SHAPE: An Unified Approach to Evaluate the Contribution and Cooperation of Individual Modalities

Pengbo Hu[1,2*] , **Xingyu Li**[2*] and **Yi Zhou**[3,2†]

[1]School of Mathematical Sciences, University of Science and Technology of China, Hefei, China
[2]Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai, China
[3]National Engineering Laboratory for Brain-inspired Intelligence Technology and Application, School of Information Science and Technology, University of Science and Technology of China, Hefei, China
pbhu@mail.ustc.edu.cn, xyli1905@bsbii.cn, yizhoujoey@gmail.com

Abstract

As deep learning advances, there is an ever-growing demand for models capable of synthesizing information from multi-modal resources to address the complex tasks raised from real-life applications. Recently, many large multi-modal datasets have been collected, on which researchers actively explore different methods of fusing multi-modal information. However, little attention has been paid to quantifying the contribution of different modalities within the proposed models. In this paper, we propose the **SH**apley v**A**lue-based **PE**rceptual (SHAPE) scores that measure the marginal contribution of individual modalities and the degree of cooperation across modalities. Using these scores, we systematically evaluate different fusion methods on different multi-modal datasets for different tasks. Our experiments suggest that for some tasks where different modalities are complementary, the multi-modal models still tend to use the dominant modality alone and ignore the cooperation across modalities. On the other hand, models learn to exploit cross-modal cooperation when different modalities are indispensable for the task. In this case, the scores indicate it is better to fuse different modalities at relatively early stages. We hope our scores can help improve the understanding of how the present multi-modal models operate on different modalities and encourage more sophisticated methods of integrating multiple modalities.

1 Introduction

The present deep learning algorithms thrive on a wide spectrum of real-life tasks, such as those in the fields of computer vision and natural language processing, for which the essence is to learn the underlying distribution from the dataset. Inspired by these achievements, researchers march on to conquer the more complex tasks that put emphasis on reasoning and knowledge. One character of these complex tasks is that they normally involve data from distinct modalities.

For example, in the Visual Question Answering (VQA) task, one must answer the questions based on the knowledge extracted from the corresponding image [Goyal *et al.*, 2017; Antol *et al.*, 2015]. The ability to fuse multi-modal information serves as the cornerstone to developing solutions to those complex tasks, and it presents severe challenges to the deep learning community.

Recently, many large multi-modal datasets have been collected, e.g. [Vadicamo *et al.*, 2017; Zadeh *et al.*, 2018; Bowman *et al.*, 2015; Liang *et al.*, 2021], laying the ground for exploring sophisticated fusion strategies. Beyond the naive methods of early and late fusion strategies, increasing attention has been paid to developing and practicing hybrid fusing strategies [Wang, 2021; Baltrušaitis *et al.*, 2018; Wang *et al.*, 2020; Bai *et al.*, 2021]. Those efforts constantly push forward the model performance on the multi-modal datasets. Despite the success, people gradually realize a subtle limitation of the present progresses: the model performance is almost solely judged on their prediction accuracy. A higher accuracy does not necessarily translate to the better usage of multi-modal information. For example, in the CMU-MOSEI datasets, the accuracy of a uni-modal model trained on the text modality is very close to the ones of SOTA multi-modal models [Zadeh *et al.*, 2018; Delbrouck *et al.*, 2020]. Consequently, there is an urgent need for a quantitative metric that can capture how the underlying models operate on the multi-modal information.

To evaluate the individual modalities' importance, Gat et al. [2021] proposed the perceptual score, which estimates the contribution of a modality with its marginal performance gain at the presence of other modalities. However, this score cannot measure cross-modal cooperation and, in the presence of more than two modalities, its attribution of the importance to individual modalities does not correctly reflect their marginal contribution. To address the above issues, we draw inspirations from the game theory and propose the **SH**apley v**A**lue-based **PE**rceptual (SHAPE) scores to directly evaluate the marginal contribution and the cooperation of individual modalities. The validity of the SHAPE scores roots in the properties of the celebrated Shapley value [Shapley, 1953].

---

*Co-first authors.

†The corresponding author.

[1]The full version of our paper: https://github.com/zhouyilab/SHAPE/blob/main/SHAPE-IJCAI22.pdf; Implementation: https://github.com/zhouyilab/SHAPE.

With the SHAPE scores, we aim at providing additional perspectives for evaluating the performance of the multi-modal models, which in turn would help building sophisticated fusion strategies.

Our contributions are three-fold:

1. We propose the SHAPE scores, which are theoretically guaranteed to quantify the marginal contribution of individual modalities and the cross-modal cooperation.

2. With the SHAPE scores, we systematically evaluate the usage of multi-modal information by models using different fusion methods on the different multi-modal datasets for different tasks.

