An Entanglement-driven Fusion Neural Network for Video Sentiment Analysis

Dimitris Gkoumas[1,2*] , **Qiuchi Li**[3] , **Yijun Yu**[1] and **Dawei Song**[1,4†]
[1]The Open University, Milton Keynes, UK
[2]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[3]University of Padua, Padua, Italy
[4]Beijing Institute of Technology, Beijing, China
{dimitris.gkoumas, yijun.yu, dawei.song}@open.ac.uk, qiuchili@dei.unipd.it

Abstract

Video data is multimodal in its nature, where an utterance can involve linguistic, visual and acoustic information. Therefore, a key challenge for video sentiment analysis is how to combine different modalities for sentiment recognition effectively. The latest neural network approaches achieve state-of-the-art performance, but they neglect to a large degree of how humans understand and reason about sentiment states. By contrast, recent advances in quantum probabilistic neural models have achieved comparable performance to the state-of-the-art, yet with better transparency and increased level of interpretability. However, the existing quantum-inspired models treat quantum states as either a classical mixture or as a separable tensor product across modalities, without triggering their interactions in a way that they are correlated or *non-separable* (i.e., *entangled*). This means that the current models have not fully exploited the expressive power of quantum probabilities. To fill this gap, we propose a transparent quantum probabilistic neural model. The model induces different modalities to interact in such a way that they may not be separable, encoding crossmodal information in the form of non-classical correlations. Comprehensive evaluation on two benchmarking datasets for video sentiment analysis shows that the model achieves significant performance improvement. We also show that the degree of non-separability between modalities optimizes the post-hoc interpretability.

1 Introduction

Video sentiment analysis is an emerging interdisciplinary area in multimedia information processing, bringing together Artificial Intelligence (AI) and cognitive science. In particular, it studies a speaker's sentiment expressed by verbal (i.e., linguistic) and non-verbal (i.e., visual, acoustic) streams. At its core, this research area focuses on modelling interactions among distinct modalities, a.k.a., crossmodal dynamics. In recent years, research has made significant strides towards the inference of human-like sentiment judgments. In particular, neural approaches have been investigated to model crossmodal interactions by operating primarily feature-level fusion [Tsai *et al.*, 2019; Liu *et al.*, 2018; Zadeh *et al.*, 2017]. However, current models neglect other important aspects, such as model transparency, post-hoc interpretability, and how people would understand and reason about sentiment states.

The modelling of distinct modalities for sentiment analysis is a challenging problem. This is due to the spectrum of sentiment polarities that an utterance can emerge (e.g., positive, neutral, or negative), depending on the context of individual modalities. For instance, let us consider that we want to identify the binary sentiment of the utterance "Well, what a surprise!". The utterance's sentiment state is ambiguous, i.e., it could be either positive or negative, since there is no specific context to bias the sentiment state of the utterance. However, once the context (i.e., other modalities) is informative, unambiguous, and simultaneously present, the sentiment state to a particular sentiment judgment becomes apparent. This implies that we shall not consider distinct modalities in isolation. Rather, we must model modalities as *non-separable*, called *entangled*. Quantum Theory (QT) is the only theory which models non-separability. Thus, there is a reason to suppose that QT provides an adequate theory to capture the crossmodal correlations and how such correlations influence the final decision about the utterance's sentiment.

Quantum probability theory has been extensively studied in the domain of human cognition and decision making [Busemeyer and Bruza, 2012]. In particular, it has been shown that in some cases human language understanding exhibits certain non-classical properties [Bruza *et al.*, 2008; Bruza *et al.*, 2009], enabling quantum probabilities a suitable framework. Recent advances in quantum probabilistic neural models [Li *et al.*, 2019; Li *et al.*, 2021] have demonstrated improved performance and high-level explainability due to their theoretical root on the well-established quantum physics meanings. Nevertheless, they treated the interactions among quantum states as either a classical mixture of states [Li *et al.*, 2019] or a separable product of states [Li *et al.*, 2021], which cannot fully exploit the potentials of quantum probabilities in modelling the non-separability of multiple modalities. The expressiveness of quantum probabilities goes beyond classi-

[*] Contact Author

[†] Contact Author

cal correlations.

In line with the above observation, we propose a quantum probabilistic neural network, which captures non-classical correlations of sememes across distinct modalities. In particular, we transform the real-valued input features of different modalities into pure quantum states of complex values. The particular way in which the modality states interact with each other allows modelling of both classical correlation and non-separability (entanglement) between modalities in a unified framework. Our work is fundamentally different from the previous probabilistic neural network approaches in that we address the issues of contextuality, i.e., a modality activates ambiguous sentiment polarities in the context of other modalities, and non-separability, i.e., a modality cannot be separated from the rest of modalities occurred concurrently. The proposed model is empirically evaluated on two benchmark datasets for video sentiment analysis. Experimental results show that our model significantly improves performance over a wide variety of state-of-the-art (SOTA) approaches. We also show that the degree of non-separability of entangled states can be used to improve the post-hoc interpretability.

