# UniS-MMC: Multimodal Classification via Unimodality-supervised Multimodal Contrastive Learning

**Heqing Zou, Meng Shen, Chen Chen, Yuchen Hu, Deepu Rajan, Eng Siong Chng**
Nanyang Technological Universityy, Singapore
{heqing001, meng005, chen1436, yuchen005}@e.ntu.edu.sg, {asdrajan, aseschng}ntu.edu.sg

ACL 2023

https://github.com/Vincent-ZHQ/UniS-MMC

**Reported by Jiawei Cheng**

# Introduction

Despite their effectiveness in learning the correspondence among modalities, these contrastive-based multimodal learning methods still meet with problems with the sensor noise in the in-the-wild datasets

The current methods always treat each modality equally and ignore the difference of the role for different modalities, The final decisions will be negatively affected by those samples with inefficient unimodal representations and thus can not provide trustworthy multimodal representations.

# Method

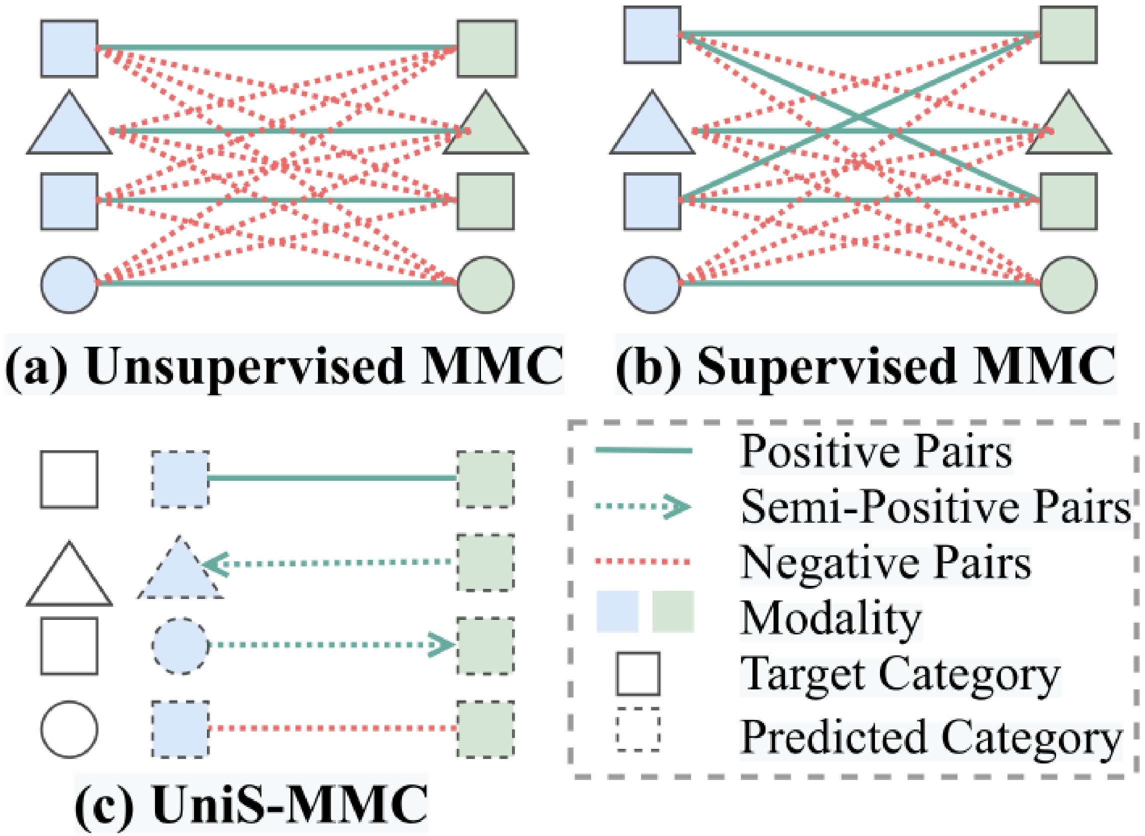



Figure 3: The relationship comparison between two modalities in training mini-batch of (a) unsupervised MMC, (b) supervised MMC and (c) UniS-MMC.

