
Research on Efficient Modality Fusion for Enhanced Uni-modal Ensemble

향상된 단일모달 모델 앙상블을 위한 효율적인 모달리티 결합에 관한 연구

Intelligent Information Processing Lab
KiHoon Lee

Large Multi-modal Model?

LLM은 옛말...이미지까지 학습한 'LMM' 뜬다

✎ 임대준 기자 | ⌚ 입력 2023.10.12 18:00 | 💬 댓글 0 | ❤️ 좋아요 0

글만 알던 생성AI...영상 보고 감정 읽는
'멀티모달'로

AI 트렌드 체크 : GPT-4V, LLM 시대를 지나 이제
'LMM'이 온다

2023.10.10. 오후 4:25

챗GPT 출시 1년 만에
기술 트렌드 확 바뀌어

'GPT-4V' 이어 '제미니' 공개 임박...오픈 소스 '라바'도 인기

사람처럼 보고 듣고 말하는 'LMM' 시대 온다

✎ 구아현 기자 | ⌚ 입력 2023.10.20 17:24 | ⌚ 수정 2023.10.20 20:52

💬 0 | ⌚ - 카 +

Large Multi-modal Model?

LLM은 옛말...이미지까지 학습한 'LMM' 뜬다

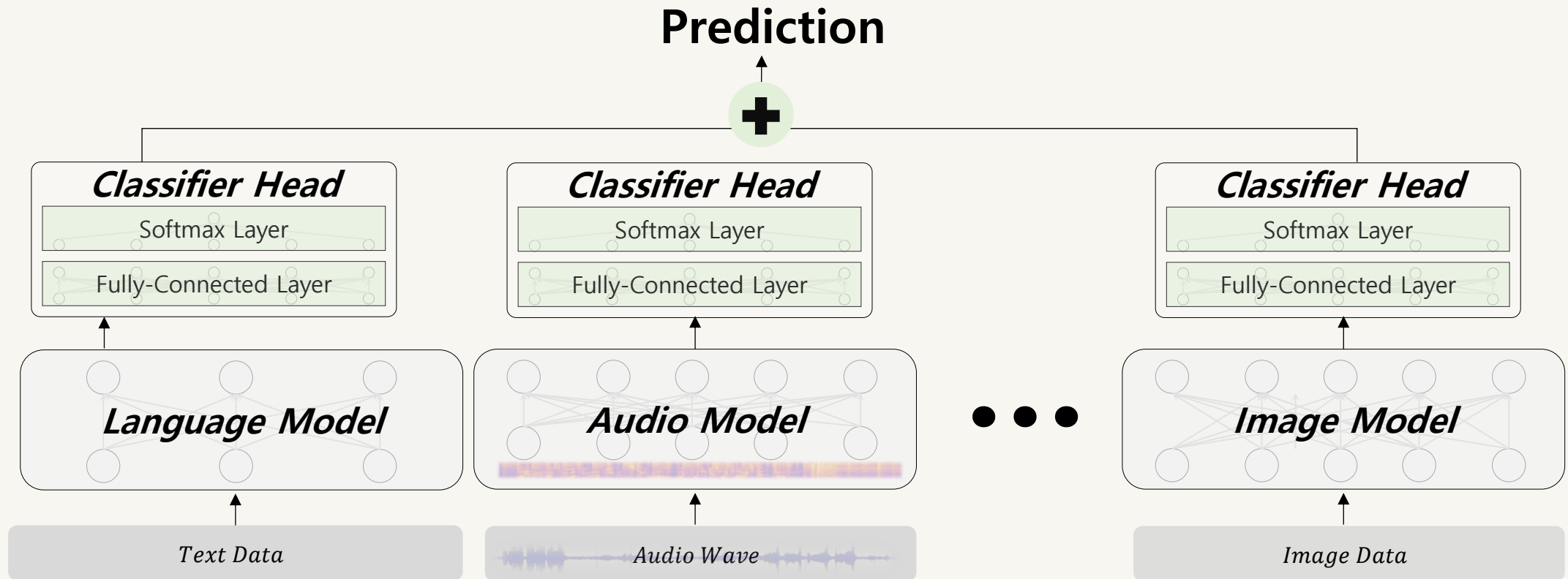


모달리티 별 적합한 딥러닝 모델

- 시계열 정보 ➡ RNN, LSTM, GRU, 1D CNN, Transformer ...
- 텍스트 데이터 ➡ BERT, RoBERTa, ALBERT, DeBERTa, DistilBERT, ELECTRA, GPT, T5, LLaMA, XLNet...
- 이미지 데이터 ➡ CNN, ResNET, VGG, ViT, BEiT, DeiT, EfficientNet, Swin Transformer, YOLO, ...
- 오디오 데이터 ➡ Speech2Text, WavLM, Wav2Vec, Whisper, Hubert...
- 멀티모달 데이터 ➡ VisualBERT, BLIP, PaLi, BeiT-3, CoCa, VLMo, ViLT, ...