2 Related Work

The application of quantum theory in representation learning began after van Rijsbergen's pioneering work [Van Rijsbergen, 2004] by integrating geometric spaces, probabilities, and logic into a unified theoretical framework. Then, the probabilities of quantum theory were exploited for various representation learning tasks [Uprety *et al.*, 2020]. Among them, quantum formalism was successfully utilised for modelling word dependencies through density matrices [Sordoni *et al.*, 2013] and formulating the semantic composition of words [Sordoni *et al.*, 2013] in IR tasks. Quantum-inspired models were also introduced to address NLP tasks. Recently, the deployment of quantum measurement into a joint complex-valued neural network led to improved performance and better interpretability [Li *et al.*, 2019].

Quantum-inspired strategies were also investigated for multimodal representation learning. Early work exploited a tensor-based representation for an image-text IR task [Wang *et al.*, 2010]. Preliminary work investigated non-classical correlations, i.e., entanglement, based on the combination of uni-modal decisions [Gkoumas *et al.*, 2018]. The notion of quantum incompatibility has been exploited to fuse decisions from different modalities for video sentiment analysis [Gkoumas *et al.*, 2021a]. Recently, a quantum-inspired network for videos sentiment analysis exploits an early fusion of tri-modals via the tensor product of modalities [Li *et al.*, 2021]. However, the interactions across different modalities are largely neglected, assuming the interactions are decomposable. Instead, inter-modal interactions are implemented in the sentiment decision process. Our model is fundamentally different from [Li *et al.*, 2021] in that we take a quantum-cognitively motivated view on the non-decomposability of cross-modality interactions, which is modelled as quantum entanglement.

3 Preliminaries on Quantum Theory

In this section, we present fundamental concepts of quantum theory (QT) [Melucci, 2015; Busemeyer and Bruza, 2012] that we exploit to construct the quantum probabilistic neural model. In consistency with the convention of QT, we adopt the widely-used *Dirac Notations*, known as "bra-ket" notation. A complex-valued *unit* vector $\vec{u}$ and its conjugate transpose $\vec{u}^{*T}$ are denoted as a *ket* $|u\rangle$ and a *bra* $\langle u|$, respectively. The inner product of two vectors $|u\rangle$ and $|v\rangle$ is defined by $\langle u|v\rangle$, while $|u\rangle\langle u|$ and $|v\rangle\langle v|$ define operators.

The starting point to modelling quantum states is a set of basis states. A basis is a set of n mutually orthogonal vectors $\{|e_j\rangle\}_{j=1}^n$ of unit length. The vector space employed in QT is a vector space over complex numbers, called Hilbert space $\mathbb{H}$, offering the structure of an inner product to enable the measurement of angles and lengths [Halmos, 1987].

Any *pure state* $|s\rangle$ of a quantum system is regarded as a linear *superposition*, i.e., an appropriate weighted sum, of one set of n basis states, represented by a unit vector in $\mathbb{H}^n$. That is, the pure state $|s\rangle$ can be written as a probability distribution of complex probability amplitudes, as follows:

$$|s\rangle = \sum_{j=1}^{n} \sqrt{r_j} e^{i\phi_j} |e_j\rangle \,, \tag{1}$$

where $\sqrt{r_j} e^{i\phi_j}$ correspond to complex probability amplitudes, r_j are non-negative scalars $\in \mathbb{R}$ satisfying $\sum_{j=1}^n |r_j| = 1$, i is imaginary number, and ϕ_j are phases $\in [0, 2\pi]$.

Quantum *measurement* with respect to a basis (i.e., a set of eigenvectors), yields one out of all the observable eigenvalues and causes the *collapse* of the state to the corresponding eigenvector. The probability of a given outcome is obtained via the projection postulate. That is, according to the Born's rule [Halmos, 1987], the probability of the pure state $|s\rangle$ to collapse onto the basis state $|e_j\rangle$ is calculated by the inner product of the two vectors as follows:

$$P(s|e_j) = |\langle e_j|s\rangle|^2. \tag{2}$$

A composite quantum system describes a compound system composed of multiple individual quantum systems. For two n_A-dimensional and n_B-dimensional spaces, the state vector $s_c \in A \otimes B$ is expressed as a linear combination of an arbitrary basis of the product space $|i\rangle \otimes |j\rangle$, as follows:

$$|s_c\rangle = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \sqrt{c_{i,j}} \, |i\rangle \otimes |j\rangle \,, \tag{3}$$

where $\sqrt{c_{i,j}}$ are probability amplitudes so that $\sum_{i,j} c_{i,j} = 1$.

