A Supplemental Material

A.1 Proposition 1 for Eqn. 13 in the paper

Proposition 1: Suppose that we have n matrices, $I_1, I_2, ..., I_n$, and n vectors, $w_1, w_2, ..., w_n$. The space of I_l is $\mathbb{R}^{d \times t_l}$ and the space of w_l is $\mathbb{R}^{1 \times t_l}$. Then

$$\Big(\bigotimes_{i=1}^{n} I_i\Big) \circ \Big(\bigotimes_{i=1}^{n} w_i\Big) = \bigotimes_{i=1}^{n} I_i \circ w_i \tag{1}$$

Proof: We use C_l and C_r to denote the left side and right side of the equation, respectively. We utilize the element-wise comparison in two tensors. Following Eqn. 7 and 8 in the paper, the $(r_1, ..., r_n)$-th entry of C_l is expressed as

$$(C_l)_{r_1,...,r_n} = \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i)_{r_i}\Big] \cdot \Big[\bigwedge_{i=1}^{n} (w_i)_{r_i}\Big] \tag{2}$$

where $(I_i)_{r_i}$ is a vector which denotes r_i-th column of the I_i, $(w_i)_{r_i}$ is the r_i-th value of the vector. Since $(w_i)_{r_i}$ is a single element, we can directly multiply it with the corresponding vector $(I_i)_{r_i}$.

$$\begin{aligned}(C_l)_{r_1,...,r_n} &= \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i \circ w_i)_{r_i}\Big] \\ &= (C_r)_{r_1,...,r_n}\end{aligned} \tag{3}$$

The proposition is proven and is used to convert Eqn. 12 to Eqn. 13 in the paper.

A.2 Proposition 2 for Eqn. 15 in the paper

Proposition 2: Suppose that we have n matrices, $I_1, I_2, ..., I_n$. The space of I_l is $\mathbb{R}^{d \times t_l}$. Then

$$\sum \Big(\bigotimes_{i=1}^{n} I_i\Big) = \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i)\mathbf{1}_{t_i}\Big] \tag{4}$$

here we use vectors $\mathbf{1}_d$ and $\mathbf{1}_{t_i}$ which consist of 1 to represent the summation operation for matrix I_i in d dimension and t_i dimensions, respectively.

Proof: We use v_l to denote the left side of the equation and v_r to denote the right side of the equation. We can express v_l as

$$v_l = \sum_{r_1=1}^{t_1} ... \sum_{r_n=1}^{t_n} C_{r_1,r_2,...,r_n} \tag{5}$$

$$C = \bigotimes_{i=1}^{n} I_i \tag{6}$$

Following Eqn. 7 and 8 in the paper, we can express $C_{r_1,r_2,...,r_n}$ as

$$C_{r_1,r_2,...,r_n} = \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i)_{r_i}\Big] \tag{7}$$

We apply Eqn. 7 to Eqn. 5,

$$\begin{aligned}v_l &= \sum_{r_1=1}^{t_1} ... \sum_{r_n=1}^{t_n} \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i)_{r_i}\Big] \\ &= \mathbf{1}_d \Big[\sum_{r_1=1}^{t_1} ... \sum_{r_n=1}^{t_n} \bigwedge_{i=1}^{n} (I_i)_{r_i}\Big] \\ &= \mathbf{1}_d \Big[\bigwedge_{i=1}^{n} (I_i)\mathbf{1}_{t_i}\Big] = v_r\end{aligned} \tag{8}$$

The proposition is proven and is used to convert Eqn. 13 to Eqn. 15 in the paper.

A.3 Learning Curves

We show the learning curves of the CIDEr on the validation set in Fig. 1 and observe that the L-HOCA-UBT performs better than HOCA-UBT and HOCA-U when the training converges.

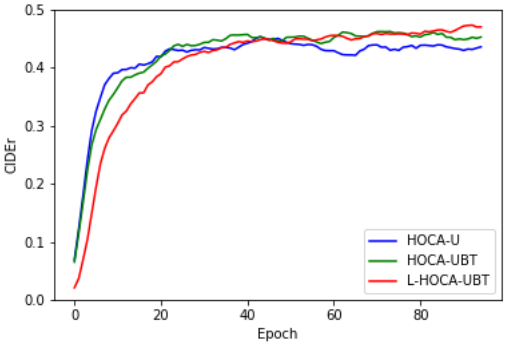

Figure 1: Learning curves of different methods on MSR-VTT, where the rank of L-HOCA-UBT is 1. Note that we use greedy search during training while beam search during testing, so the testing scores are higher.

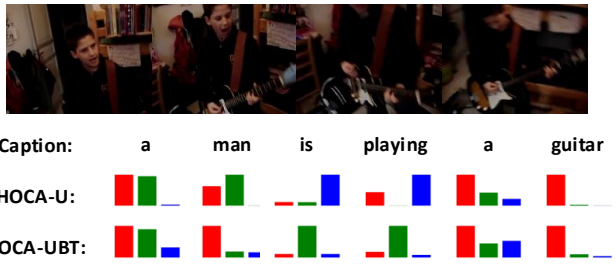

Figure 2: Visualization of the attention weights in multiple attentive fusion (MAF) module, the red bar denotes image modality, the green bar denotes motion modality, the blue bar denotes audio modality.

A.4 Visualization of Attention Weights

We also perform visualization of the attention weights in multiple attentive fusion (MAF) module. As shown in Fig. 2, HOCA-UBT obtains a

more accurate ratio of each modality than HOCA-U, i.e. for the word “man”, HOCA-U obtains a higher score of motion modality, which violates human subjective understanding.