MULTIMODAL SENTIMENT ANALYSIS ON UNALIGNED SEQUENCES VIA HOLOGRAPHIC EMBEDDING

Yukun Ma, Bin Ma

Alibaba Group
{yukun.ma,bin.ma}@alibaba-inc.com

ABSTRACT

Multimodal sentiment analysis is built on fusion of inputs from multiple modalities. However, at the core of existing fusion method is the dot product between a key vector and a query vector and relies on multiple neural network layers to model the high-order correlation. In this paper, we present a method based on holographic reduced representation which is a compressed version of the outer product to model facilitate higher-order fusion across multiple modality. Experiment shows that our proposal performs promisingly on benchmark multimodal sentiment analysis data sets with improved efficiency.

1. INTRODUCTION

The spoken language of human beings is naturally multimodal – the textual input, which was the mainstream of natural language processing, does not function alone to support a comprehensive spoken language understanding (SLU) system. It has drawn more and more attention to enrich the textual representation with information from additional modality such as human voices and vision. Voices, for example, might contain acoustic cues related with the speaker's emotion, while spoken language might be further complemented by the facial expression or body language being sighted. These characteristics of multimodality have been leveraged to benefit a wide range of different tasks including sentiment analysis[1, 2], speech recognition [3, 4], and gesture recognition[5].

As one of the main tasks of multimodal SLU, mutlimodal sentiment analysis targets at accurately perceiving the sentimental status of a speaker. The core issue faced by a multimodal sentiment analysis system is modeling of inter-modal and intra-modal dependency. As the inputs are typical streams of texts, audios or videos, capturing the temporal dependency is of most importance to understanding and analyzing the language. For example, a smirking face co-occurring with abnormal intonation and a positive phrase might indicate sarcasm or opposite sentiment. To address such issue, previous work have explored fusing information from the sequence level pre-computing uni-modal utterance and combining the representation at different phases [6, 7]. In addition, the cross-modal temporal dependency can be handled by temporal fusion [8, 9, 10] of modalities at each individual time step. The temporal dependency can be established by either hard or soft alignment [1, 11]. To model the correlation across modalities, one prevalent choice of fusion operation by existing work is the dot-product followed by layers of neural network that results in low parameter-efficiency.

Although it is noted that the fusion based on outer product produces improved performance over dot product based methods [6, 7], it is too costly to directly apply outer product in the unaligned sequences. In this paper, we propose to use holographic representation to model the interaction of multimodal sequences. As a compressed outer product, holographic representation reduces the dimension while retains the bit-wise interaction pattern at large and has demonstrated capable of encoding relations between objects [12]. We experiment with three different fusion stages to incorporate the holographic reduced representation (HRR) in a cross-modal transformer, and we provide experimental analysis to validate its effectiveness and efficiency. We demonstrate that the proposed method requires few neural network layers to account for high-order fusion which reduces the inference time and comparable (or even better) performance than the state-of-the-art methods on benchmark data sets.

2. RELATED WORK

2.1. Multimodal Sentiment

It has been shown that learning to represent the multimodal information via fusion of heterogeneous sources plays a critical role in multimodal analysis. Existing work has explored a wide range of possibilities for different fusion models. Various approaches es have been explored to fuse the multimodal information. Early works inject visual cues into the contexts while learning word embeddings [13]. In particular, it is popular to adopt multimodal analysis for sentiment or emotion recognition tasks [14, 15, 16, 6, 17, 7, 2, 11]. Of these methods, early fusion [15, 18] concatenates the representation across different multimodalities right after input layer. Since the input layers are mostly divided into tem-

poral steps, it needs to align the sequences across different modalities, which can be achieved by various ways including force alignment [16, 17, 8], contortionist temporal classification method [1]. Another line of work fuse the multimodal information at the decision phase. A set of sub-models are trained separately and then jointly predict the final label via majority voting or weighted averaging [19]. Intermediate fusion attempts to model both inter- and intra-modality dependency [6, 7]. These approaches have demonstrated the benefits brought by modeling high order interaction via outer product, but suffering from low efficiency as well as the explicit requirement for sequence alignment. To ease the requirement for explicit alignment, cross-modal attention models [1, 11] have been explored to ease these problems. These attention models the cross-modal dependency with a Query-Key attention matrix, but their approaches fail to inherit the high-order interaction pattern introduced by intermediate fusion approaches.