Table 1: Contrastive settings.

| Uni-Prediction | Modality $a$ | Modality $b$ | Category |
|---|---|---|---|
| 0 | True | True | Positive |
| 1 | True | False | Semi-positive |
| 2 | False | True | |
| 3 | False | False | Negative |

# Method

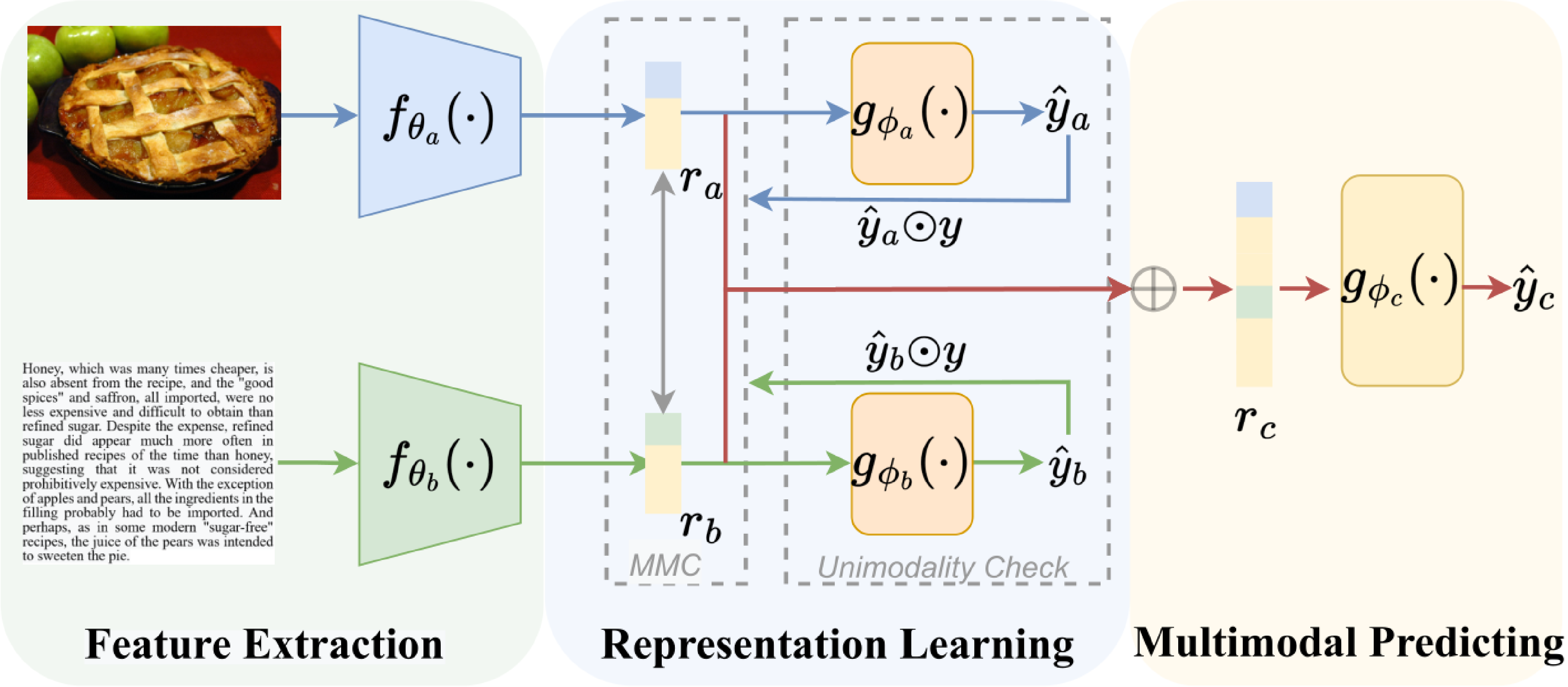
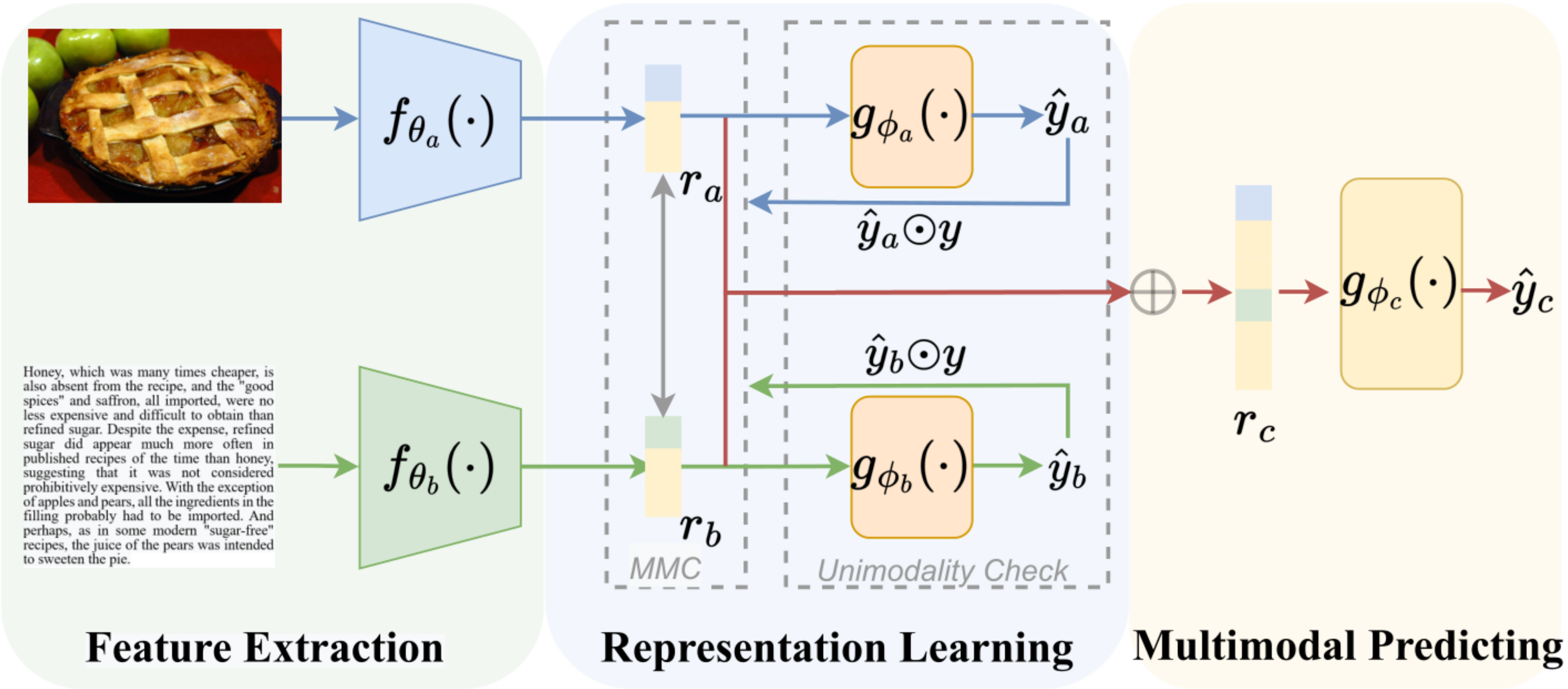