3. We extend the perceptual score to a more informative fine-grained version. Moreover, we demonstrate that the perceptual score implicitly measures the correlation among modalities within the dataset.

2 Related Work

2.1 Multi-modal Fusion Method

The fusion strategies can be roughly classified as early fusion, late fusion, and hybrid fusion [Baltrušaitis *et al.*, 2018]. In the early fusion, one directly concatenates inputs from individual modalities to build a unified input embedding. In this case, a single pathway model is enough to make a prediction, making the training pipeline easier [Kiela *et al.*, 2019]. The late fusion usually processes individual modalities in different pathways. The fusion operation is typically implemented on the output from each pathway. The later fusion strategy ignores the low-level feature interaction, making it hard to learn the joint representation of modalities. However, in practice, later fusion enables a model using the different pre-trained models for different pathways, making it more flexible than the early fusion [Rodrigues Makiuchi *et al.*, 2019]. Hybrid fusion has no specific recipe. The different modalities can be integrated at different stages, on different feature levels, and among different modules. It could effectively promote collaboration among individual modalities. Generally, the optimal hybrid fusion strategy depends on the type of the underlying task. For example, in the VQA task, the co-attention-based fusion methods are proved to be better than other strategies [Liu *et al.*, 2019].

2.2 Modality Contribution Attribution

Presently, most evaluation works focus on the feature importance. For example, Lundberg and Lee [2017] proposed the SHapley Additive exPlanations (SHAP) value, which can estimate feature importance using an approximation of the features Shapley values. For evaluating modal importance, Hessel and Lee [2020] adopted an indirect attribution method. They designed a tool called the empirical multimodally-additive function projection (EMAP), which is an effective model that elimates the cross-modal cooperation in the reference multi-modal model. Through comparing the performance of the reference model and its EMAP, one could infer how much the cross-modal cooperation contribute.

The work most relevant to ours is the perceptual score, which is proposed by Gat *et al.* [2021] to assess the importance of a modality in the input data for a given model.

They adopted a permutation-based approach for removing the modality influence. Specifically, let x^M denote the feature of target modality M in the input data point. In order to remove M's influence, x^M is substituted with a randomly chosen one from another data point, i.e., x'^M. The resulting classifier on the permutated data is called the permuted classifier. The perceptual score for M is defined as the normalized difference between the accuracies of the normal classifier and the corresponding permuted classifier.

It is worth noting that the perceptual score does not account for the marginal contribution of the target modality M. Instead, it measures the M's contribution conditioned on all other modalities in the dataset. The two types of contributions are the same only in the two-modal case. In the presence of more modalities, it requires a more sophisticated method to compute the marginal contribution. Moreover, there still lacks a quantitative method that isolates the cooperation between different modalities.

2.3 Shapley Value

The Shapley value [Shapley, 1953] was initially proposed in the coalitional game theory for fair attribution of payouts among the players depending on their contribution to the total payout. A multi-player game is usually defined as a real-valued utility function V on its participants. Let $\mathcal{N} := \{1, \dots, n\}$ denote the set of players in the game. The player can form different coalitions (i.e., subsets of $\mathcal{N}$) whose net effect can be either cooperative or competing. Consider a player i and a subset $S \subseteq \mathcal{N}/\{i\}$, the marginal contribution of i to S can be evaluated by

$$m_i(S; V) = V(S \cup \{i\}) - V(S).$$

Since a player can interact with any other players, in order to evaluate the marginal contribution of i in the game $V(\mathcal{N})$ one needs to average over all possible $m_i(S; V)$. This is exactly how the Shapley value ϕ_i is defined,

$$\phi_i(\mathcal{N}; V) := \sum_{S \subseteq \mathcal{N}/\{i\}} \frac{|S|!(n - |S| - 1)!}{n!} m_i(S; V).$$

A fair attribution of payouts is considered to satisfy the properties of Efficiency, Symmetry, Dummy and Additivity, and the Shapley value is the only attribution method that has those four properites [Molnar, 2020]. More precisely,

Efficiency. All the individual rewards sum to the overall net payouts, i.e., $\sum_i \phi_i = V(\mathcal{N}) - V(\emptyset)$, where $V(\emptyset)$ indicates the baseline reward without any player.

Symmetry. If two palyers i, j contribute equally to all possible coalitions, i.e., $V(S \cup \{i\}) = V(S \cup \{j\}), \forall S \subseteq \mathcal{N}/\{i, j\}$, then they have the same reward, $\phi_i = \phi_j$.

Dummy. A dummy player i, which does not contribute to any coalition, i.e., $V(S \cup \{i\}) = V(S) + V(\{i\}), \forall S \subseteq \mathcal{N}/\{i\}$, has the reward as if playing alone, $\phi_i = V(\{i\})$.