When a composite quantum system evolves under a Hamiltonian that includes interactions between individual subsystems, the resulting state of the composite system is no longer *separable*, a.k.a. *entanglement* [Melucci, 2015]. That is, Eq. 3 cannot be expressed as a tensor product of individual systems. The bipartite Von Neumann *entanglement entropy* [Vivo *et al.*, 2016] is a measurement of the degree of quantum entanglement for a composite pure state. For instance, for an arbitrary bipartite quantum state consisting of two sub-states, i.e., $A \in \mathbb{H}^{n_A}$ and $B \in \mathbb{H}^{n_B}$, the Von Neumann entropy is calculated as follows:

$$S = S_A = S_B = -\sum_{i}^{min(\mathbb{H}^{n_A},\mathbb{H}^{n_B})} |a_i| log(|a_i|), \quad (4)$$

where a_i are singular values of the Schmidt decomposition of the bipartite quantum state over the sub-system A or B. Eq. 4 makes clear that the entanglement entropy is the same regardless of whether the decomposition is over A or B sub-system. Crucially, if the entropy S is zero, there is no entanglement.

4 Task Formulation

The goal is to infer the sentiment of video utterances. The multimodal dataset consists of N labeled video utterances $U = (U_1, ..., U_i, ..., U_N)$. Each utterance U_i is associated with linguistic, visual, and acoustic features, denoted as $U_i = (U_i^l, U_i^v, U_i^a)$. The corresponding labels for the N utterances are denoted as $y = (y_1, ..., y_i, ..., y_N)$, $y_i \in \mathbb{R}$. Essentially, the objective is to establish a function, mapping each video utterance U_i to its corresponding sentiment label y_i.

5 Entanglement-driven Fusion Neural Network

We propose a quantum probabilistic neural model, namely, Entanglement-driven Fusion Neural Network (EFNN).

5.1 General Architecture

The architecture of EFNN is illustrated in Fig. 1. In particular, we deploy an intermediate fusion strategy into a neural network modelling paradigm that incorporates QT-inspired complex representation of information, composite quantum system and entanglement to capture the non-separability of modalities, and quantum measurement for abstract sentiment concept extraction.

Specifically, EFNN first takes multimodal information, i.e., linguistic, visual, and acoustic, and feed it into three separated neural branches, one for each modality (see Fig. 1). At the outset, multimodal information is projected into a common-dimensional space, and then a preparation step converts the information to its quantum analogues, i.e., quantum states. Afterwards, we operate a pairwise fusion of modalities, i.e., *linguistic-visual*, *linguistic-acoustic*, and *visual-acoustic*, via the tensor product of bi-modals (any of two modalities). A weight vector captures the correlations within the bi-modal tensor-based representations (see Fig. 1, Entangled Bi-Modals step). A set of parameterized measurements map the complex-valued representation to a real-valued high-level representation via the quantum measurement postulate. Then, a row-max pooling operator is applied, followed by a fully connected layer passed to a *softmax* function for classification. In the remaining part of the section, we elaborate on the methodology according to the above procedural steps.

5.2 Preparation of States

In this work, each utterance is modelled as a uni-modal pure quantum state into modality-specific Hilbert spaces $\mathbb{H}_m$, where $m \in \{l, v, a\}$, for linguistic, visual and audio modalities. In line with previous works [Li *et al.*, 2019], we consider the exponential form of complex numbers to express quantum states: $z = re^{i\theta}$, where amplitude r is a real non-negative coefficient, phase $\theta \in [0, 2\pi)$, and i is imaginary number satisfying $i^2 = -1$.

Then, according to Eq. 1, the modality-specific pure state of an utterance $|u_m\rangle$ could generally be expressed by the following modulus-augment form:

$$\begin{aligned} |u_m\rangle &= [r_{1,m}e^{i\theta_{1,m}}, r_{2,m}e^{i\theta_{2,m}}, ..., r_{d,m}e^{i\theta_{d,m}}]^T \\ &= [r_{1,m}, r_{2,m}, ..., r_{d,m}]^T \odot e^{i[\theta_{1,m}, \theta_{2,m}, ..., \theta_{d,m}]^T} \end{aligned} \quad (5)$$

where d is the dimension of modality features and $\odot$ refers to element-wise vector product. In the modulus-argument form, any operation on the complex numbers will lead to a non-linear combination of the constituent moduli and arguments. This implies that a non-linear feature combination is inherently produced when we assign Eq. 5 with linguistic, visual, and acoustic features.

