Probabilistic Multimodal Learning with Bayesian Disagreements

Supplementary Material

In the supplementary section, we provide additional details about the method. We also provide additional details about the results.

1. Additional Details about the Methodology

1.1. Attention Reuse and Partial Sum algorithm

The algorithm to describe the Attention Reuse and Partial Sum Mechanism (Section 3.1.2) is described in Algorithm 1.

Algorithm 1 Attention Reuse and Partial Sum

Require: $Q, K, V \in \mathbb{R}^{n \times b \times d}$, number of modalities M_n

Given: $Q \in \mathbb{R}^{n \times b \times d}$, $K \in \mathbb{R}^{n \times b \times d}$, $V \in \mathbb{R}^{n \times b \times d}$ {n: batch-size, b: number of tokens, d: feature dimension}; M_n is the number of modalities

Define:

$M_I = \{0, 1, \ldots, M_n - 1\}$

$I_M = \{I_1, I_2, \ldots, I_{M_n}\}$ {I_i represents the indices of tokens for i-th modality}

$\mathcal{P} \leftarrow \mathcal{P}(I_M) \setminus \{\emptyset\}$ {Power set of I_M, excluding the empty set}

$\mathcal{P}_M \leftarrow \mathcal{P}(M_I) \setminus \{\emptyset\}$ {Power set of M_I, excluding the empty set}

Step 1: Full Attention Calculation

$\mathcal{A} \leftarrow \exp\left(\frac{Q \cdot K^\top}{\sqrt{d}}\right)$

Step 2: Initialize containers

$M \leftarrow [\,]$; $E \leftarrow [\,]$; $O \leftarrow [\,]$ {Array to store output of attention}

Step 3: Compute intermediate attention outputs

for each $i \in \{0, 1, \ldots, M_n - 1\}$ **do**

  $M[i] \leftarrow \mathcal{A}[:, I_M[i]] \cdot V[I_M[i], :]$

  $E[i] \leftarrow \sum_{j \in I_M[i]} \mathcal{A}[:, j]$

end for

Step 4: Reuse attention and compute output

for each $(I, J) \in (\mathcal{P}, \mathcal{P}_M)$ **do**

  O.append $\left(\frac{\sum_{i \in J} M[i]}{\sum_{i \in J} E[i]}\right)$ {Row-wise division}

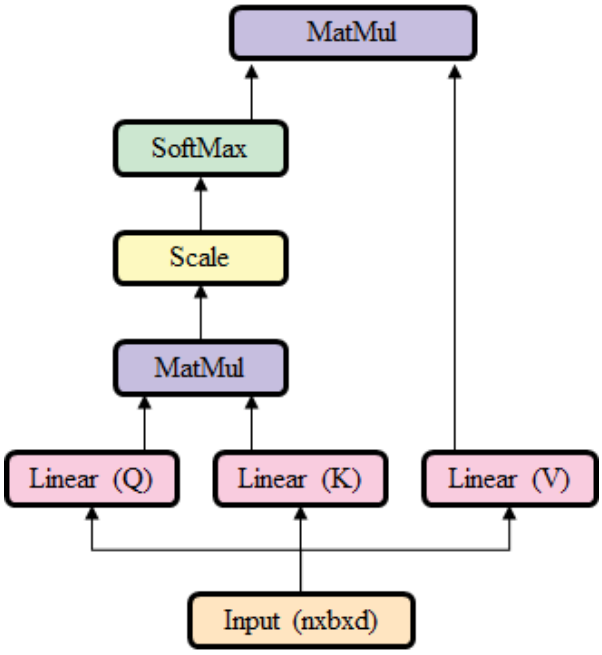
  $O[-1] \leftarrow O[-1][I, :]$ {Slice}

end for

Note: $\cdot$ operator represents matrix multiplication.

1.2. Probabilistic Attention

A visualization to show the difference between the Traditional Attention block and the proposed Probabilistic Attention block is shown in Figure 1. Figure 1(a) shows the traditional attention block and Figure 1(b) shows the proposed probabilistic attention block.



(a) Traditional Attention block

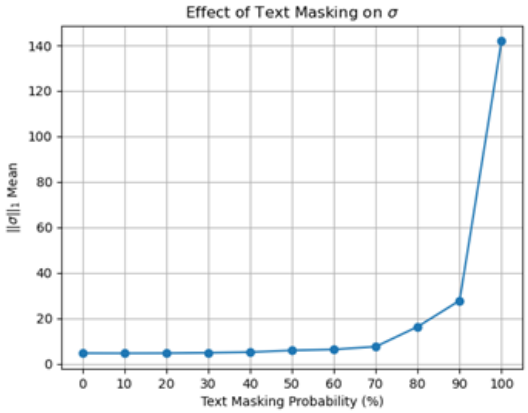
(b) Proposed Probabilistic Attention block

Figure 1. Comparison of our proposed probabilistic attention block and traditional attention. n is batch size, b is the number of tokens and d is the dimensionality.