Additivity. Consider a combined game $U(S) = V(S) + W(S), \forall S \subseteq \mathcal{N}$, its reward is the sum of the rewards for individual games, i.e. $\phi_i(U) = \phi_i(V) + \phi_i(W)$.

Those four properties in turn could be viewed as axioms defining a fair attribution method, and it has been proved that
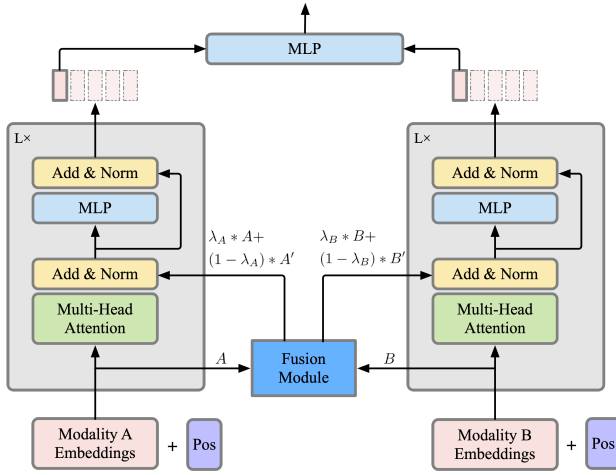
Figure 1: Schematic diagram for the dense fusion strategy, implemented with the transformer encoder blocks.

they uniquely lead to the definition of the Shapley value [Weber, 1988].

3 Method

3.1 Models for Comparison

In deep learning, diverse fusion methods have been developed to address the multi-modal learning tasks [Baltrušaitis *et al.*, 2018; Wang, 2021]. Among them, the Early Fusion and Late Fusion strategies are widely recognized. Besides these two, the rest methods cannot be easily categorized. Depending on specific implementation, the fusion may happen at one particular stage or at multiple stages. In addition, different works usually adopt distinct network architectures, making it infeasible to compare them on the same ground.

The present work is devoted to evaluate the contribution and cooperation of individual modalities under different fusion methods on different multi-modal datasets for different tasks. To achieve this goal, we need (i) a common network architecture that ensures decent performance on different datasets for different tasks. (ii) a unified framework to account for different fusion types.

To address point (i), we adopt the transformer [Vaswani *et al.*, 2017] architecture, which is capable of handling different modalities. We are aware that using a common architecture for different tasks would potentially degrade the performance. For point (ii), we introduce the Dense Fusion strategy that builds upon the transformer architecture (see Figure 1). The fusion happens at every stage. We also consider one of its variations, which includes weights (λ's) to control the fusion intensity. Those weights are learned during the training. Such a variation is referred as the Dynamic Dense Fusion.

To summarize, in our experiments, we consider the Early Fusion, Late Fusion, as well as two variations of the Dense Fusion strategy. All the network models are built upon the transformer architecture. Details can be found in the experiment setup and in the supplementary material.

3.2 The SHAPE Scores

The SHAPE scores are built upon the Shapley value with identifying the set of modalities as players and model accuracy as the utility function. We distinguish two types of SHAPE scores, one captures the marginal contribution of individual modalities, and the other captures the cross-modal cooperation. The SHAPE scores are easy to compute under the general multi-modal learning scenario, no matter how many modalities are involved.

In the following, we use $V_f(\mathcal{D})$ to denote the accuracy of model f on the dataset $\mathcal{D}$. When discussing the contribution of different modalities, e.g., M_1 and M_2, we shall slightly change the symbol to $V_f(M_1, M_2)$ for emphasizing the involved modalities. The set of modalities in $\mathcal{D}$ is denoted as $\mathcal{M} := \{M_1, \ldots, M_m\}$.

SHAPE Score for Modality's Marginal Contribution

For the purpose of evaluating the marginal contribution of a modality M_i, we use the scaled Shapley value

$$\mathcal{S}_{M_i; f; \mathcal{D}} := \frac{1}{Z_f} \phi_{M_i}(\mathcal{M}; V_f), \tag{1}$$

where the normalization factor Z_f is introduced to align the scores for different models and thus enable cross-model comparison. Similar to the modal-normalization in [Gat *et al.*, 2021], we choose Z_f to be the model accuracy on the full dataset $\mathcal{D}$, i.e., $Z_f = V_f(\mathcal{D})$. For simplicity, we will use the abbreviated symbol $\mathcal{S}_{M_i}$, when there is no risk of confusion.

As an example, for the two-modal case, we have

$$\mathcal{S}_{M_1} = \frac{1}{2Z_f}[V(M_1, M_2) - V(0_1, M_2) \\ + V(M_1, 0_2) - V(0_1, 0_2)],$$

where $Z_f = V(M_1, M_2)$. Extensions to three or more modalities cases are straightforward, and we leave them to the supplementary material.