In Eq. 5, the first vector, i.e., $r_m = [r_{1,m}, r_{2,m}, ..., r_{d,m}]^T$, corresponds to amplitudes, where the moduli r is a real-valued vector of unit length. To construct amplitudes, we transform the input real-valued features to their quantum analogues as follows. Suppose the input word-level features are $X^l \in \mathbb{R}^{L \times d_l}$, $X^v \in \mathbb{R}^{L \times d_v}$, $X^a \in \mathbb{R}^{L \times d_a}$, where d_l, d_v, d_a represents feature dimensions for linguistic, visual, and acoustic modalities respectively, and L is the sequence length, i.e., total number of words in an utterance. At the outset, we project the input features into the same dimension d via convolutional neural networks [Lai *et al.*, 2015] from the respective input features with Rectified Linear Unit (ReLU) as the activation function in the last hidden layer, to ensure all elements $\{r_{i,m}\}_{i=1}^d$ are non-negative: $\hat{m} = ReLU(CNN_m(X^m)) \in \mathbb{R}^d$, where $m \in \{l, v, a\}$. Despite the projection of modalities into a common-dimensional space, the convolutional neural networks CNN_m capture local structure of words in an utterance. Then, we normalize the outputs to create vectors of unit length: $r_m = \frac{\hat{m}}{||\hat{m}||^2}$.

The second vector, i.e., $\theta = [\theta_{1,m}, \theta_{2,m}, ..., \theta_{d,m}]^T$, is also real-valued, with all its elements in $[0, 2\pi]$. The assignment of the phases θ is an open research question. In this work, to enable each utterance to carry *temporal information*, i.e., we assign the position of words in a sentence to the phase part. With this way, we capture the global structure of words in an utterance. The phase θ is hence calculated by

$$\theta = \theta(k) = f_{pe}(k), \quad (6)$$

where $f_{pe}(k)$ defines a map $f_{pe} : \mathbb{N} \to \mathbb{R}^d$ from a discrete position index to a $d-$dimensional real-valued vector.

5.3 Entanglement-driven Modality Fusion

After the transformation of feature inputs to quantum states into uni-modal Hilbert spaces, we feed them into the modality fusion component (see Fig. 1). In particular, we deploy a fusion module, which takes the utterance states of *pairwise* modalities, i.e., *linguistic-visual*, *linguistic-acoustic*, *visual-acoustic*. For each pairwise of states, a composite

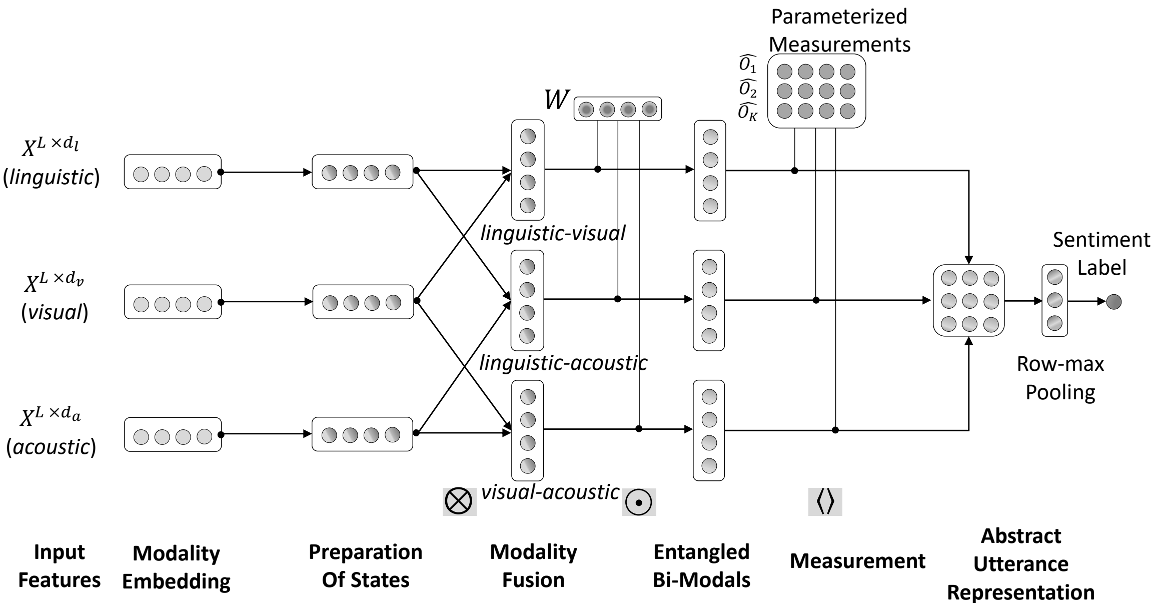


Figure 1: Entanglement-driven Fusion Neural Network (EFNN) architecture. The symbol $\otimes$ stands for the tensor product of vectors, $\odot$ the element-wise vector product, and $\langle\rangle$ the inner product of vectors. Different shades imply transformations.