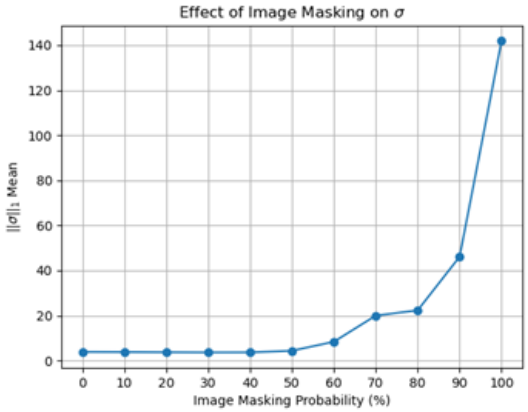
2. Additional Details about the Results

2.1. Effects of σ

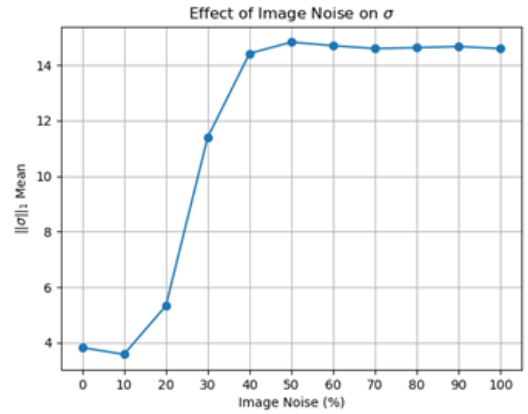
We analyze the sigma (standard deviation) value with increasing amounts of degradation in the input data (Figure 2). We experiment with three settings. First is increasing the text masking ratio (Figure 2(a)). Second is increasing the image masking ratio (Figure 2(b)). Lastly, we increasingly add Gaussian noise (Figure 2(c)). The



(a) Increasing text masking ratio



(b) Increasing image masking ratio



(c) Increasing noise in image

Figure 2. Sigma values captured by our PMBD model under various amounts of degradation severity.

X-axis has increasing amounts of degradation. In all cases, we can see that our model outputs sigma values that have approximately a monotonic behavior with the degradation severity. This implies that as the reliability decreases (that is, the degradation increases), the predicted sigma value also increases.

We also qualitatively visualize the sigma values outputted by our model using the PMBD-PA model (Figure 3). We can see that the informative regions of the image have lower sigma values. Lower sigma values are indicated by blue color. The text regions in the images usually show lower sigma values indicating a high information region.

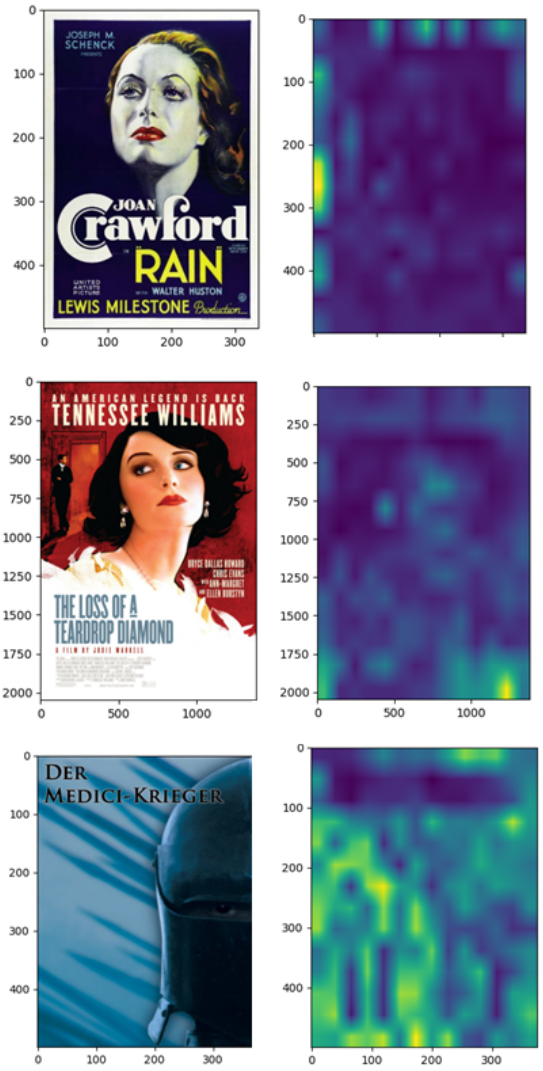


Figure 3. Qualitative visualizations of sigma values for various images in the MM-IMDB dataset.

2.2. Results on Epic-Kitchens dataset

In the paper in Figure 5, we show the results of different methods on the Action recognition task under noisy conditions. However, we only show the results on the

Action score. In this Supplementary Material, in Table 1, we show the results on the Verb, Noun and Action recognition under noisy conditions to validate that the trends in Action are similar to the trends in Verb and Noun recognition.

Table 1. Accuracy on Epic-Kitchens dataset. ‡ refers to supervised fine-tuning.

Method	**Verb**	**Noun**	**Action**
Background Noise SNR – 30			
CoMM‡	0.651	0.515	0.379
ProbCoMM‡	0.688	0.523	0.380
PMBD‡	0.711	0.568	0.403
PMBD-PA‡	0.723	0.577	0.412
Background Noise SNR – 20			
CoMM‡	0.512	0.451	0.281
ProbCoMM‡	0.515	0.459	0.293
PMBD‡	0.632	0.504	0.341
PMBD-PA‡	0.641	0.509	0.350
Background Noise SNR – 15			
CoMM‡	0.501	0.443	0.272
ProbCoMM‡	0.504	0.449	0.284
PMBD‡	0.628	0.500	0.336
PMBD-PA‡	0.638	0.502	0.347
Background Noise SNR – 10			
CoMM‡	0.496	0.439	0.267
ProbCoMM‡	0.498	0.440	0.278
PMBD‡	0.625	0.498	0.333
PMBD-PA‡	0.636	0.499	0.345
Gaussian Blur in Video			
CoMM‡	0.421	0.415	0.243
ProbCoMM‡	0.452	0.429	0.256
PMBD‡	0.495	0.478	0.272
PMBD-PA‡	0.500	0.486	0.280