Baseline Values. The above 0_i refers to the absence of modality M_i and should be implemented using some predefined baseline values. In the case where only one modality is absent, it is natural to substitute zeros for the missing inputs. However, when both modalities are absent, i.e., $V(0_1, 0_2)$, an all-zero input would produce a single deterministic prediction that is sensitive to specific training settings, such as the model initialization. We argue that it is better to predict the majority class in this case. This majority-classifier is stable and could account for the possible unbalances in the dataset.

SHAPE Score for Cross-modal Cooperation

The Shapley value also enables us to directly measure the cross-modal cooperation, which evaluates the degree to which the model exploits the complementary information of a modality provided to the other modalities.

Intuitively, to compute the cooperation among a set of modalities $A \subseteq \mathcal{M}$, we need to subtract the marginal contributions of all $M_i \in A$ from the marginal contribution of A itself. Treating A as a single effective player, its marginal contribution can be evaluated by $\phi_A(\mathcal{M}; V_f)$. On

the other hand, the marginal contribution of $M_i \in A$ should be $\phi_{M_i}(\mathcal{M}/A \cup M_i; V_f)$. Consequently, the cooperation score is defined as

$$\mathcal{C}_{A;f;\mathcal{D}} = \phi_A(\mathcal{M}; V_f) - \sum_{M_i \in \mathcal{A}} \phi_{M_i}(\mathcal{M}/A \cup M_i; V_f). \quad (2)$$

Note that the set of "players" in the above ϕ_{M_i} is different from the one in Equation (1). For simplicity, we usually omit the subscripts $f; \mathcal{D}$ and write the score as $\mathcal{C}_A$.

For a concrete example, let's consider the two-modal case. After a few steps of simple algebra, we arrive at

$$\begin{aligned} \mathcal{C}_{\{M_1, M_2\}} = {}& V(M_1, M_2) - V(M_1, 0_2) - V(0_1, M_2) \\ & + V(0_1, 0_2), \end{aligned}$$

where 0_1 and 0_2 denote the baseline values for modalities M_1, M_2, respectively.

3.3 Fine-grained Perceptual Scores

In this subsection, we proceed to have a closer look at the perceptual score [Gat *et al.*, 2021]. First, we will show that this score implicitly measures the correlation among different modalities within the dataset $\mathcal{D}$. Then we propose a fine-grained version of the perceptual score. In the following, we only keep the key steps and leave the detailed deduction to the supplementary material.

Let (x, y) denote a data sample in $\mathcal{D}$, where $x \in \mathcal{X}$ is the input feature and $y \in \{0, \ldots, k-1\}$ is the corresponding label. In order to explicitly refer the modalities, we will decompose x as $(x^{\mathcal{M}_1}, x^{\mathcal{M}_2})$, where $\mathcal{M}_1 \subseteq \mathcal{M}$ and $\mathcal{M}_2 = \mathcal{M}/\mathcal{M}_1$ are the subsets of modalities. Recall that the model accuracy is defined as $V_f(\mathcal{D}) = \mathbb{E}_{\mathcal{D}}[\mathbb{I}_{y^*=y}]$, where $\mathbb{I}$ is the identity function and y^* refers to the model prediction at input x. Simple algebra will lead us to

$$V_f(\mathcal{D}) = \int_{\mathcal{X}} dx^{\mathcal{M}_1} dx^{\mathcal{M}_2} \mathrm{Pr}(y*, x^{\mathcal{M}_1}, x^{\mathcal{M}_2}),$$

where $\mathrm{Pr}(y*, x^{\mathcal{M}_1}, x^{\mathcal{M}_2})$ is the probability distribution on $\mathcal{D}$.

The permutation-based step in the perceptual score would alter the data distribution, resulting in a new effective dataset with distribution $\mathrm{Pr}(y, x^{\mathcal{M}_1})\mathrm{Pr}(x^{\mathcal{M}_2})$. We refer to this effective dataset as $\mathcal{D}^p$. Now, the perceptual score for $\mathcal{M}_2$ is

$$\begin{aligned} P_{\mathcal{M}_2;f;\mathcal{D}} &= \frac{1}{Z_f}(V_f(\mathcal{D}) - V_f(\mathcal{D}^p)) \\ &= \frac{1}{Z_f} \int_{\mathcal{X}} dx^{\mathcal{M}_1} dx^{\mathcal{M}_2} P(x^{\mathcal{M}_2}) \Delta_{\mathcal{M}_1, \mathcal{M}_2; f}, \end{aligned}$$

where $\Delta_{\mathcal{M}_1, \mathcal{M}_2; f} = P(y^*, x^{\mathcal{M}_1} | x^{\mathcal{M}_2}) - P(y^*, x^{\mathcal{M}_1})$. Clearly, apart from the dependence on the model f through y^*, the term $\Delta_{\mathcal{M}_1, \mathcal{M}_2; f}$ effectively measure the correlation between $\mathcal{M}_1$ and $\mathcal{M}_2$ in the dataset $\mathcal{D}$.