2.2. Holographic Embedding

The proposed method is based on holographic reduced representation (HRR). HRR has been successfully adopted together with neural networks in various settings. Early work of holographic recurrent network [20] uses HRR for the recurrent connection of a RNN. Later, HRR is further exploited for forming an associative memory [21] in LSTMs. Most recently, the work of holographic knowledge embedding[12] leverages the circular correlation of HRR to represent the relational tuples of knowledge graph, while in the community of recommend system, HRR is used as part of a factorization machine to model the item-item and user-item associative memory [22].

3. METHODOLOGY

We incorporate the HRRs into cross-modal transformers at three different fusion stages: "late fusion", "early fusion", and "intermediate fusion" from literature. One notable characteristic that might differ from literature is that "late fusion" does not refer to fusing the decisions resulted from a set of sub-optimal models. Instead, our late fusion merges the output of uni-modal transformer encoder before they are passed to the MLP for prediction, i.e., it is later than all the transformer layers. Similarly, the early fusion means the fusion happens right after the input layer while the intermediate fusion refers to fusions after the unimodal transformer but before the cross-modal transformer.

3.1. Holographic Reduced Embedding (HRR)

For a given pair of input sequences $X = \{x_1, x_2, x_3, ...x_n\}$, and a target modality sequence $Y = \{y_1, y_2, ...y_n\}$, resulted from different modalities, we compute the interaction of X

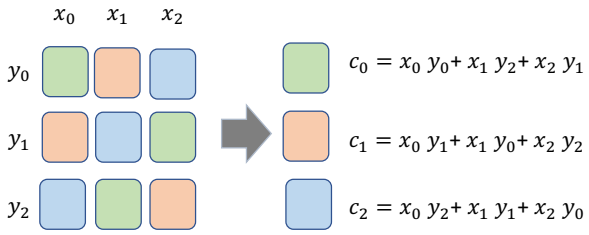

Fig. 1: Circular Convolution. Summation on the same color.

and Y as a tensor C as defined in Equation 1.

$$C_{ij} = x_i \circledast y_j \tag{1}$$

, each entry of C being obtained by taking circular convolution operator the corresponding elements in X and Y. The circular convolution operator was initially proposed as way to compute the holographic reduced representation (HRR) [23].

$$C_{ij}^k = [x \circledast y]_k = \sum_{i=0}^{d=1} x_i y_{(k-i) \bmod d}] \tag{2}$$

Noted that, compared with outer product,circular convolution does not introduce higher dimensionality into the model. The resulted representation has the same dimension as the original space of x and y. The computation of circular convolution and circular correlation can be implemented naively using brutal force iteration which results in exponential complexity. One efficient implementation is based on fast Fourier transformation and its inverse[20].

$$x \circledast y = \mathscr{F}^{-1}(\mathscr{F}(x) \odot \mathscr{F}(y)) \tag{3}$$

Equation 3 illustrates the fast Fourier transformation as an efficient implementation of circular convolution, $\odot$ denote the Hadamart product. With fast Fourier transformation, the operator can be implemented in quasi-linear complexity. Note that, since $x \circledast y = y \circledast x$, circular convolution can be used to represent symmetric relations which is particularly suitable to represent the symmetric relation between cross-modal objects. In contrast, the other alternative, circular convolution is non-commutative so that it can be used to represent the asymmetric association.