Figure 2: The framework for our proposed UniS-MMC.

# Method



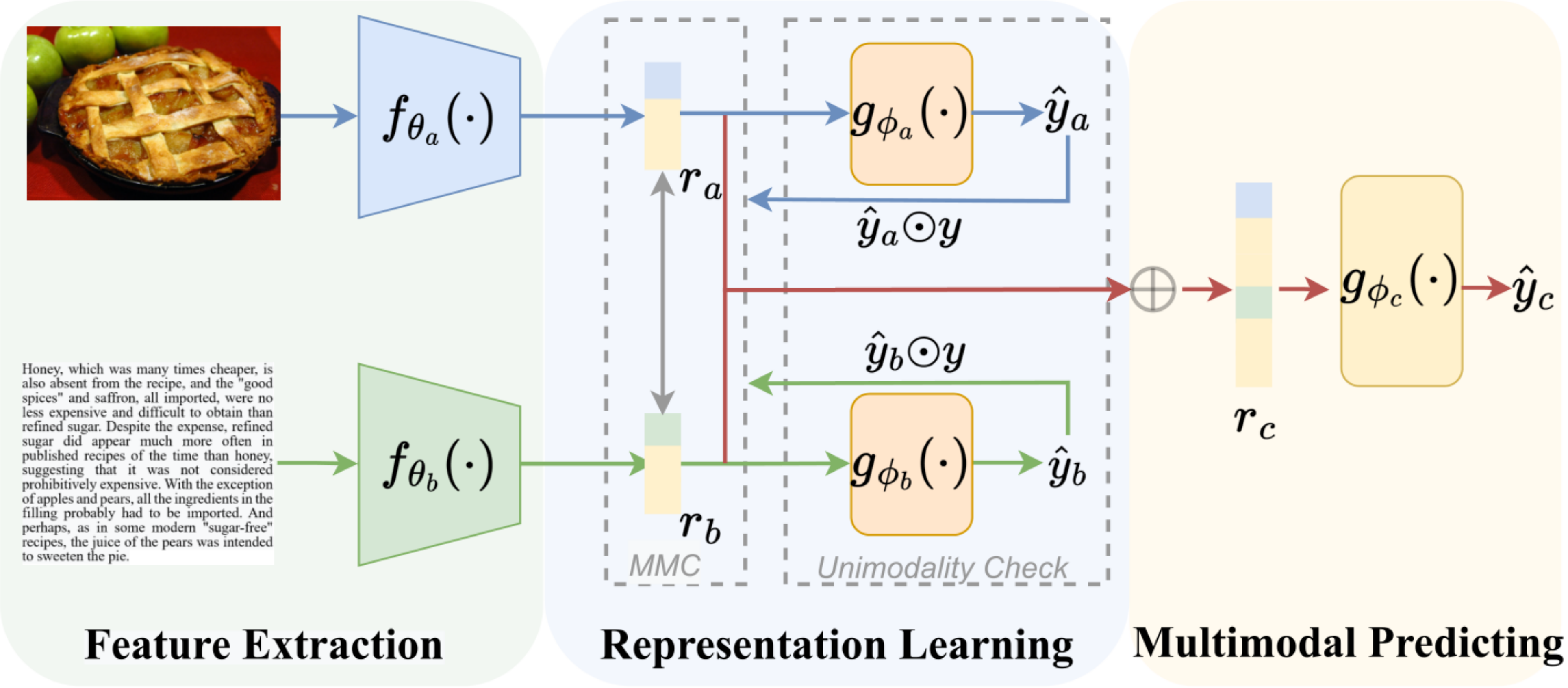

$$\mathcal{L}_{uni} = -\sum_{m=1}^{M}\sum_{k=1}^{K} y^k \log p_m^k, \qquad (1)$$

$$\mathcal{L}_{b-mmc} = -\log \frac{\sum_{n\in\mathbb{P},\mathbb{S}}(\exp(\cos(r_a^n, r_b^n)/\tau)}{\sum_{n\in\mathbb{B}}(\exp(\cos(r_a^n, r_b^n)/\tau)}, \qquad (2)$$

$$\mathcal{L}_{mmc} = \sum_{i=1}^{M}\sum_{j>i}^{M} \mathcal{L}_{b-mmc}(m_i, m_j), \qquad (3)$$

$$\mathcal{L}_{multi} = -\sum_{k=1}^{K} y^k \log p_k^k, \qquad (4)$$

# Method




$$\mathcal{L}_{UniS-MMC} = \mathcal{L}_{uni} + \mathcal{L}_{multi} + \lambda \mathcal{L}_{mmc}, \qquad (5)$$