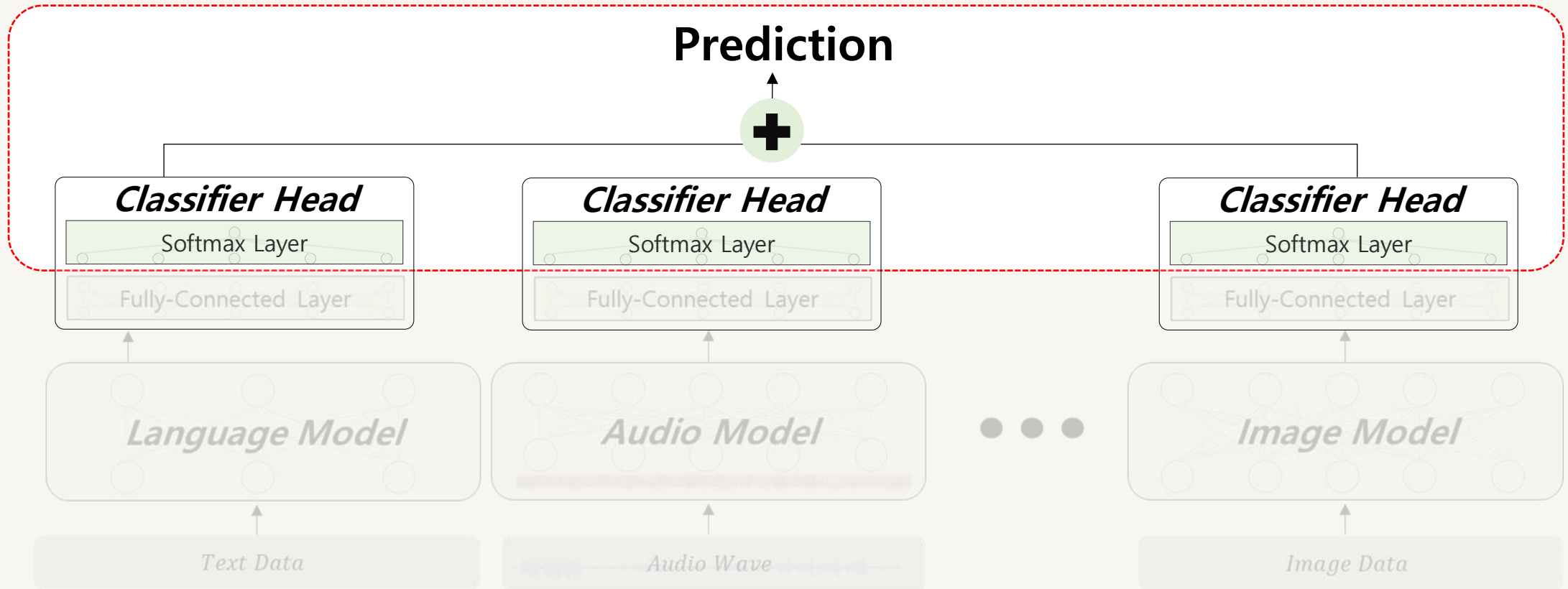
단일모달 앙상블 방안

- 한정된 자원을 사용하여 효과적으로 문제를 해결



단일모달 앙상블 방안

- 한정된 자원을 사용하여 효과적으로 문제를 해결



Graduation Paper

- **Tittle**

- 향상된 단일모달 모델 앙상블을 위한 효율적인 모달리티 결합에 관한 연구
- Research on Efficient Modality Fusion for Enhanced Uni-modal Ensemble

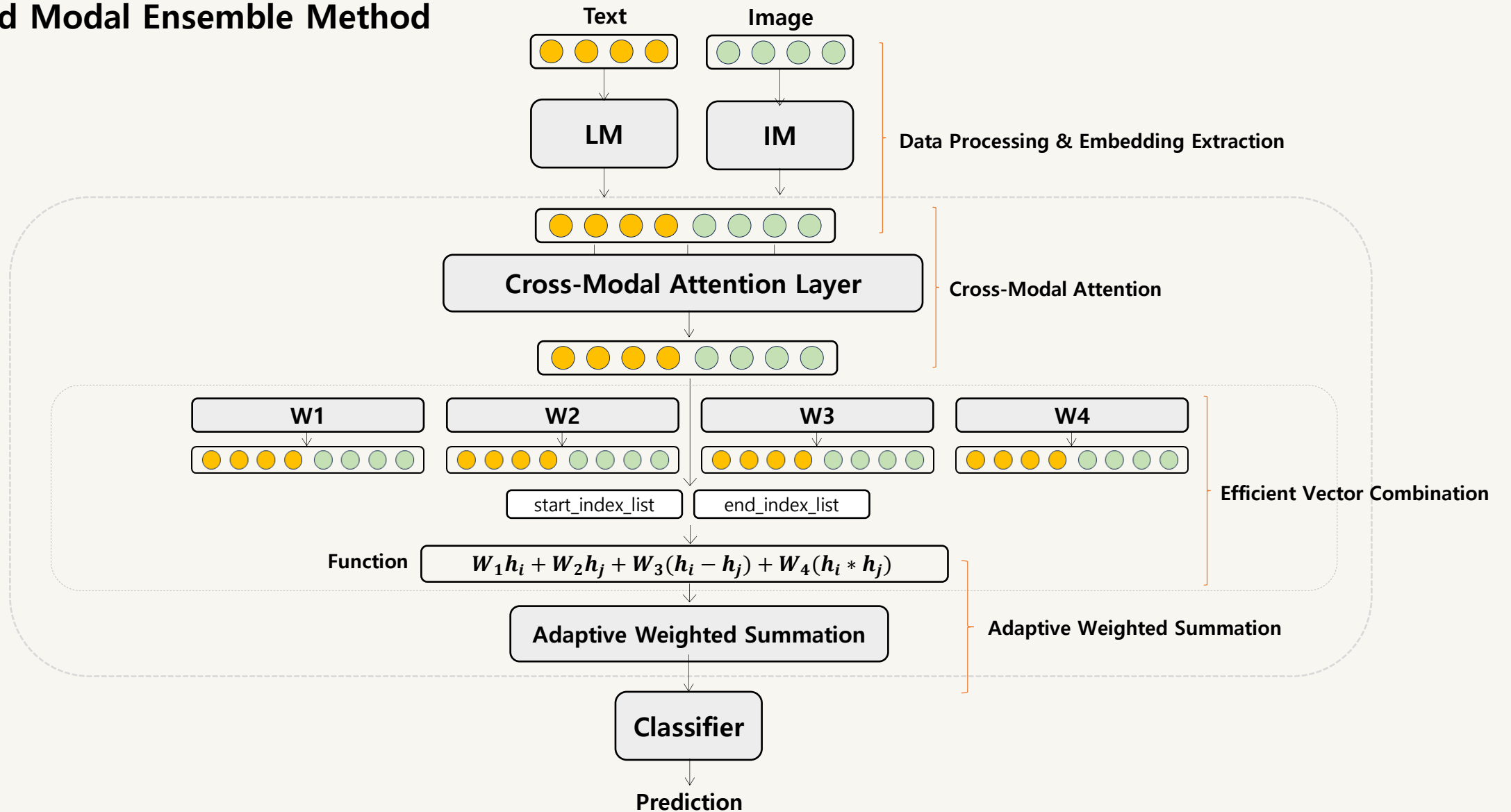
- **Research Motivation**

- 실증적인 연구 부족
 - 데이터간의 결합 과정에서 정보 손실 문제를 최소화하는 연구
 - 데이터간의 이질성을 최소화할 수 있는 벡터 조합 연구
 - 단일 모달을 결합하여 멀티모달 문제를 해결하는 연구