3.2. Crossmodal HRR

Each pair of x and y is now represented by HRR as a form of associated memory. To represent an arbitrary element of a given modality with its counterpart from the other modality,

we could simply calculate the weighted sum over the columns or rows of C. For example,

$$\hat{x}_i = \sum_j \beta_{ij} C_{ij} \tag{4}$$

, where the $beta_i$ is an attention weight assigned to y_i given x_i. We design a co-attention matrix, denoted as B, for computing the weights between two modalities. Each entry of B is defined as in Equation 5

$$B_{ij} = W(x_i \circ y_j) \tag{5}$$

where $\circ$ denotes concatenation by the last dimension. Note that, here we assume each pair of x and y has the same dimensionality. Once we have computed the co-attention matrix B, we can obtain $\beta_{i,j}$ by computing the softmax of B_i.

As the circular convolution is symmetric, the same matrix C and co-attention matrix B could be reused by both modalities for computational efficiency. The associative memory between two modality can be decoded (as shown in Equation6) to reconstruct the set of items by an operation called circular correlation[20], denoted as $*$.

$$x_i * (x_i \circledast y_j) \approx y_j \tag{6}$$

The decoding can also be achieved for a set of x_i and y_j association as in Equation 7

$$x_i * \sum_{x_i, y_j} (x_i \circledast y_j) \approx y_j \tag{7}$$

3.3. Cross-modal Transformer with HRR

Note that each $\hat{x_i}$ will be later passed to the attention model of transformer. and the linear transformation will be applied to the weighted sum of $\hat{x_i}$. Now, let us examine these operation closely. Assuming the input query for a given X is y_q at the moment, the output from attention model (we assume single head for simplicity) is

$$\hat{X_q} = \sum_j \alpha_i^q \hat{x_i} \tag{8}$$

Namely, the output is a weighted sum of the associated memory. By applying a linear transformation is to have

$$\hat{X_q} = \sum_i \alpha_i^q h \hat{x_i} = h \sum_i \sum_j \alpha_i^q \beta_j (x_i \circledast y_j) \tag{9}$$

,where the linear transformation matrix h is learnable through training process. As pointed in literature[22], the learning of h is be simulating the decoding process of circular correlation [20]. Since the h can be deemed combined with a positional encoding, this forms a position-sensitive decoding of the associated memory, which is desirable in recovering the "entities" (e.g., visual objects, textual words, or audible events). Namely, it will retrieve the y_j with the highest association strength regarding a relation specified by h.

To incorporate HRRs with the input or output of transformer layer. We designed the three fusion methods as discussed in the beginning of this section. However, it should be noted that, for early fusion and intermediate fusion, cross-modal HRRs are concatenated with the input to the subsequent transformer layer. Let X denote unimodal embedding (from either input layer or uni-modal transformer encoder) of an interested modality, and Y and Z being the uni-modal embeddings of the other two modalities. The fused result can be denoted as $X \circ (X \circledast Y) \circ (X \circledast Z)$.

The fused input for Y and Z can be computed in the same way. We make late fusion as an exception by taking a ultra-late fusion on only the outputs of the classification tokens of each modalities. Namely, the output of late fusion is defined as $x_c \circ y_c \circ z_c \circ (x_c \circledast y_c) \circ (y_c \circledast z_c) \circ (z_c \circledast x_c))$.

4. EXPERIMENTS

4.1. Data Sets

Following the experiment protocol set up by previous work[1], we evaluate the proposed method as well as state of the art baselines on two benchmark data sets: CMU MOSI [14], and CMU MOSEI [24]. All the three datasets consist of texts, audio, and videos. To extract the textual features, GLOVE word embedding [1] is used. Acoustic features are extracted from COVAREP while visual features are extracted by Facet[2]. Note that, as the main focus of this paper is on unaligned sequences, all sequences are not aligned. The two CMU data sets (i.e., MOSI and MOSEI) are both created by CMU researchers in similar manner. MOSI data set contains 2,199 video clips of human monologues. Each utterance in these clips are manually labeled with a sentiment strength ranging from -3 to 3 with 3 being the most strongly positive and -3 being the most negative. We simply round the continuous value to get the discrete labels of 7 classes. CMU-MOSEI data set is much larger in size, consisting of 23,454 YouTube video clips on topics of movie review. We follow the original split of the data sets, and tune the hyperparameters on the validation set.