Since $P(x^{\mathcal{M}_2})$ is non-negative, if the part $x^{\mathcal{M}_2}$ comes from the same class as $x^{\mathcal{M}_1}$, then we would expect $\Delta_{M_1, M_2; f} \geq 0$. We refer to this case as the in-class permutation. Likewise, in the out-class permutation, $x^{\mathcal{M}_2}$ comes from a different class from $x^{\mathcal{M}_1}$ and we expect $\Delta_{M_1, M_2; f} \leq 0$.

The in-class and out-class cases provide conflicting driving forces that would cancel each other and cannot be reflected by the original perceptual score. Based on this observation, we propose to use the in-class and out-class perceptual scores in place of the original one. They are defined as following:

$$P_{\mathcal{M}_i;f;\mathcal{D}}^{in/out} := \frac{1}{Z_f} \left[V_f(\mathcal{M}) - V_f([\mathcal{M}/\mathcal{M}_i]^{\pm}) \right],$$

where the $[\mathcal{M}/\mathcal{M}_i]^{\pm}$ operator removes the $\mathcal{M}_i$ modality either by the in-class permutation $(+)$ or the out-class permutation $(-)$. Z_f is the normalization factor as usual. We only consider the model-normalization here. Finally, for simplicity, we usually omit the subscripts $f; \mathcal{D}$ as long as it raises no confusion.

4 Experiments

4.1 Setup

For all models, we use 4-layer transformer encoder as the backbone. The number of attention heads is 8, and the hidden dimension is 1024. All models are trained using the Adam optimizer with learning rate warm-up in the first epoch. We encode text data with the GloVe [Pennington *et al.*, 2014], and audio data with the MFCC [Tiwari, 2010]. As for the image data, we employ the patch embedding technique, following the ViT Model [Dosovitskiy *et al.*, 2020]. It is worth noting that our models achieve comparable performance with the supervision-based SOTA models.

4.2 Datasets

Real-life applications raise diverse demands for multi-modal data. Depending on specific tasks, some must be handled by combining all the involved modalities, while in the others, additional modalities merely provide complementary information. Furthermore, different modalities in a multi-modal dataset might be only weakly correlated to each other, or they could be different representations that provide precisely the same information. To account for those complicated situations, we deliberately choose three datasets, as will be introduced below.

B-T4SA. The dataset [Vadicamo *et al.*, 2017] is collected from Twitter, including the posted text (textual modality T) and associated images (visual modality V). Each data pair is annotated with a sentiment label (positive, negative and neutral). The task is to predict this label using both the text and image inputs. We note that the text data normally align with their sentiment labels while a large number of images seem to be irrelevant to the labels. B-T4SA represents the situation that different modalities are complementary and weakly correlated to each other.

CMU-MOSEI. This dataset [Zadeh *et al.*, 2018] is the largest dataset for sentence-level sentiment analysis and emotion recognition. It contains more than 65 hours of annotated video data, consisting of audio, video and text modalities. In our experiments, we only use the audio (A) and text (T) modalities. The text modality includes the scripts that are generated from the audio data. In this way, the CMU-MOSEI dataset stands for the situation that different modalities are complementary and highly correlated.

B-T4SA		SHAPE scores			(fine-grained) perceptual scores					
	Acc.	$\mathcal{S}_V$	$\mathcal{S}_T$	$\mathcal{C}_{VT}$	P_V	P_V^{in}	P_V^{out}	P_T	P_T^{in}	P_T^{out}
Unimodal V	41.95	–	–	–	–	–	–	–	–	–
Unimodel T	94.89	–	–	–	–	–	–	–	–	–
Early Fusion	95.01	0.52	64.40	1.04	$0.02_{\pm0.01}$	$0.03_{\pm0.01}$	$0.03_{\pm0.02}$	$65.00_{\pm0.26}$	$-0.04_{\pm0.14}$	$97.34_{\pm0.09}$
Late Fusion	95.22	0.00	65.00	0.00	$2.09_{\pm0.01}$	$-0.01_{\pm0.01}$	$0.61_{\pm0.01}$	$64.95_{\pm0.27}$	$-0.02_{\pm0.11}$	$97.43_{\pm0.12}$
Dense Fusion	94.97	0.24	64.66	0.48	$0.01_{\pm0.01}$	$0.01_{\pm0.01}$	$0.02_{\pm0.01}$	$64.84_{\pm0.24}$	$0.14_{\pm0.13}$	$97.23_{\pm0.09}$
Dynamic Fusion	95.36	0.01	65.04	0.01	$0.01_{\pm0.01}$	$0.01_{\pm0.01}$	$0.00_{\pm0.01}$	$65.02_{\pm0.30}$	$0.05_{\pm0.14}$	$97.57_{\pm0.17}$