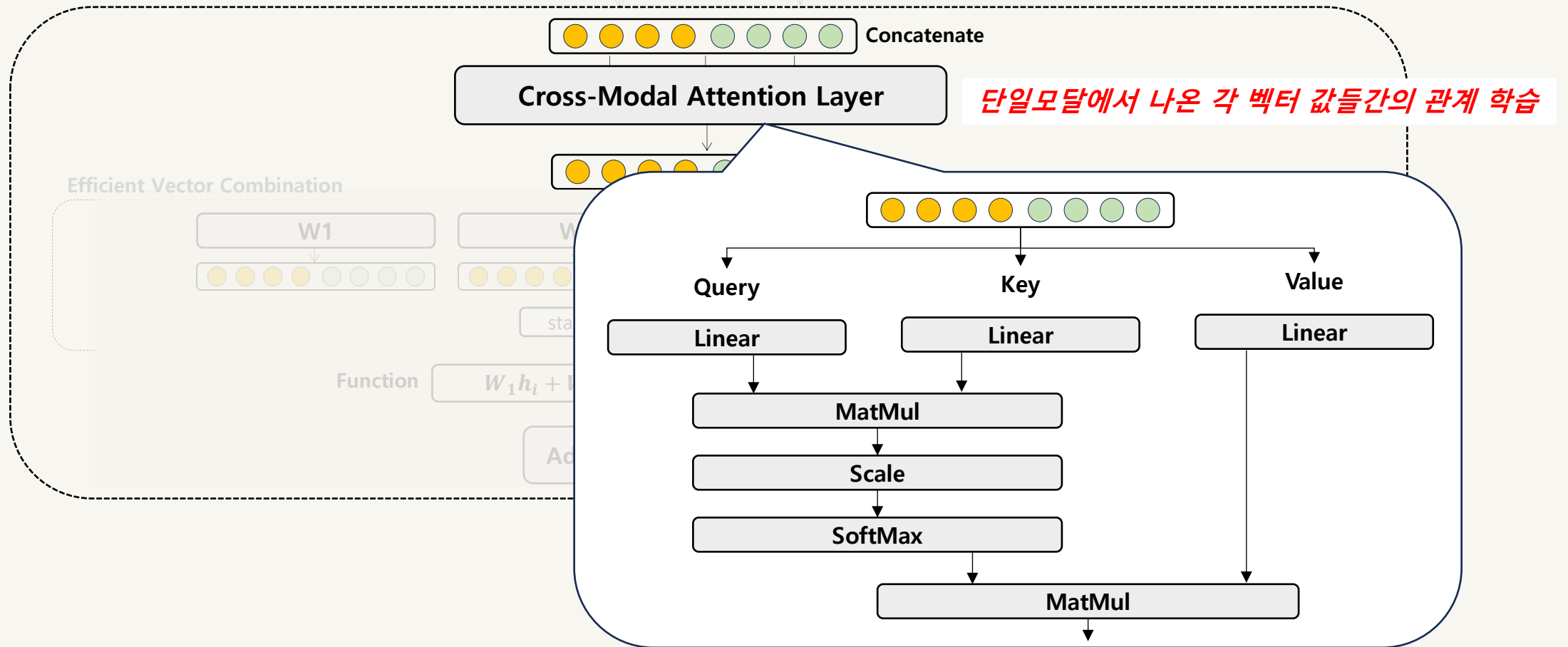
- **Research Objective**

- 단일 모달리티 모델의 앙상블을 통한 효율적인 모달리티 결합 방안 개발
- 데이터 간의 이질성을 최소화하여 모델의 학습 및 추론 성능을 향상시키는 방법 연구
- 제안하는 모달리티 결합 방안의 성능을 실증적으로 검증
- 모든 유형의 데이터와 딥러닝 네트워크에 적용 가능한 범용성 있는 결합 방안을 제안