but separable state is created by computing the tensor product of them. The composite separable state is defined on a d^2-dimensional joint space $\mathbb{H}_{m_1,m_2} := \otimes_2(H_m)_2$ and formulated as

$$|u_{m_1,m_2}\rangle = |u_{m_1}\rangle \otimes |u_{m_2}\rangle \,, \tag{7}$$

where m_1, m_2 are any of two modalities, and $\otimes$ defines the outer product of two states.

Then, a complex-valued neural layer W is injected to induce interactions of pairwise modalities (see Fig. 1), as follows:

$$|\widehat{u_{m_1,m_2}}\rangle = W \odot |u_{m_1,m_2}\rangle \,, \tag{8}$$

where W in $\mathbb{H} \in \mathbb{R}^{d^2}$ is a shared weight vector, and $\odot$ stands for element-wise vector product. The output is an unnormalized vector $|\widehat{u_{m_1,m_2}}\rangle$, which is then normalized to get a unit vector in $\mathbb{H}_{m_1,m_2} \in \mathbb{R}^{d^2}$, i.e., a valid quantum state: $|\widehat{u_{m_1,m_2}}\rangle = \frac{|\widehat{u_{m_1,m_2}}\rangle}{||\widehat{u_{m_1,m_2}}||^2}$, in short $|\widehat{u_{m_1,m_2}}\rangle = |u_{m_1,m_2}\rangle$.

From the representation point of view, Eq. 7 can be considered as a weighted linear transformation layer. From the quantum point of view, W can be realized as a unitary operator U [Banchi *et al.*, 2018]. Throughout the bipartite modality interaction process, W acts as a quantum Hamiltonian control on different Hilbert spaces, i.e., $\mathbb{H}_{m_1}$, $\mathbb{H}_{m_2}$, and *entanglement* is hence generated after the transformation. This means that the output after the transformation cannot be written in the decomposable form, thus giving the potential to capture nonclassical correlations across pairwise modalities.

5.4 Measurement

The measurement component acts upon the set of three non-separable pairwise modalities to identify the discriminating information for sentiment classification. In particular, a set of parameterized measurements $\{O_k\}_{k=1}^{K}$ are performed on the set of non-separable pairwise modalities (see Fig. 1), generating a sequence of positive scalars for each pair of modalities,

$$P(k) = |\langle O_k|u_{m_1,m_2}\rangle|^2, \tag{9}$$

where m_1, m_2 are any pair of modalities and each O_k represents an abstract sentiment concept. The output is a $K \times 3$ matrix of positive real values produced by measurement. Each value corresponds to the likelihood of a non-separable pairwise modality state collapsing to a basis state O_k, which is in effect a basis context representing abstract sentiment concepts. Note that the measurement component can be thought of as a dictionary learning approach.

Then a row-wise maximum pooling operator is conducted to cascade the three sequences of abstract concepts into one high-level utterance representation (see Fig. 1). Finally, the high-level representation is passed to a fully connected layer followed by a softmax classifier.

6 Experiments

6.1 Experimental Settings

We performed experiments on two widely used benchmarking video sentiment analysis datasets: CMU Multimodal Opinion-level Sentiment Intensity (CMU-MOSI) [Zadeh *et al.*, 2016], and the largest available dataset for multimodal sentiment analysis, CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [Zadeh *et al.*, 2018c]. We refer the curious readers who are interested in the feature extraction process to [Gkoumas *et al.*, 2021b].

In line with [Gkoumas *et al.*, 2021b], we chose Early-Fusion LSTM (EF-LSTM) and Late-Fusion LSTM (LF-LSTM) as baselines, Multi-Attention Recurrent Network (MARN) [Zadeh *et al.*, 2018b], Memory Fusion Network

(MFN) [Zadeh *et al.*, 2018a] and Contextual GRU with Attention (c-GRU) [Ghosal *et al.*, 2018] as advanced LSTM approaches, Tensor Fusion Network (TFN) [Zadeh *et al.*, 2017] and Low-rank Multimodal Fusion (LMF) [Liu *et al.*, 2018] as tensor-based fusion approaches, Multimodal Transformer (MulT) [Tsai *et al.*, 2019] as a seq-to-seq approach, and QMF [Li *et al.*, 2021] as a quantum-inspired fusion approach.