Table 1: Accuracy (Acc.), the SHAPE scores ($\mathcal{S}_V$, $\mathcal{S}_T$, $\mathcal{C}_{VT}$) and the (fine-grained) perceptual scores ($P_{V/T}$, $P_{V/T}^{in}$, $P_{V/T}^{out}$) on the B-T4SA dataset for multi-modal models using different fusion strategies. Dynamic Fusion stands for the Dynamic Dense Fusion. The results of unimodal models are included as baselines. V and T indicate the visual and text modalities, respectively.

CMU-MOSEI		SHAPE scores			(fine-grained) perceptual scores					
	Acc.	$\mathcal{S}_A$	$\mathcal{S}_T$	$\mathcal{C}_{AT}$	P_A	P_A^{in}	P_A^{out}	P_T	P_T^{in}	P_T^{out}
Unimodal A	71.04	–	–	–	–	–	–	–	–	–
Unimodel T	80.20	–	–	–	–	–	–	–	–	–
Early Fusion	80.22	-0.01	11.46	-0.02	$1.21_{\pm0.17}$	$-0.62_{\pm1.50}$	$-0.30_{\pm0.16}$	$21.76_{\pm0.64}$	$-0.47_{\pm0.24}$	$59.85_{\pm1.66}$
Late Fusion	80.67	0.04	11.89	0.09	$0.72_{\pm0.09}$	$2.12_{\pm0.22}$	$2.38_{\pm0.31}$	$22.04_{\pm0.76}$	$2.91_{\pm1.23}$	$50.68_{\pm1.45}$
Dense Fusion	79.04	0.19	9.93	0.38	$-0.08_{\pm0.20}$	$0.65_{\pm0.47}$	$0.53_{\pm0.33}$	$19.04_{\pm0.90}$	$0.90_{\pm1.19}$	$46.95_{\pm1.28}$
Dynamic Fusion	80.05	-0.21	11.47	-0.42	$2.41_{\pm0.14}$	$0.47_{\pm0.17}$	$0.65_{\pm0.22}$	$21.99_{\pm0.87}$	$0.31_{\pm0.95}$	$54.60_{\pm1.75}$

Table 2: The same as Table 1, but for the MOSEI dataset. The invoved modalities are audio (A) and text (T).

SNLI		SHAPE scores			(ffine-grained) perceptual scores					
	Acc.	$\mathcal{S}_{T_p}$	$\mathcal{S}_{T_h}$	$\mathcal{C}_{T_p T_h}$	P_{T_p}	$P_{T_p}^{in}$	$P_{T_p}^{out}$	P_{T_h}	$P_{T_h}^{in}$	$P_{T_h}^{out}$
Unimodal T_p	33.84	–	–	–	–	–	–	–	–	–
Unimodel T_h	66.32	–	–	–	–	–	–	–	–	–
Early Fusion	82.47	21.34	38.25	42.52	$42.89_{\pm0.52}$	$42.89_{\pm0.30}$	$42.50_{\pm0.15}$	$59.84_{\pm0.25}$	$42.42_{\pm0.76}$	$68.53_{\pm0.32}$
Late Fusion	69.77	6.41	45.82	10.55	$6.97_{\pm0.40}$	$6.94_{\pm0.30}$	$7.30_{\pm0.47}$	$52.33_{\pm1.18}$	$7.07_{\pm0.79}$	$60.69_{\pm0.46}$
Dense Fusion	84.91	25.49	35.26	50.95	$47.65_{\pm0.28}$	$47.73_{\pm0.37}$	$47.58_{\pm0.18}$	$61.04_{\pm0.25}$	$47.50_{\pm0.62}$	$67.87_{\pm0.45}$
Dynamic Fusion	85.02	25.61	35.19	52.06	$48.40_{\pm0.25}$	$48.24_{\pm0.31}$	$48.66_{\pm0.38}$	$61.06_{\pm0.47}$	$48.09_{\pm0.68}$	$67.54_{\pm0.49}$

Table 3: The same as Table 1, but for the SNLI dataset. The invoved modalities are the premise text (T_p) and the hypothesis text (T_h).