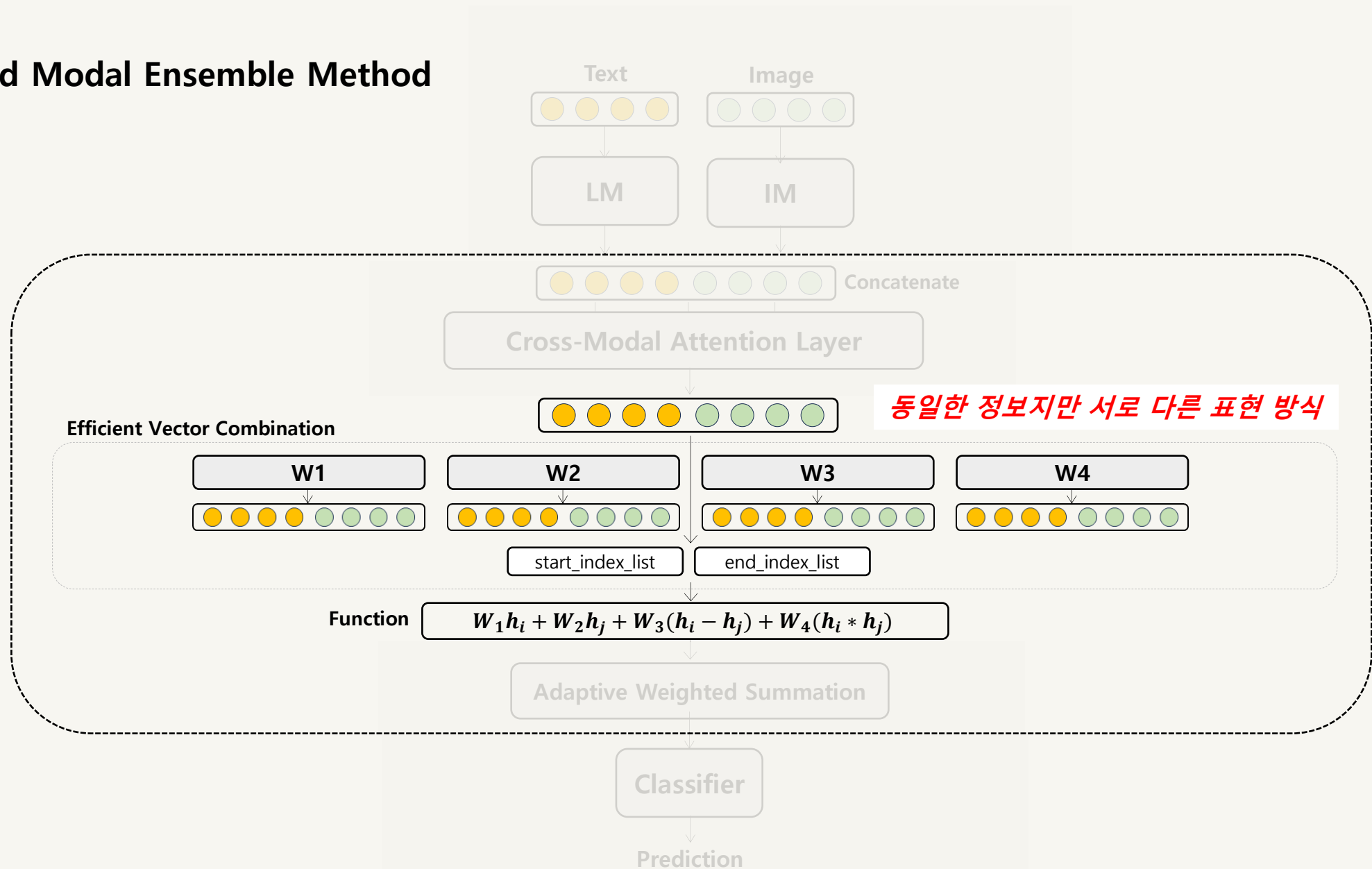
Unified Modal Ensemble Method



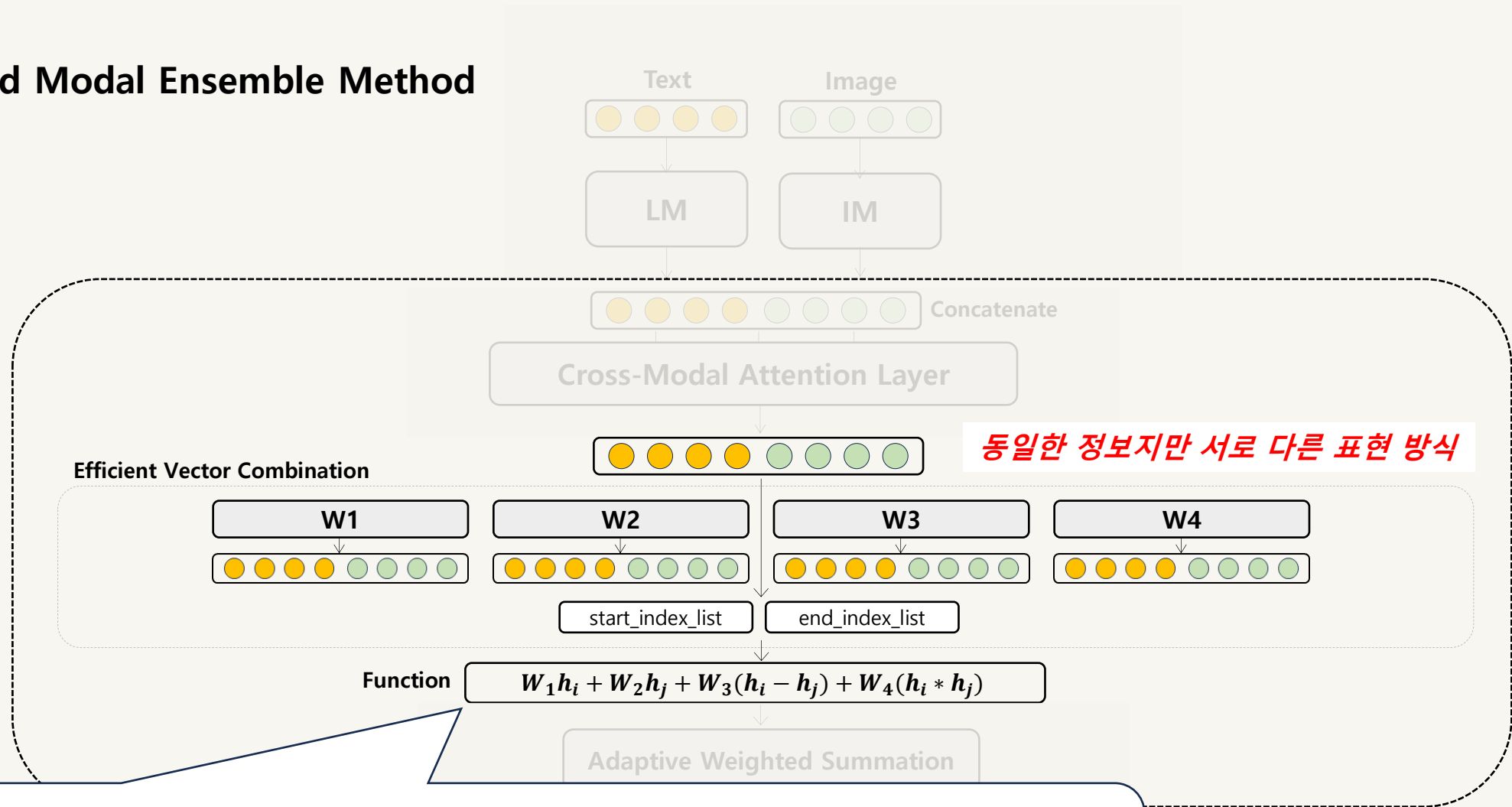
Unified Modal Ensemble Method



Unified Modal Ensemble Method