To evaluate the effectiveness of our model on CMU-MOSI and CMU-MOSEI tasks, we adopted a series of evaluation performance metrics used in prior work [Gkoumas *et al.*, 2021b; Zadeh *et al.*, 2018c; Zadeh *et al.*, 2018a], including: binary accuracy (i.e., Acc_2 : positive sentiment if $values \geq 0$, and negative sentiment if $values < 0$), 7-class accuracy (i.e., Acc_7 : sentiment score classification in $Z \cap [-3, 3]$), $F1$ score, Mean Absolute Error (MAE) of the score, and the Pearson's correlation ($Corr$) between the model predictions and regression ground truth. For all the metrics, the higher values denote a better performance, except MAE where the lower values denote better performance.

A grid search for the best hyper-parameters was conducted for all models. At each search, the models were trained for 100 epochs. Out of 50 searches, the model with the lowest validation loss was used to produce the test performance. The parameters in the proposed EFNN model were determined by the set of hyper-parameters $\Theta = \{D, K\}$, where D is the embedding dimension of input features into same dimensional spaces and K is the number of measurement vectors. For both datasets, we searched over a parameter pool.

We trained EFNN by feeding the real and imaginary parts of the complex-valued layers as different input parts and simulated complex operations using real values. EFNN was hence trained via the backpropagation algorithm. Measurements were initialized from standard normal distributions. All the parameters were trainable with respect to $L1$-loss defined on the extracted features. We chose Adam as the optimization algorithm.

6.2 Performance Analysis

Table 1 shows the comparison results between EFNN and the SOTA baseline approaches for the CMU-MOSI task. The approaches that apply attention mechanism to align pairwise modalities, i.e., c-GRU and MulT, exhibit the highest binary accuracy as compared to the rest of the baselines. TFN achieves the highest accuracy for Acc_7 among the baselines.

Finally, the results show that the proposed EFNN is the most effective approach for the CMU-MOSI task. In particular, it achieves an increased binary accuracy 80.9% as compared to 78.7% of MulT, which is the next best model in terms of binary accuracy. That is a significant improvement of 2.7% (t-test<.05). Overall, EFNN shows performance improvements for all evaluation metrics.

Table 2 presents the results for the CMU-MOSEI task. All approaches attain an improved performance compared to that of the CMU-MOSI dataset. We suspect this is because CMU-MOSEI is a much larger dataset. c-GRU is the most effective model among the baselines for the CMU-MOSEI task. MulT achieves similar performance to c-GRU, without a significant difference. EFNN gains an increased binary accuracy of 82.8% as compared to 80.7% of c-GRU, which is a significant improvement of 2.6% (t-test<.05). Finally, EFNN achieves an improvement for all evaluation metrics on CMU-MOSEI.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
Baseline					
EF-LSTM	32.7	75.8	75.6	1.00	0.63
LF-LSTM	32.7	76.2	76.2	0.99	0.62
LSTM					
MARN	31.8	76.4	76.2	0.98	0.63
MFN	31.9	76.2	75.8	0.99	0.62
c-GRU	33.8	78.2	78.1	0.95	0.68
Tensor					
TFN	34.9	75.6	75.5	1.01	0.61
LMF	30.5	75.3	75.2	1.02	0.61
Seq-to-Seq					
MulT	33.6	78.7	78.4	0.96	0.66
Quantum					
QMF	34.2	78.1	77.9	0.99	0.67
EFNN	**35.9**	**80.9**	**80.8**	**0.91**	**0.69**
$(\Delta\%)$	2.8%	2.7%	2.9%	3.7%	2.2%
$(\Delta_{EF}\%)$	8.9%	6.3%	6.4%	9.5%	8.7%

Table 1: Effectiveness on CMU-MOSI. Best results are highlighted in bold. $(\Delta\%)$ and $(\Delta_{EF}\%)$ indicate absolute relative percentage improvement over the next best model and the baseline EF-LSTM.

In summary, EFNN significantly outperforms the baselines for both CMU-MOSI and CMU-MOSEI tasks. The analysis of results has shown that EFNN is capable of coping with both balanced and skewed datasets.

6.3 Ablation Test

We also carried out an ablation test on CMU-MOSEI to investigate the effect of introduced quantum components. In particular, to examine the effectiveness of convolution neural networks CNN_m projecting modalities to common-dimensional spaces, we replace the component with GRU layers (a.k.a. *EFNN*$_{gru}$). Furthermore, we would like to investigate the impact of non-separable modalities, by introducing two other variants of EFNN, after removing the weight vector W (see Figure 1): a) *EFNN*$_{tensor}$ fuses all modalities into a unified tensor-based representation, i.e., trimodal fusion; and b) *EFNN*$_{con}$ concatenates all modalities into a vector representation, and then the outputs interact with the measurement component. Moreover, we also consider the impact words' position in an utterance by initializing phases from standard normal distributions (a.k.a. *EFNN*$_{rand}$). We finally replace the measurements with a convolutional neural network (CNN), whereby the K filters of CNN serve as K measurements, in order to investigate the impact of the measurement component (a.k.a. *EFNN*$_{cnn}$).