SNLI. The SNLI dataset [Bowman *et al.*, 2015] is a collection of 570k human-written English sentence pairs. Each pair contains a premise and a hypothesis, whose relation is labeled as either entailment, contradiction, or neutral. We shall view the premise and hypothesis as two distinct modalities, i.e., T_p and T_h. The underlying task is to infer the relationship between the premise and the hypothesis. Clearly, this task intrinsically relies on both modalities and, thus, represents the situation that different modalities are indispensable.

4.3 Contribution and Cooperation

In this subsection, we investigate the SHAPE scores across different models on the three multi-modal datasets, demonstrating their ability in capturing the marginal contribution of individual modalities and the cross-modal cooperation.

Table 1 summarizes the results for B-T4SA dataset. We find the unimodal model for T performs as good as multi-modal models, while the unimodal model for V only slightly suppasses the random guess. The SHAPE scores align with these observations. The contribution scores $\mathcal{S}_V$ for V modal-

ity are near zero for all the fusion strategies, indicating V almost does not contribute to the task. Further, the cooperation scores $\mathcal{C}_{VT}$ are very small as well. Hence, as expected, nearly no cooperation happens between the two modalities. At last, we note that $\mathcal{C}_{VT}$ for early fusion and dense fusion are evidently higher than the other two models. This may be attributed to the fact that both the early and the dense fusion models encourage the fusion between the two modalities.

The A modality in the CMU-MOSEI dataset encodes richer information than its text counterparts by including features such as the tone and the talking speed. On the other hand, the T modality is better at representing semantic information. One would expect a multi-modal model to actively integrate the two modalities and make better predictions. Unfortunately, it is not the case as indicated by the SHAPE scores (see Table 2). Values of $\mathcal{S}_A$ and $\mathcal{C}_{AT}$ are very small, indicating the audio data seldom contribute to the model performance and there is little cooperation between the A and T modalities. In addition, the marginal contributions of T modality, $\mathcal{S}_T$, are also relatively small. At first sight,

it may seem contradictory that both modalities have limited contribution to the performance, while the model achieves a decent accuracy. Closer investigation reveals that this is a consequence of the unbalanced dataset. Simply by predicting positive, one could get an accuracy of 71.04%, which is the accuracy of the unimodal model for A. This phenomenon coincides with the common belief that the deep neural networks are lazy, tending to take shortcuts. It could be the case that the multi-modal models simply exploit the unbalance of the dataset and pay little effort in mining the cooperation.

In Table 3, we show the results for the natural language inference task on the SNLI dataset, for which both modalities are indispensable. We find that both unimodal models perform poorly. Interestingly, the unimodal model for T_h significantly outperforms the random guess. This implies that there could be implicit biases within the dataset despite that it is balanced. As expected, the behavior of SHAPE scores is different from previous cases. $\mathcal{S}_{T_p}$ and $\mathcal{S}_{T_h}$ becomes comparable, with $\mathcal{S}_{T_h}$ being generally higher. Hence, both modalities contribute, and the hypothesis is more important for model prediction. More strikingly, the cooperation scores $\mathcal{C}_{T_pT_h}$ are also high. This suggests that the models actively exploit the cooperation between T_p and T_h for predictions. At last, in the late fusion case, $\mathcal{S}_{T_p}$ and $\mathcal{C}_{T_pT_h}$ are much lower than the other cases. The lack of cooperation could be attributed to the nature of the late fusion strategy, i.e., the fusion only happen at the last step. This helps explain the low accuracy of the late fusion model and suggest that, in this case, fusion at earlier stages is a better choice.

4.4 Compared with the Perceptual Score

In this subsection, we compare our SHAPE scores with the perceptual score as well as with its fine-grained versions.

Firstly, we observe that the perceptual score P coincides with our marginal contribution SHAPE score $\mathcal{S}$ in the sense that their values are highly correlated. In the B-T4SA dataset, both $\mathcal{S}_V$ and P_V are close to zero, and $\mathcal{S}_T$ (P_T) is significantly higher than $\mathcal{S}_V$ (P_V). Similar pattern is found in the CMU-MOSEI dataset. For the SNLI dataset, $\mathcal{S}$ and P are both considerably higher than zero, except for the Late Fusion case. Besides, both scores for T_h modality are higher than the ones for T_p modality.

Despite the similarities, we emphasize the superiority of our SHAPE scores for (i) guaranteed by the game theory, the SHAPE score $\mathcal{S}$ can correctly capture the marginal contribution no matter how many modalities are involved. (ii) the SHAPE score $\mathcal{C}$ can directly quantify the cooperation among modalities in addition to their individual importance.