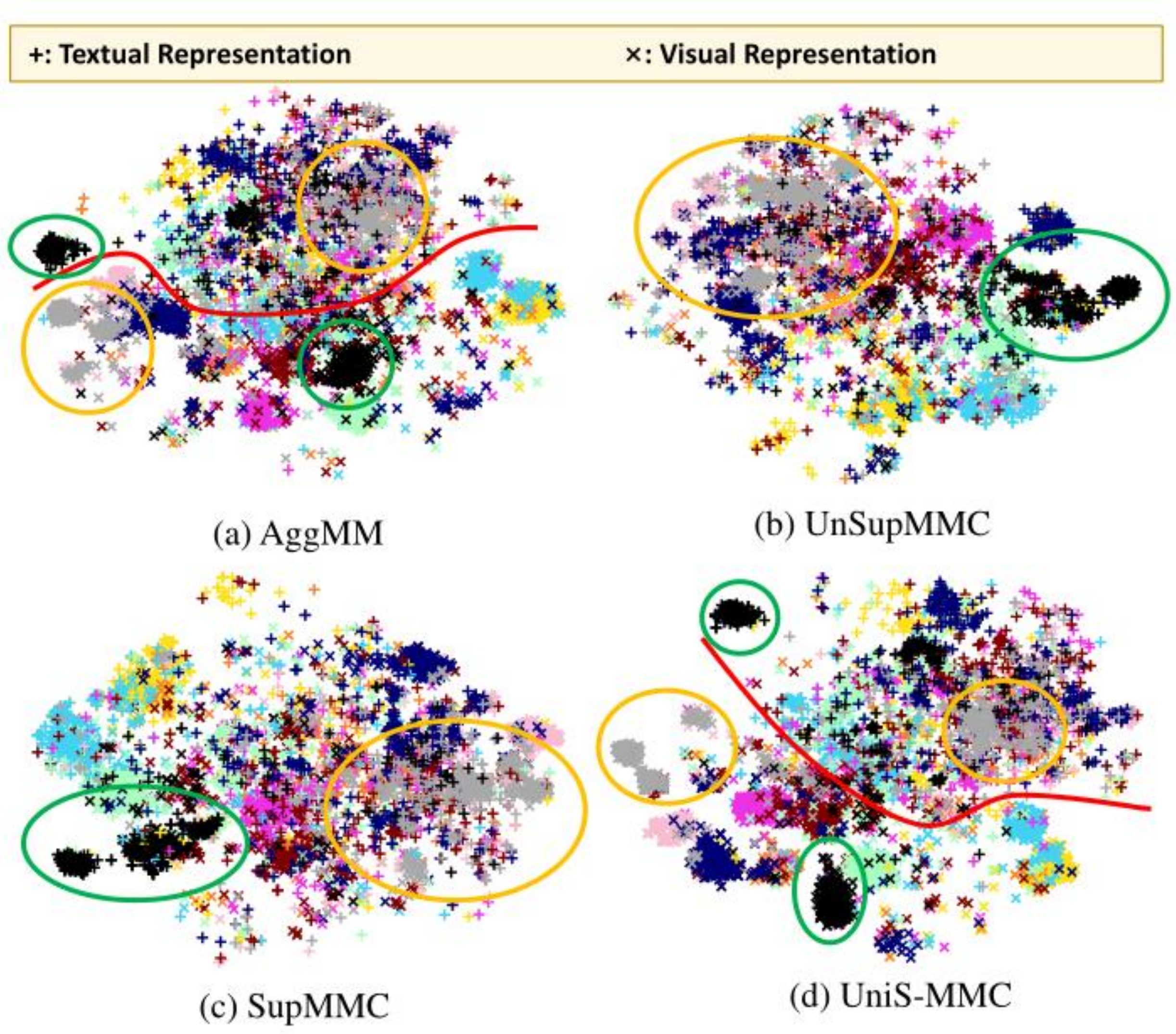
# Experiments

Table 2: Comparison of multimodal classification performance on a) Food101 and b) N24News.

| a) Model | Fusion | | Backbone | | Acc |
|---|---|---|---|---|---|
| | AGG | ALI | Image | Text | |
| MMBT | Early | ✗ | ResNet-152 | BERT | $92.1_{\pm 0.1}$ |
| HUSE | Early | ✓ | Graph-RISE | BERT | 92.3 |
| ViLT | Early | ✓ | ViT | BERT | 92.0 |
| CMA-CLIP | Early | ✓ | ViT | Transformer | 93.1 |
| ME | Early | ✗ | DenseNet | BERT | 94.6 |
| AggMM | Early | ✗ | ViT | BERT | $93.7_{\pm 0.2}$ |
| UnSupMMC | Early | ✓ | ViT | BERT | $94.1_{\pm 0.7}$ |
| SupMMC | Early | ✓ | ViT | BERT | $94.2_{\pm 0.2}$ |
| **UniS-MMC** | Early | ✓ | ViT | BERT | $\mathbf{94.7}_{\pm 0.1}$ |