On these two data sets, 7-class and binary accuracy (Acc[7] and Acc[2]) as well as the F-score are reported based on discretizing the continuous sentiment score. Also, mean absolute error (MAE) along with the correlation of predication and gold truth score are also included as part of the experiment protocol.

[1] https://nlp.stanford.edu/data/glove.840B.300d.zip

[2] https://imotions.com/

4.2. Baselines

We retain the three baselines based on connectionist-temporal-classification (CTC) alignment: early fusion LSTM (EF-LSTM), Recurrent Attended Variation Embedding Network (RAVEN) [9] and multimodal cyclic translation network (MCTN) [10]. In addition, we compare with the multimodal transformer (MULT)[1] as well as the recently proposed multimodal temporal graph attention network (MTGAT)[11]. Note that, we have attempted to re-run the experiments of MULT and MTGAT. We follow the suggested best-performed setting reported in their paper. For MULT, the implementation is the same as being released by the author while we do re-implement the MTGAT following the instructions in the original paper. We refer to our proposed approach as holographic embedding multimodal transformer (HEMT). We add three prefixes: early fusion (EF), late fusion (LF), and intermediate fusion (IF) to indicate the way of fusion.

4.3. Analysis

We first evaluate our proposed model on two CMU data sets. The results are shown in Table 1 and Tabel 2. On the MOSI data set, it can be seen that our proposed methods outperform the MTGAT model on all the metrics except MAE. Similar results have also been observed on MOSEI data set which is larger in size. Since our method largely reused the multimodal transformer architecture, the results suggests the holographic representation could effective replace the cross-modal encoder. It captures the high-order dependency without recursively computing attention matrices. On this regression task, the proposed early fusion has demonstrate higher performance than the other two models. But, in our experiment, we found even the other two HRR fusion methods perform closely as compared with our own run MULT baselines.

Metric	Acc_7	Acc_2	F1	MAE	Corr
CTC+EF-LSTM	31.0	73.6	74.5	1.078	0.542
CTC+MCTN	32.7	75.9	76.4	0.991	0.613
CTC+RAVEN	31.7	72.7	73.1	1.076	0.544
MULT	39.1	81.1	81.0	0.889	0.689
MULT(our run)	31.6	75.3	75.3	1.083	0.580
MTGAT	37.2	81.5	81.7	**0.881**	0.701
LF-HEMT	34.4	79.2	79.4	0.951	0.673
IF-HEMT	34.4	79.8	80.2	1.028	0.646
EF-HEMT	**39.2**	**82.3**	**82.5**	0.901	**0.701**

Table 1: Comparison of SOTA methods on CMU MOSI data set.

Since HRR captures higher order correlation via circular convolution, it, to some extent, relieves the requirement for recursively infer the dependency. During the training and evaluating process, we also have observed an improvement of computational efficiency due to the removal of multiple cross-modal attention layers. In order to compare the inference efficiency, we report the inference time measured for models that achieving best performance in Table 1, Table 2. We use Tesla V100-SXM2 GPU and set the batch size to 16. We run the experiment for 10 times each and take the average inference time for fair comparison. Table 3 shows the inference time(ms/batch) on the three data sets. It shows that, to process the same amount of data, the proposed fusion methods requires only 1/3 of the time required by MULT. On CMU data sets, the proposed EF-HEMT model achieces higher performance while consuming only 1/2 the time required by MTGAT. As compared wit h MULT, LF-HEMT and EF-HEMT used only 1/3 of the number of transformer layers to achieve a comparable performance while IF-HEMT have to used 1/2 of the number of layers.