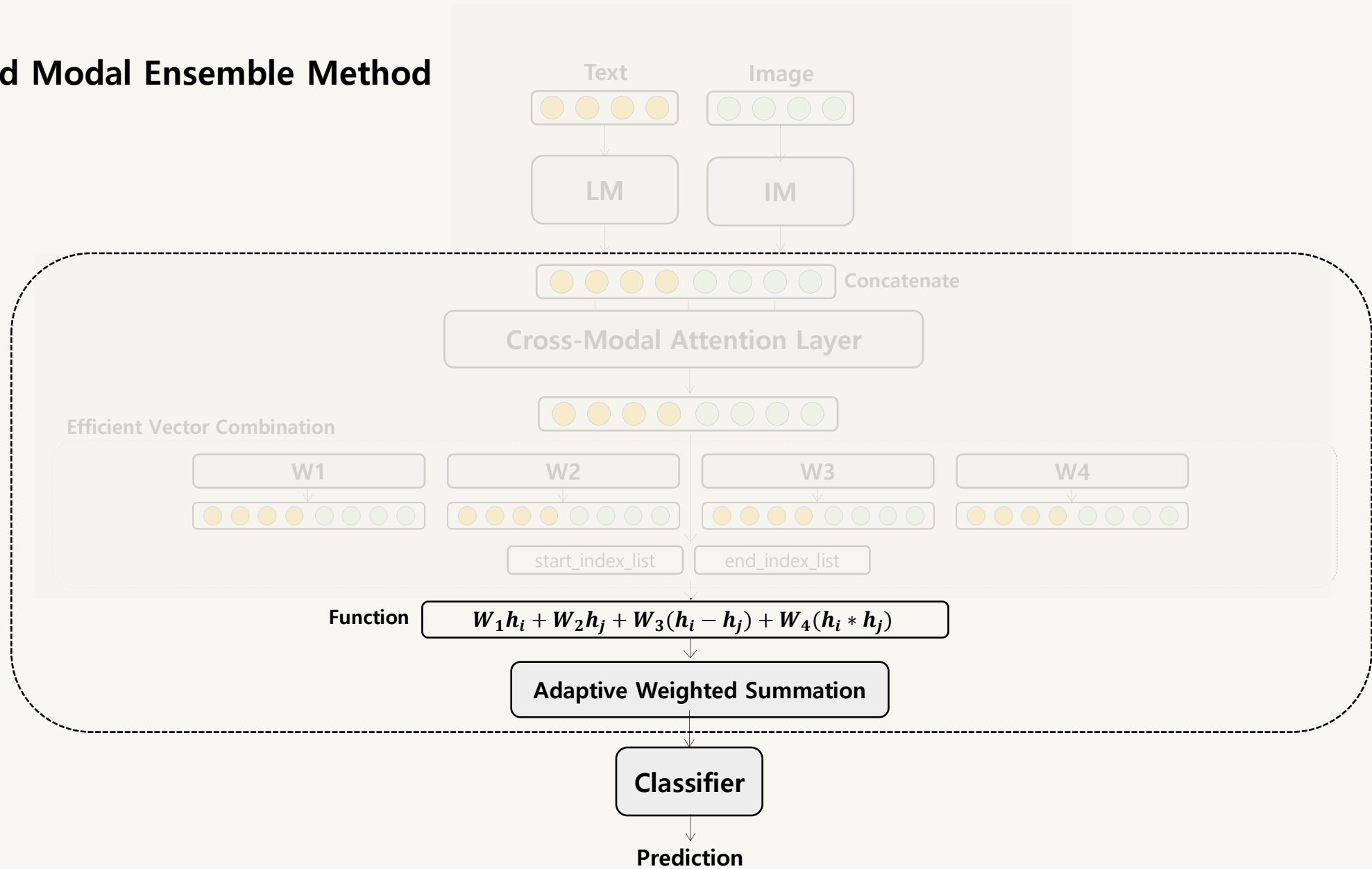
Unified Modal Ensemble Method



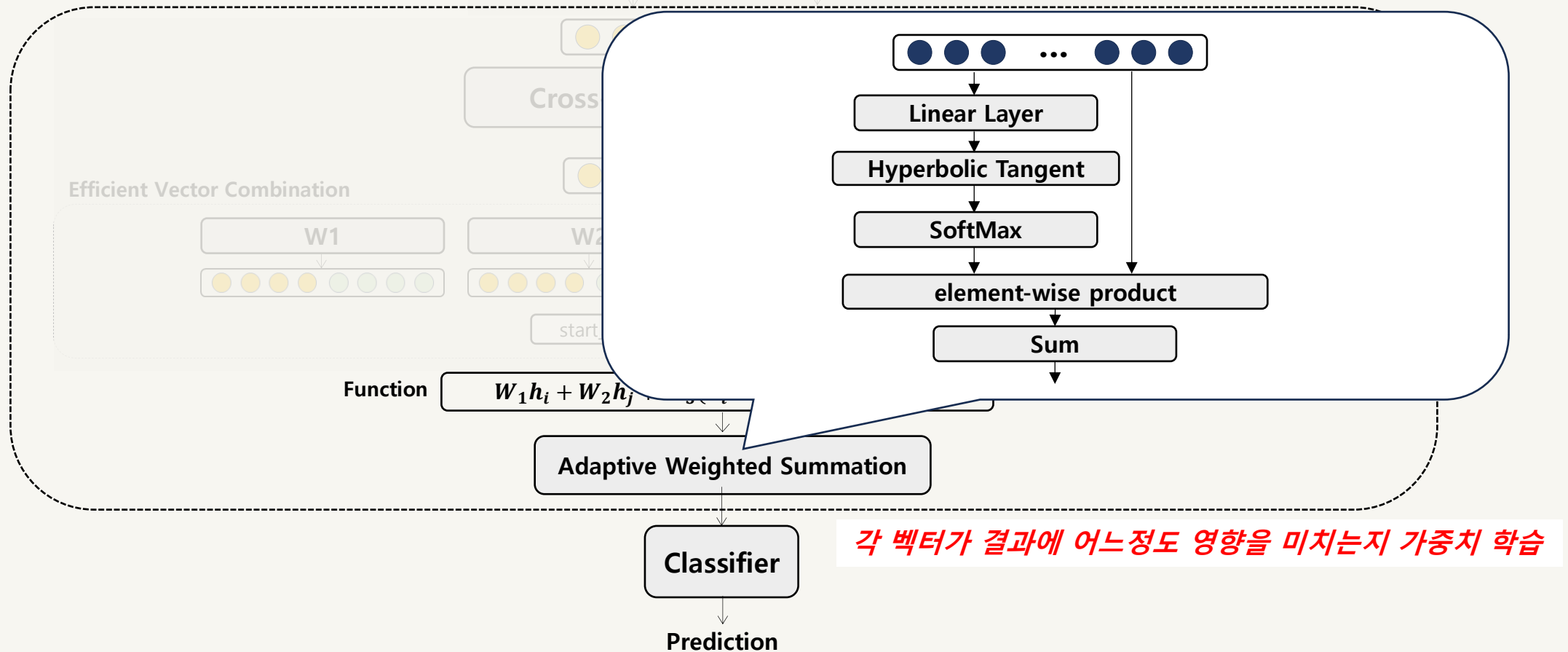
$$h(i, j) = F(h_i^k, h_j^k) = W[h_i, h_j, h_i - h_j, h_i * h_j]$$