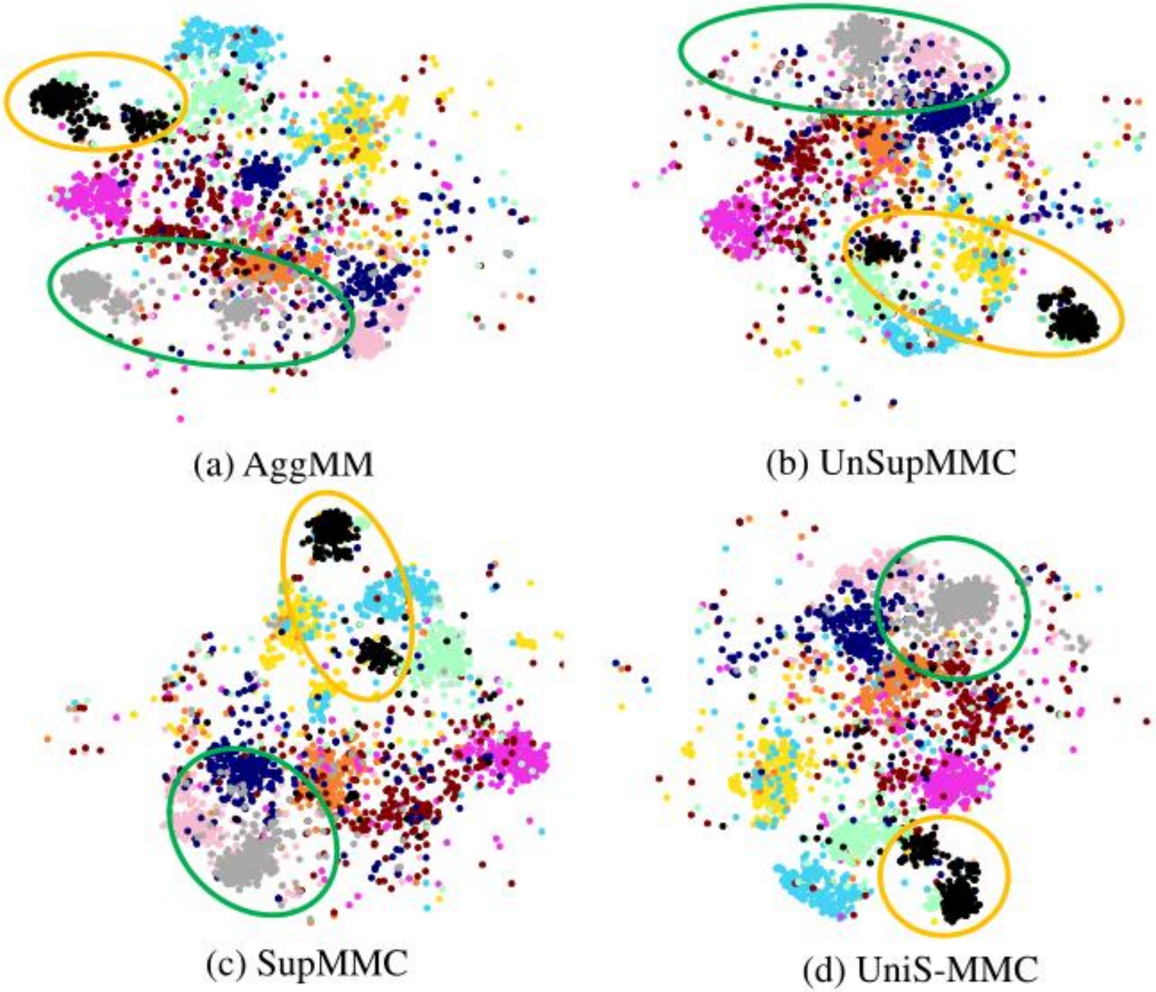
| b) Model | Fusion | | Backbone | | Multimodal | | |
|---|---|---|---|---|---|---|---|
| | AGG | ALI | Image | Text | Headline | Caption | Abstract |
| N24News | Early | ✗ | ViT | RoBERTa | 79.41 | 77.45 | 83.33 |
| AggMM | Early | ✗ | ViT | BERT | $78.6_{\pm 1.1}$ | $76.8_{\pm 0.2}$ | $80.8_{\pm 0.2}$ |
| UnSupMMC | Early | ✓ | ViT | BERT | $79.3_{\pm 0.8}$ | $76.9_{\pm 0.3}$ | $81.9_{\pm 0.3}$ |
| SupMMC | Early | ✓ | ViT | BERT | $79.6_{\pm 0.5}$ | $77.3_{\pm 0.2}$ | $81.7_{\pm 0.8}$ |
| UniS-MMC | Early | ✓ | ViT | BERT | $\mathbf{80.2}_{\pm 0.1}$ | $\mathbf{77.5}_{\pm 0.3}$ | $\mathbf{83.2}_{\pm 0.4}$ |
| AggMM | Early | ✗ | ViT | RoBERTa | $78.9_{\pm 0.3}$ | $77.9_{\pm 0.3}$ | $83.5_{\pm 0.2}$ |
| UnSupMMC | Early | ✓ | ViT | RoBERTa | $79.9_{\pm 0.2}$ | $78.0_{\pm 0.1}$ | $83.7_{\pm 0.3}$ |
| SupMMC | Early | ✓ | ViT | RoBERTa | $79.9_{\pm 0.4}$ | $77.9_{\pm 0.2}$ | $84.0_{\pm 0.2}$ |
| **UniS-MMC** | Early | ✓ | ViT | RoBERTa | $\mathbf{80.3}_{\pm 0.1}$ | $\mathbf{78.1}_{\pm 0.2}$ | $\mathbf{84.2}_{\pm 0.1}$ |

# Experiments



# Experiments

· **Multimodal Representation**



(a) AggMM (b) UnSupMMC (c) SupMMC (d) UniS-MMC

# Experiments

Table 3: Comparison to unimodal learning and the baseline model on N24News.