Metric	Acc_7	Acc_2	F1	MAE	Corr
CTC+EF-LSTM	46.3	76.1	75.9	0.680	0.585
CTC+MCTN	48.2	79.3	79.7	0.631	0.645
CTC+RAVEN	45.5	75.4	75/7	0.664	0.599
MULT	50.7	81.6	81.6	**0.591**	0.694
MULT(our run)	49.9	81.8	81.6	0.608	0.680
MTGAT(our run)	49.2	79.9	80.3	0.649	0.638
IF-HEMT	49.4	81.7	81.8	0.613	0.675
LF-HEMT	49.8	81.6	81.6	0.610	0.685
EF-HEMT	**51.2**	**81.9**	**82.15**	0.597	**0.699**

Table 2: Comparison of SOTA methods on CMU MOSEI data set.

Model	MOSI	MOSEI
MULT	1503	1497
MTGAT	809	930
IF-HEMT	752	644
LF-HEMT	452	433
EF-HEMT	492	495

Table 3: The inference time (ms/batch) of all models achieving comparable performance (as reported in Table 2 and 1) with batch size=16 and 2 GPUs in use.

5. CONCLUSION

In this paper, we introduce holographic reduced representation for modeling cross-modal dependency. The proposed cross modal HRR creates associative memories of pairs of inputs from different modalities. We demonstrate that the proposed method could effectively capture high-order dependency and ease the needs of high-order attention model which results in significant reduction of computing time. We demonstrate that the proposed method could achieve comparable or even superior performance as compared with state-of-the-art baselines.

6. REFERENCES

[1] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 2019 ACL*, vol. 2019, 2019, p. 6558.

[2] Z. Pan, Z. Luo, J. Yang, and H. Li, "Multi-modal attention for speech emotion recognition," *Proc. Interspeech 2020*, pp. 364–368, 2020.

[3] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *2015 ICASSP*, 2015, pp. 2130–2134.

[4] S. Palaskar, R. Sanabria, and F. Metze, "End-to-end multimodal speech recognition," in *2018 ICASSP*, 2018, pp. 5774–5778.

[5] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: multimodal transfer module for cnn fusion," in *Proceedings of the ICCV 2020*, 2020, pp. 13 289–13 299.

[6] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017.

[7] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 2018 ACL*, 2018, pp. 2247–2256.

[8] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv:1806.06176*, 2018.

[9] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the 2019 AAAI*, vol. 33, no. 01, 2019, pp. 7216–7223.

[10] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6892–6899.

[11] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, and L.-P. Morency, "Mtgat: Multimodal temporal graph attention networks for unaligned human multimodal language sequences," *arXiv preprint arXiv:2010.11985*, 2020.

[12] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proceedings of the 2016 AAAI*, vol. 30, no. 1, 2016.

[13] A. Lazaridou, M. Baroni *et al.*, "Combining language and vision with a multimodal skip-gram model," in *HLT-NAACL*, 2015.

[14] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[15] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 ICDM*, 2016, pp. 439–448.

[16] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[17] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the 2018 AAAI*, vol. 32, no. 1, 2018.

[18] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *2017 ICME*. IEEE, 2017, pp. 949–954.

[19] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency, "Deep multimodal fusion for persuasiveness prediction," in *Proceedings of the 2016 ACM MM*, 2016, pp. 284–288.

[20] T. A. Plate, "Holographic recurrent networks," *NIPS*, pp. 34–34, 1993.

[21] I. Danihelka, G. Wayne, B. Uria, N. Kalchbrenner, and A. Graves, "Associative long short-term memory," in *International Conference on Machine Learning*, 2016, pp. 1986–1994.

[22] Y. Tay, S. Zhang, A. T. Luu, S. C. Hui, L. Yao, and T. D. Q. Vinh, "Holographic factorization machines for recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5143–5150.

[23] T. A. Plate, "Holographic reduced representations," *IEEE Transactions on Neural networks*, pp. 623–641, 1995.

[24] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 2018 ACL*, 2018, pp. 2236–2246.