$$= W_1 h_i + W_2 h_j + W_3 (h_i - h_j) + W_4 (h_i * h_j)$$

Unified Modal Ensemble Method



Unified Modal Ensemble Method



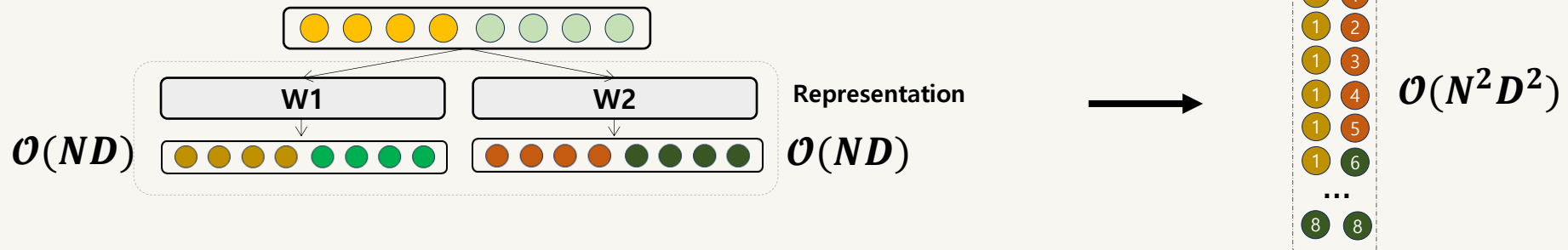
Computational Complexity

- Simple Vectors Pair Combine Method

- $$h(i, j) = F(h_i, h_j) = W_1 h_i * W_2 h_j$$

- x = input sequence $\{x_1, x_2, \dots, x_N\}$
- $h(i, j)$ = representaion된 $x(i)$ 와 $x(j)$ 의 조합 결과
- N = length of input sequence ($x = \{x_1, x_2, \dots, x_N\}$)
- D = Dimension(Hidden State Vector's depth)

- W = D – dimension matrix (layer)
- k = number of layers
- F = Feed Forward Network



재표현된 벡터간의 단순 쌍 조합은 $O(N^2 D^2)$ 의 계산복잡도를 가짐

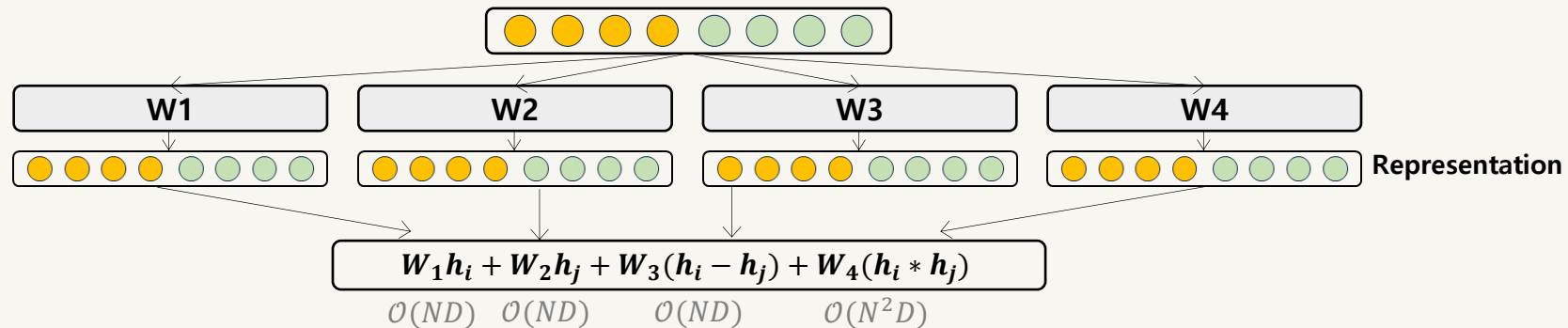
즉, 벡터가 길수록 막대한 컴퓨팅 자원 소모

Computational Complexity

Efficient Vector Combination

$$h(i, j) = F(h_i^k, h_j^k) = W[h_i, h_j, h_i - h_j, h_i * h_j] = W_1 h_i + W_2 h_j + W_3 (h_i - h_j) + W_4 (h_i * h_j)$$

- x = input sequence $\{x_1, x_2, \dots, x_N\}$
- $h(i, j)$ = representation된 $x(i)$ 와 $x(j)$ 의 조합 결과
- N = length of input sequence ($x = \{x_1, x_2, \dots, x_N\}$)
- D = Dimension(Hidden State Vector's depth)
- W = $D - \text{dimension matrix (layer)}$
- k = number of layers
- F = Feed Forward Network



계산 복잡도 감소

$$\mathcal{O}(N^2 D^2) \rightarrow \mathcal{O}(N^2 D)$$

$$* F(H_i, H_j) = \frac{2}{1 + 2^{-2W(H_i, H_j, H_i - H_j, H_i \circ H_j)}} - 1$$