| Dataset | Text | Image-only | BERT-based | | | RoBERTa-based | | |
|---|---|---|---|---|---|---|---|---|
| | | | Text-only | AggMM | **UniS-MMC** | Text-only | AggMM | **UniS-MMC** |
| **N24News** | Headline | $54.1_{\pm 0.2}$ | $72.1_{\pm 0.2}$ | $78.6_{\pm 1.1}$ | $\mathbf{80.2}_{\pm 0.1}$ ↑ 1.6 | $71.8_{\pm 0.2}$ | $78.9_{\pm 0.3}$ | $\mathbf{80.3}_{\pm 0.1}$ ↑ 1.4 |
| | Caption | | $72.7_{\pm 0.3}$ | $76.8_{\pm 0.2}$ | $\mathbf{77.5}_{\pm 0.3}$ ↑ 0.7 | $72.9_{\pm 0.4}$ | $77.9_{\pm 0.3}$ | $\mathbf{78.1}_{\pm 0.2}$ ↑ 0.3 |
| | Abstract | | $78.3_{\pm 0.3}$ | $80.8_{\pm 0.2}$ | $\mathbf{83.2}_{\pm 0.4}$ ↑ 2.4 | $79.7_{\pm 0.2}$ | $83.5_{\pm 0.2}$ | $\mathbf{84.2}_{\pm 0.1}$ ↑ 0.7 |

# Experiments

Table 4: Ablation study on N24News.

| **Method** | Headline | | Caption | | Abstract | |
|---|---|---|---|---|---|---|
| | BERT | RoBERTa | BERT | RoBERTa | BERT | RoBERTa |
| AggMM | $78.6_{\pm 1.1}$ | $78.9_{\pm 0.3}$ | $76.8_{\pm 0.2}$ | $77.9_{\pm 0.3}$ | $80.8_{\pm 0.2}$ | $83.5_{\pm 0.2}$ |
| $+ L_{uni}$ | $79.4_{\pm 0.4}$ | $79.4_{\pm 0.3}$ | $77.3_{\pm 0.2}$ | $77.9_{\pm 0.1}$ | $82.5_{\pm 0.3}$ | $84.1_{\pm 0.2}$ |
| $+ C_{Semi}$ | $80.1_{\pm 0.1}$ | $80.0_{\pm 0.3}$ | $77.3_{\pm 0.2}$ | $78.0_{\pm 0.3}$ | $82.7_{\pm 0.4}$ | $84.2_{\pm 0.3}$ |
| $+ C_{Neg}$ | $\mathbf{80.2}_{\pm 0.1}$ | $\mathbf{80.3}_{\pm 0.1}$ | $\mathbf{77.5}_{\pm 0.3}$ | $\mathbf{78.1}_{\pm 0.2}$ | $\mathbf{83.2}_{\pm 0.4}$ | $\mathbf{84.2}_{\pm 0.1}$ |