Next, we turn to the fine-grained perceptual scores. As argued in section 3.3, the raw perceptual score implicitly measures the correlation among modalities in the dataset. However, the positive and negative correlations are mixed and cancel each other, causing a considerable loss of information. To be concrete, let's consider the Late Fusion case for the SNLI dataset. With only the raw perceptual scores, we at best might conclude that the T_h modality is more important than the T_p one in the contribution to model performance. On the other hand, we find $P_{T_h}^{in}$ is much smaller than $P_{T_h}^{out}$, indicating that the model prediction almost relies solely on the T_h

modality. The small values of $P_{T_p}^{in}$ and $P_{T_p}^{out}$ further consolidate this conclusion. Putting it differently, the Late Fusion model tends to use only the T_h modality while ignoring the T_p modality. Consequently, the cooperation between the two modalities should be small in the Late Fusion case. This is just what the cooperation SHAPE score $C_{T_pT_h}$ reveals to us. Under more complicated scenario with three or more modalities, it would be harder to capture the cross-modal cooperation using perceptual scores. On the contrary, it is always straightforward to quantify the cross-modal cooperation with the SHAPE score $\mathcal{C}$.

5 Discussion

The above experiment results suggest that the way a model exploits multi-modal information depends on various factors, including the intrinsic correlation among individual modalities in the dataset, the type of the task, and the fusion strategy in use. Based on this observation, we want to draw attention to dataset and task in multi-modal learning research. There is little hope to train a model that can exploit cross-modal cooperation on a dataset where different modalities are nearly independent. Likewise, if the task can be easily handled with only a single modality, one must try hard to design an algorithm to exploit the possible cross-modal cooperation. Furthermore, the bias and unbalance in dataset also affect multi-modal learning. In many cases, the improvement in the accuracy of a multi-modal model could be illusive. Our SHAPE score $\mathcal{C}$ can directly quantify the cooperation among individual modalities and, hence, could help diagnose the fusion strategy.

At last, it is worth noting that though the present SHAPE scores are defined using the accuracy metric as the utility function, there is no limitation of what function should be used. In fact, the SHAPE scores can readily adapt to more complex metrics such as the F1 score.

6 Conclusion

In this paper, we present the SHAPE scores that quantify the marginal contribution of individual modalities and their collaboration. Properties of the Shapely value provides theoretical guarantee for the validity of our SHAPE scores and the experiments empirically verify their effectiveness. We would like to suggest using the SHAPE scores in addition to the accuracy metric when analyzing a multi-modal model, as they are easy to compute and can extract key information about how the model use multi-modal data.

We are aware that the present SHAPE scores are not yet perfect. The choice of baseline values deserve more deliberate consideration. In the future work, we will keep improving the SHAPE scores and systematically evaluate the SOTA multi-modal models.

Acknowledgments

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[Bai *et al.*, 2021] Chongyang Bai, Haipeng Chen, Srijan Kumar, Jure Leskovec, and V. S. Subrahmanian. M2P2: Multimodal Persuasion Prediction using Adaptive Fusion. *arXiv preprint arXiv:2006.11405*, December 2021.

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[Bowman *et al.*, 2015] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

[Delbrouck *et al.*, 2020] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955*, 2020.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Gat *et al.*, 2021] Itai Gat, Idan Schwartz, and Alex Schwing. Perceptual score: What data modalities does your model perceive? *Advances in Neural Information Processing Systems*, 34, 2021.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[Hessel and Lee, 2020] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! *arXiv preprint arXiv:2010.06572*, 2020.

[Kiela *et al.*, 2019] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[Liang *et al.*, 2021] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. *arXiv preprint arXiv:2107.07502*, November 2021.

[Liu *et al.*, 2019] Fei Liu, Jing Liu, Zhiwei Fang, Richang Hong, and Hanqing Lu. Densely connected attention flow for visual question answering. In *IJCAI*, pages 869–875, 2019.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

[Molnar, 2020] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[Rodrigues Makiuchi *et al.*, 2019] Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda. Multimodal fusion of bert-cnn and gated cnn representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 55–63, 2019.

[Shapley, 1953] LS Shapley. The value of n-person games. *Annals of Math. Studies*, 28:307–317, 1953.

[Tiwari, 2010] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1):19–22, 2010.

[Vadicamo *et al.*, 2017] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-media learning for image sentiment analysis in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 308–317, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2020] Weiyao Wang, Du Tran, and Matt Feiszli. What Makes Training Multi-Modal Classification Networks Hard? *arXiv preprint arXiv:1905.12681*, April 2020.

[Wang, 2021] Yang Wang. Survey on deep multi-modal data analytics: collaboration, rivalry, and fusion. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1s):1–25, 2021.

[Weber, 1988] Robert J Weber. Probabilistic values for games. *The Shapley Value. Essays in Honor of Lloyd S. Shapley*, pages 101–119, 1988.

[Zadeh *et al.*, 2018] 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.