Experiment

- Dataset

- VQA v2 (Visual Question Answering)

- 데이터 개수

	Train	Validation	Test
Images	82,783	40,540	81,434
Questions	443,757	214,354	447,793

- 성능지표

$$accuracy = \min\left(\frac{\text{humans that provided that answer}}{3}, 1\right)$$

- Example



Question	Answer (Label)
What color are the dishes?	id1: pink and yellow id2: yellow, pink ... id10: pink, yellow, and blue
How many cookies can be seen?	id1: 2 id2: 2 ... id10: 2
What is the green stuff?	id1: broccoli id2: broccoli ... id10: broccoli

Experiment

- **Dataset**
 - **VQA v2** (Visual Question Answering)
 - **Baseline**

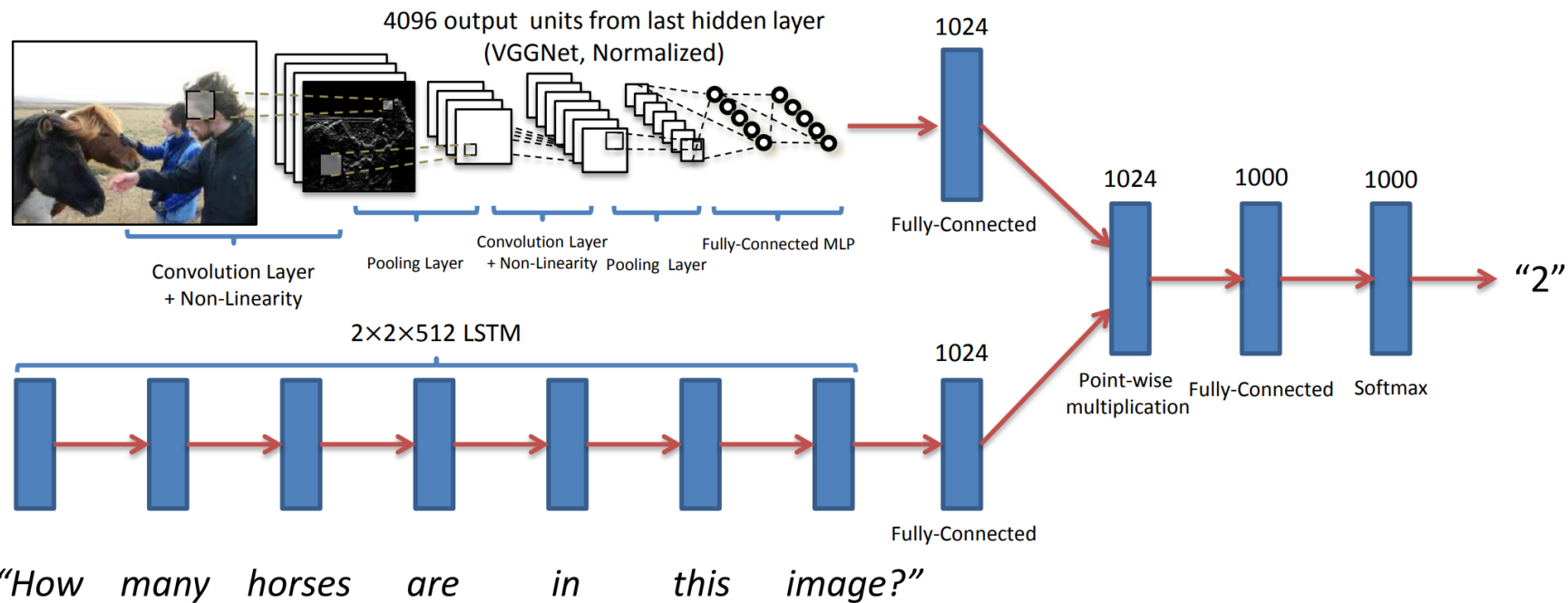


Fig. 8: Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [48] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers.

Experiment-1

- Hyperparameter**

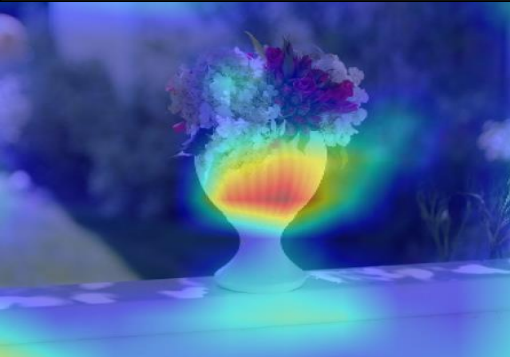
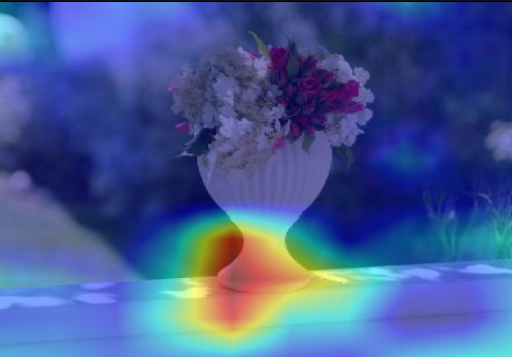
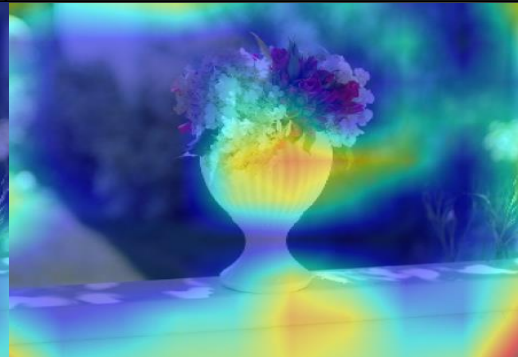
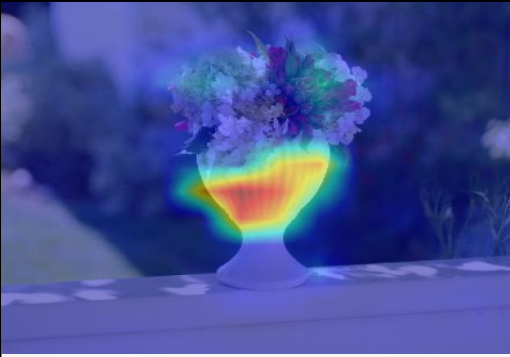
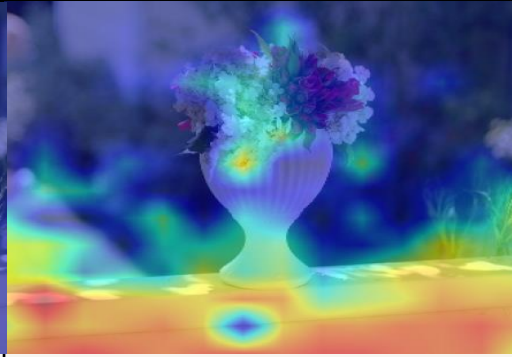
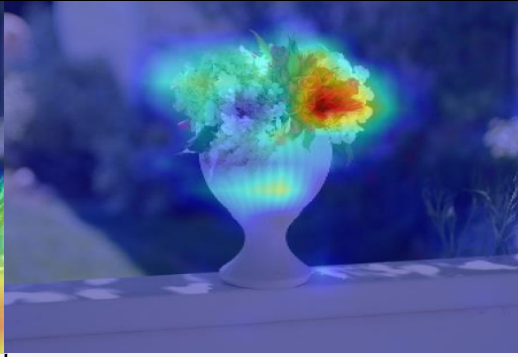
- Image_size: 384
- Learning Rate: $2e-5$
- Batch_size: 32
- Epoch: 10
- num_class: 1000(87.47%)
- Criterion: CrossEntropyLoss
- Optimizer: Adam