# Experiments

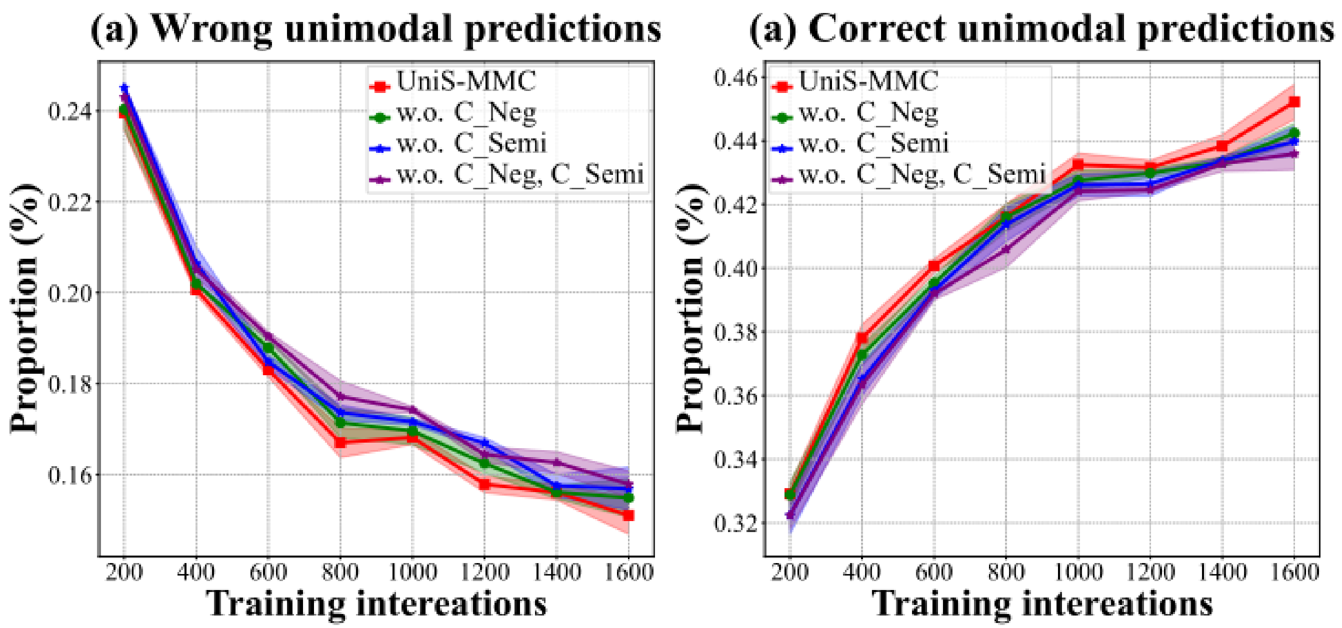



Figure 6: As the training progresses, the change of the proportion of both wrong (left), both correct (right) unimodal predictions of the validation set (N24News): the complete method (UniS-MMC), remove negative pair (w.o. C_Neg), remove semi-positive pair (w.o. C_Semi) and remove both (w.o. C_Neg,C_Semi).

# Experiments

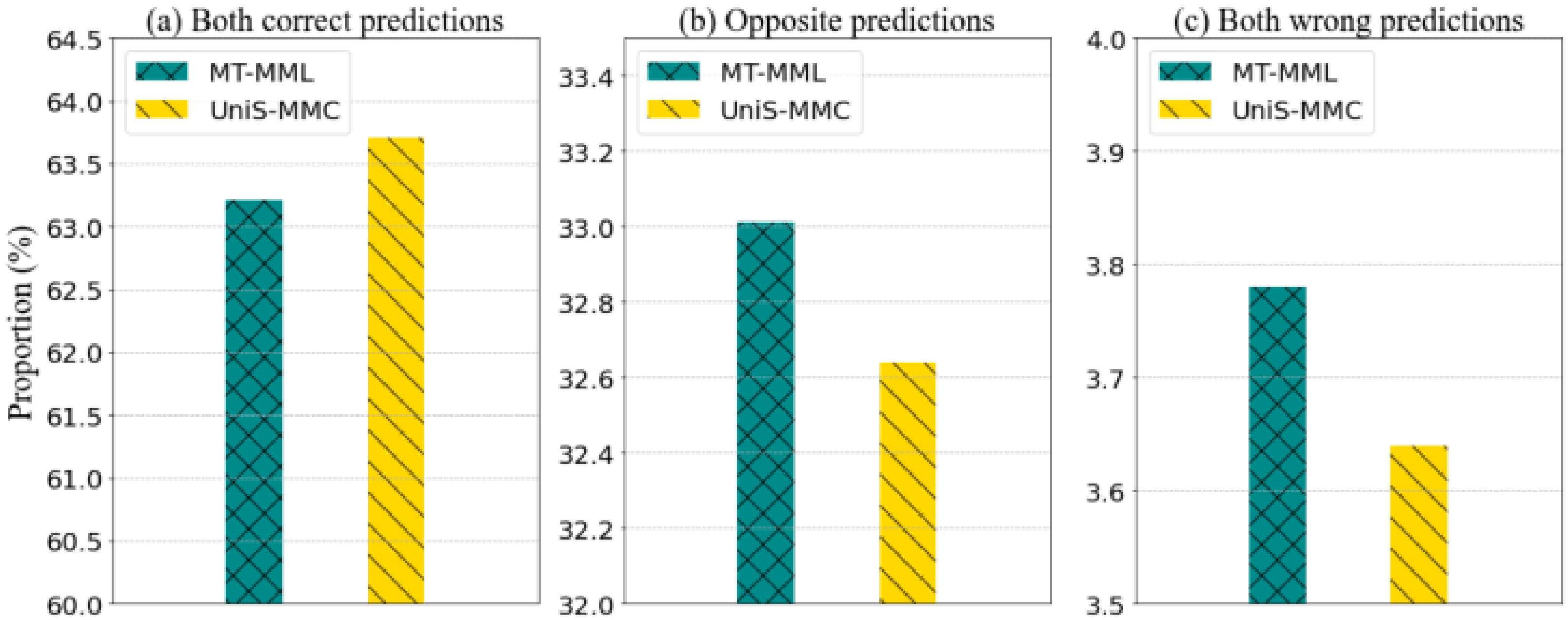



Figure 7: Consistency comparison of unimodal prediction between MT-MML and the UniS-MMC.