the complementarity and uniqueness among different modalities.
- Apply the learned representations to support downstream applications.

**Related work**
- Traditional work:
  - Bayesian-based fusion
  - Sparse representation-based fusion
- Deep learning-based work:
  - Early fusion
  - Late fusion
  - Intermediate fusion

**Multimodal Fusion**

- Early fusion:
  - Net 1
  - Net 2
  - Result
- Late fusion:
  - Net 1
  - Net 2
  - Result
- Intermediate fusion:
  - Net 1
  - Net 2
  - Result
Our Method

- **Challenges I**
  - Feature redundancies
    - Irrelevant information
      - Caused by a general task-irrelevant feature extractor
    - Repetitive information
      - Similar information in the modal.

- **Motivation I**
  - Irrelevant information + Repetitive information (I+R)
    - Accumulation of redundancies → Serious semantic bias of fusion representations
  - Existing methods cannot be directly used to simultaneously deal with I+R

Fig. 2 Illustration of the redundancies in multimodal fusion.
Our Method

- **Challenges I**
  - Feature redundancies
    - Irrelevant information
      - Caused by a general task-irrelevant feature extractor
    - Repetitive information
      - Similar information in the modal.

- **Motivation I**
  - Irrelevant information + Repetitive information (I+R)
    → Accumulation of redundancies → Serious semantic bias of fusion representations
  - Existing methods cannot be directly used to simultaneously deal with I+R

Simultaneously deal with both irrelevant and repetitive information

Fig. 2 Illustration of the redundancies in multimodal fusion.
Our Method

- **Challenges II**
  - Feature homogeneity
    - Unified data structure
      - Easy to achieve feature interaction
  - Feature heterogeneity
    - Diverse data structure
      - Difficult to achieve feature interaction

- **Motivation II**
  - Existing methods fall short in processing data with diverse structures

Fig. 3  Feature homogeneity and feature heterogeneity.
Our Method

- **Challenges II**
  - Feature homogeneity
    - Unified data structure
      - Easy to achieve feature interaction
  - Feature heterogeneity
    - Diverse data structure
      - Difficult to achieve feature interaction

- **Motivation II**
  - Existing methods fall short in processing data with diverse structures

*Fig. 3 Feature homogeneity and feature heterogeneity.*
Our Method

- **Countering Modal Redundancy and Heterogeneity (CMRH)**
  - Unified Feature Interaction Module (UFIM)
    - Orthogonal attention component
    - Interactive feedback mechanism
    \[ \text{Countering heterogeneity} \]
  - Self-Correcting Transformer Module (SCTM)
    - Modified transformer
    - Fusion representation correction
    \[ \text{Countering redundancy} \]

- **Contributions**
  - First work that comprehensively understands the modal redundancy problem
  - A unified multimodal fusion strategy to counter modal redundancy and heterogeneity
  - Experiments on four cross-domain datasets show the effectiveness of CMRH
Our Method

- **Unified Feature Interaction Module (UFIM)**
  - Orthogonal attention component

  **Step 1.** Obtain the fine-grained attention map
  
  \[ M_{X,Y} = \text{Softmax} \left( \frac{E_X^i \otimes E_Y^i}{\sqrt{C}} \right) \]

  **Step 2.** Obtain the fine-grained attention-based representations

  \[
  \begin{align*}
  E_X^{i'} &= M_{X,Y} \otimes E_X^i \\
  E_Y^{i'} &= M_{X,Y}^\top \otimes E_Y^i
  \end{align*}
  \]

- Interactive feedback mechanism

  **Step 3.** Fine-grained attention-based representations are fed back to the original feature

  \[
  \begin{align*}
  \hat{E}_X^i &= E_X^i + \alpha \cdot E_Y^{i'} \\
  \hat{E}_Y^i &= E_Y^i + \alpha \cdot E_X^{i'}
  \end{align*}
  \]

Fig. 4 An illustration of our proposed UFIM.
Our Method

- **Self-Correcting Transformer Module (SCTM)**
  - Modified transformer

**Step 1.** Transfer features to a consistent dimension and obtain attention-based feature maps

\[
\begin{align*}
M'_X &= T(\text{Func}_X(X'), \text{Func}_Y(Y'), \text{Func}_Y(Y')) \\
M'_Y &= T(\text{Func}_Y(Y'), \text{Func}_X(X'), \text{Func}_X(X'))
\end{align*}
\]

**Substep 1.** Features are equally divided into blocks

**Substep 2.** Each block is concatenated with the position embedding of this block and the modal-identity embedding of the current modality.

**Substep 3.** Obtain the attention-based feature map

Fig. 5 Illustrated of the modified transformer.
Our Method

- **Self-Correcting Transformer Module (SCTM)**
  - Modified transformer

  **Step 2.** Feed the obtained attention-based feature maps back to their original modal features

  \[
  \begin{align*}
  \hat{X}' &= \mathcal{M}'_X + X' \\
  \hat{Y}' &= \mathcal{M}'_Y + Y'
  \end{align*}
  \]

- Fusion representation correction

  **Step 3.** Obtain the fusion representation

  \[
  \mathcal{P} = \mathcal{F}(\hat{X}', \hat{Y}')
  \]

  **Step 4.** Calculate the element-wise weighted average feature map as the weights of fusion representation

  \[
  \hat{\mathcal{P}} = \mathcal{P} \odot \text{Norm}(\frac{\mathcal{M}'_X + \mathcal{M}'_Y}{2})
  \]

Fig. 6 An illustration of our proposed SCTM.
Task 1: Hand Gesture Recognition

- **Dataset**
  - EgoGesture

- **Implementation**
  - I3D for RGBs and depth maps
  - Apply UFIMs to the last three inception modules
  - Improve the fusion module with SCTM

- **Analysis**
  - With UFIM:
    - Perform other non-interactive methods and MMTM
    - Properly feature interactions help improve the expressiveness of representations
  - With SCTM
    - Perform more effectively (compared with I3D late fusion)
    - Correcting the redundancy makes the final representation more can represent the fused modalities
  - With UFIM and SCTM
    - Outperform the top performer MMTM by 1.09%
    - Mutually compatible

![An overview of the improved hand gesture recognition framework.](image)
Experiment

- **Task 2: House Price Prediction**
  - Dataset: NYC and BEIJING dataset

- **Task 3: Action Recognition**
  - Dataset: NTU-RGBD

- **Task 4: Traffic Accident Forecast**
  - Dataset: NYC and SIP dataset

- **Discussion of the UFIM and SCTM**
  - Location of the interaction
    - Optimal location for module insertion is the tail layer of the model
  - Number of the interaction
    - NOT More than one-third of the entire model layers
  - Selection of weight parameters
    - setting $\alpha$ at about 0.2 and fine-tuning it according to the actual task.
Conclusion

- We propose a **unified multimodal fusion strategy**, including two well-designed modules, UFIM and SCTM, for addressing both modal **heterogeneity** and **redundancy** by exploiting the inter-modal complementarity.

- UFIM and SCTM can be **flexibly applied** to existing multimodal fusion networks at a relatively **low cost**.

- Extensive experiments on **four different cross-domain datasets** from the fields of hand gesture recognition, house price prediction, action recognition, and traffic accident forecast show the **effectiveness** of the proposed modules.
Thanks for your listening!

For more details, please refer to our paper!

Reporter: Pengkun Wang
pengkun@mail.ustc.edu.cn