The results of the ablation test, illustrated in Table 3, show that each component plays a crucial role in the EFNN. In particular, the comparison with EFNN$_{rand}$ shows the effectiveness of modelling words' position into the phase part of complex-valued representations. At the same time, the decreased performance of EFNN$_{tensor}$ and EFNN$_{con}$ reveals

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
Baseline					
EF-LSTM	45.7	78.2	77.1	0.69	0.57
LF-LSTM	47.1	79.2	78.5	0.66	0.61
LSTM					
MARN	47.7	79.3	77.8	0.65	0.63
MFN	47.4	79.9	79.1	0.65	0.63
c-GRU	48.4	80.7	80.2	0.63	0.67
Tensor					
TFN	47.3	79.3	78.2	0.66	0.62
LMF	47.6	78.2	77.6	0.66	0.62
Seq-to-Seq					
MulT	46.6	80.2	79.8	0.64	0.65
Quantum					
QMF	47.2	79.8	79.4	0.65	0.66
EFNN	**50.2**	**82.8**	**82.6**	**0.60**	**0.69**
(Δ%)	3.6%	2.6%	2.9%	5.4%	2.5%
(Δ_{EF}%)	9.0%	5.6%	6.7%	15.5%	16.8%

Table 2: Effectiveness on CMU-MOSEI. Best results are highlighted in bold. (Δ%) and (Δ_{EF}%) indicate absolute relative percentage improvement over the next best model and the baseline EF-LSTM.

Approach	Acc_7	Acc_2	$F1$	MAE	$Corr$
EFNN$_{gru}$	-1.5%	-1.3%	-1.3%	-1.2%	-1.1%
EFNN$_{tensor}$	-2.2%	-1.8%	-1.8%	-1.5%	-1.7%
EFNN$_{con}$	-2.8%	-2.5%	-2.5%	-2.4%	-2.2%
EFNN$_{random}$	-1.5%	-1.2%	-1.2%	-1.2%	-1.4%
EFNN$_{cnn}$	-1.9%	-1.4%	-1.4%	-1.6%	-1.7%

Table 3: Ablation test on CMU-MOSEI. Absolute relative percentage difference from EFNN.

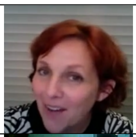
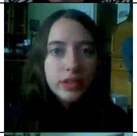
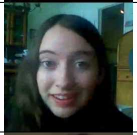
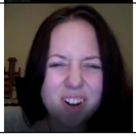
Linguistic	Visual	Sentiment
The story was all right.		Positive
I do not wanna see any more of this.		Negative
The voice acting was phenomenal!		Positive
Yeap a horrible protagonist!		Negative

Table 4: The first two examples illustrate cases of the most entangled bi-modals for the linguistic and visual modalities. The last two examples illustrates cases of the least entangled bi-modals.

the superiority of encoding the non-classical correlations (i.e., entanglements) between modalities. Moreover, the comparison with EFNN$_{cnn}$ shows the usefulness of trainable measurements. Finally, EFNN$_{gru}$ shows that convolutional neural networks could be a more appropriate way to project modalities into common-dimensional spaces. Overall, the ablation test reveals that the entanglement-driven fusion component plays the most crucial role in the architecture of EFNN.

6.4 Post-hoc Interpretability

Further, we evaluated the post-hoc interpretability by investigating the bi-modal correlations within composite utterance states after the modality context interaction. In particular, according to Eq.4, we calculated the degree of quantum entanglement for bipartite composite utterance states of linguistic and visual modalities.

Table 4 illustrates some examples of the most and least entangled linguistic-visual modalities according to entanglement entropy. The most entangled pairs are those that one of two modalities is ambiguous or uninformative. For example, in Table 4, the linguistic content of the first utterance is ambiguous, while the visual content of the second utterance is uninformative. By contrast, when the context of both modalities is informative, unambiguous, and simultaneously present (e.g., the last two utterances in Table 4), the entanglement entropy is close to zero. In those cases, the composite representation is separable, and there is no need to exploit the quantum probabilistic interpretation. However, through the concept of non-separability, EFNN is able to capture both separable and non-separable bi-modal interactions, as a generalization of existing probabilistic modality fusion approaches. This attribute is the core reason that EFNN has achieved performance improvement.