- Results**

Text Model	Image Model	Method	Accuracy
LSTM	VGGNet	Concatenation	0.35895 (± 0.0183)
LSTM	VGGNet	Element Wise Product	0.36382 (± 0.0157)
LSTM	VGGNet	Cross-Modal Attention	0.39814 (± 0.0102)
LSTM	VGGNet	+ Efficient Vector Combination	0.47285 (± 0.0092)
LSTM	VGGNet	+ Adaptive Weighted Summation (UME)	0.50481 (± 0.0194)

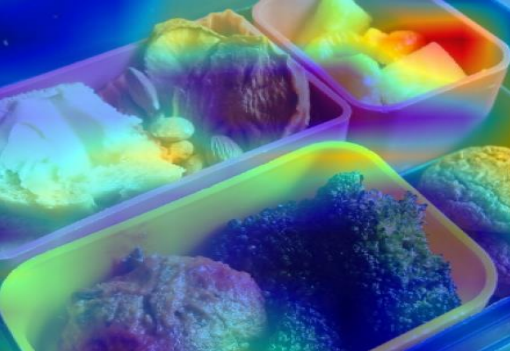
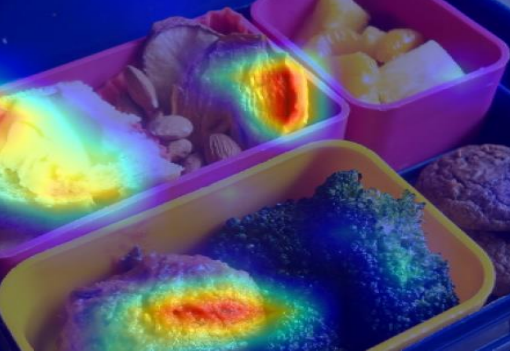
Experiment-2

- Visualization Analysis

LSTM+VGG	"what color is the vase?" <u>white</u>	"what is the vase sitting on?" <u>railing</u>	"are all the flowers white?" <u>no</u>
Baseline			
	<u>white</u>	table	<u>no</u>
UME Method			
	<u>white</u>	<u>railing</u>	<u>no</u>

Experiment-2

- Visualization Analysis

LSTM+VGG	"what color are the dishes?" <u>pink and yellow</u>	"how many cookies can be seen?" <u>2</u>	"what is the green stuff?" <u>broccoli</u>
Baseline	 <u>yellow</u>	 <u>3</u>	 <u>broccoli</u>
UME Method	 <u>pink and yellow</u>	 <u>3</u>	 <u>broccoli</u>

Experiment-3

- **Hyperparameter**

- Image_size: 384
- Learning Rate: 2e-5
- Batch_size: 16
- Epoch: 10
- num_class: 3128(93.25%)
- Criterion: CrossEntropyLoss
- Optimizer: AdamW

- **Results**

Text Model	Image Model	Method	Accuracy
BERT	ViT	Baseline	0.47186 (± 0.0093)
BERT	ViT	UME Method	0.55331 (± 0.0132)
RoBERTa	ViT	Baseline	0.49412 (± 0.0105)
RoBERTa	ViT	UME Method	0.57921 (± 0.0148)
DeBERTa_V3	BeiT	Baseline	0.58251 (± 0.0089)
DeBERTa_V3	BeiT	UME Method	0.67917 (± 0.0121)

Conclusion

- 단순 결합 대비 효율적인 계산 복잡도를 가진 조합 방안 제안
- 동일한 모델과의 비교를 통해 벡터 조합 방안의 효과 검증
- Grad-Cam 시각화를 통해 벡터 조합 방안의 효과 검증
- 모든 모델에 적용 가능한 일반성과 범용성 있는 조합 방안 제안

Future Works

- 3가지 이상의 유니모달 데이터를 사용한 작업으로 추가 검증
- 대형 멀티모달 모델(LMM) 사전학습 방식으로 채택하여 효과 검증

논문 추가 계획

- 레퍼런스 추가
- 실험 추가 - 언어모델 3개(BERT, RoBERTa, DeBERTa)와 비전모델 2개(ViT, BeiT) 전체 앙상블
- 실험 결과 추가 - 그래프 등 직관적으로 비교가능한 시각화 결과 추가
- 알고리즘 추가

Thank you

Intelligent Information Processing Lab
KiHoon Lee