7 Conclusion

In this work, we have introduced a transparent and joint quantum probabilistic neural model for video sentiment analysis. The concept and formalism of non-separability as quantum entanglement to fuse bi-modalities enables the model to capture both classical and non-classical correlations between modalities. The information encoded in the non-classical correlations yielded significant improvements in performance over a range of strong baselines. Besides, non-classical correlations were quantified by an appropriate measure, which optimized post-hoc interpretability. In the future, we would like to extend the framework for video emotion detection in conversations that requires leveraging the conversational context and evolution of systems.

Acknowledgements

This work is funded in part by the National Key Research and Development Program of China (grant No. 2018YFC0831704), Royal Society project on Big Code (grant No. IES/R1/191138), EPSRC STRIDE project (grant No. EP/T017465/1), and the Trusted Programming Lab at Huawei Ireland Research Center.

References

[Banchi *et al.*, 2018] Leonardo Banchi, Edward Grant, Andrea Rocchetto, and Simone Severini. Modelling non-markovian quantum processes with recurrent neural networks. *New Journal of Physics*, 20(12):123030, 2018.

[Bruza *et al.*, 2008] Peter Bruza, Kirsty Kitto, and Doug McEvoy. Entangling words and meaning. In *Quantum Interaction: Proceedings of the Second Quantum Interaction Symposium (QI-2008)*, pages 118–124. College Publications, 2008.

[Bruza *et al.*, 2009] Peter Bruza, Kirsty Kitto, Douglas Nelson, and Cathy McEvoy. Is there something quantum-like about the human mental lexicon? *Journal of mathematical psychology*, 53(5):362–377, 2009.

[Busemeyer and Bruza, 2012] Jerome R Busemeyer and Peter D Bruza. *Quantum models of cognition and decision*. Cambridge University Press, 2012.

[Ghosal *et al.*, 2018] Deepanway Ghosal, Md Shad Akhtar, Dushyant Chauhan, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Contextual inter-modal attention for multi-modal sentiment analysis. In *proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3454–3466, 2018.

[Gkoumas *et al.*, 2018] Dimitris Gkoumas, Sagar Uprety, and Dawei Song. Investigating non-classical correlations between decision fused multi-modal documents. In *International Symposium on Quantum Interaction*, pages 163–176. Springer, 2018.

[Gkoumas *et al.*, 2021a] Dimitris Gkoumas, Qiuchi Li, Shahram Dehdashti, Massimo Melucci, Yijun Yu, and Dawei Song. Quantum cognitively motivated decision fusion for video sentiment analysis. *arXiv preprint arXiv:2101.04406*, 2021.

[Gkoumas *et al.*, 2021b] Dimitris Gkoumas, Qiuchi Li, Christina Lioma, Yijun Yu, and Dawei Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197, 2021.

[Halmos, 1987] P.R. Halmos. *Finite-Dimensional Vector Spaces*. Undergraduate Texts in Mathematics. Springer, New York, USA, 1987.

[Lai *et al.*, 2015] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *AAAI-2015*, 2015.

[Li *et al.*, 2019] Qiuchi Li, Benyou Wang, and Massimo Melucci. Cnm: An interpretable complex-valued network for matching. *arXiv preprint arXiv:1904.05298*, 2019.

[Li *et al.*, 2021] Qiuchi Li, Dimitris Gkoumas, Christina Lioma, and Massimo Melucci. Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion*, 65:58–71, 2021.

[Liu *et al.*, 2018] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

[Melucci, 2015] Massimo Melucci. *Introduction to information retrieval and quantum mechanics*. Springer, 2015.

[Sordoni *et al.*, 2013] Alessandro Sordoni, Jing He, and Jian-Yun Nie. Modeling latent topic interactions using quantum interference for information retrieval. In *ACL-2013*, pages 1197–1200, 2013.

[Tsai *et al.*, 2019] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[Uprety *et al.*, 2020] Sagar Uprety, Dimitris Gkoumas, and Dawei Song. A survey of quantum theory inspired approaches to information retrieval. *ACM Computing Surveys (CSUR)*, 53(5):1–39, 2020.

[Van Rijsbergen, 2004] Cornelis Joost Van Rijsbergen. *The geometry of information retrieval*. Cambridge University Press, 2004.

[Vivo *et al.*, 2016] Pierpaolo Vivo, Mauricio P Pato, and Gleb Oshanin. Random pure states: Quantifying bipartite entanglement beyond the linear statistics. *Physical Review E*, 93(5):052106, 2016.

[Wang *et al.*, 2010] Jun Wang, Dawei Song, and Leszek Kaliciak. Tensor product of correlated text and visual features. In *In QI'10*. Citeseer, 2010.

[Zadeh *et al.*, 2016] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

[Zadeh *et al.*, 2018a] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Zadeh *et al.*, 2018b] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Zadeh *et al.*, 2018c] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.