What Makes Multimodal Learning Better than Single (Provably)

Anonymous Author(s)
Affiliation
Address
email

Abstract

The world provides us with data of multiple modalities. Intuitively, models fusing data from different modalities outperform unimodal models, since more information is aggregated. Recently, joining the success of deep learning, there is an influential line of work on deep multimodal learning, which has remarkable empirical results on various applications. However, theoretical justifications in this field are notably lacking. *Can multimodal provably perform better than unimodal?*

In this paper, we answer this question under a most popular multimodal learning framework, which firstly encodes features from different modalities into a common latent space and seamlessly maps the latent representations into the task space. We prove that learning with multiple modalities achieves a smaller population risk than only using its subset of modalities. The main intuition is that the former has more accurate estimate of the latent space representation. To the best of our knowledge, this is the first theoretical treatment to capture important qualitative phenomena observed in real multimodal applications. Combining with experiment results, we show that multimodal learning does possess an appealing formal guarantee.

1 Introduction

The world provides us with multimodal data, which allows us to build multi-sensor systems to perceive it. As deep learning develops, more and more multimodal applications have sprung up like mushrooms after a spring rain. For example, PixelPlayer [34] learns to find the sound source from the image and separate the sounds into different components which can represent the sound from each pixel. [16] designs a reliable and efficient end-to-end learnable 3D object detector to perform accurate localization by exploiting both LIDAR as well as cameras. CLIP [22] demonstrates that simple pre-training task between text and image is an efficient way to learn SOTA image representations and benchmarks over 30 different vision datasets.

Deep multimodal learning shows great power in practice. However, from a theoretical standpoint, our understanding towards deep multimodal learning is extremely limited. Recently, there has been a research aiming to obtain the principled understanding of multimodal learning in theory. [25] provides theoretical guarantee for an information theory based approach proposed for semi-supervised multimodal learning. Besides, while very little is known about the theory of multimodal learning, there is a close cousin of the setting, called multi-view learning [30], which already has comprehensive theoretical underpinnings. Yet, two common principles, *consensus* and *complement*, which are crucial in the theoretical analysis of multi-view learning, are not applicable in multimodal learning. Toward this end, the following fundamental problem remains largely open:

Can multimodal learning provably performs better than unimodal?

In this paper, we provably answer this question from two perspectives:

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

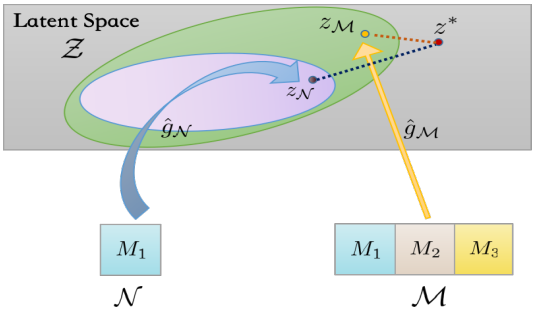


Figure 1: $\mathcal{M}$ ***vs*** $\mathcal{N}$ **modalities latent space representation**, where the latter is a subset of the former. $z_{\mathcal{M}}$, $z_{\mathcal{N}}$ and $z^\star$ are images on the latent space $\mathcal{Z}$ corresponding to the representation mappings $\hat{g}_{\mathcal{M}}$, $\hat{g}_{\mathcal{N}}$ and $g^\star$. M_i denotes modality i.

- (When) Under what conditions multimodal performs better than unimodal?
- (Why) What results in the performance gains ?

We study a widely used composite multimodal framework [35], which can seamlessly perform latent space learning and task-specific learning in a unified framework. Specifically, we firstly encode the complex data from heterogeneous sources into a common latent space $\mathcal{Z}$. The true latent representation is $g^\star$ in a function class $\mathcal{G}$, and the task mapping $h^\star$ is contained in a function class $\mathcal{H}$ defined on the latent space. Our model corresponds to recent progress of deep multimodal learning on various applications, such as video classification [28] and action recognition [14].

Under this framework, we provide the first theoretical analysis to shed light on what makes multimodal outperform unimodal. We identify the relationship between the population risk and the distance between a learned latent representation $\hat{g}$ and the $g^\star$, under the metric we will define later. Informally, closer to the true representation leads to less population loss, which indicates that a better latent representation guarantees the end-to-end multimodal learning performance. Compared to simply considering the comparison of *multi vs uni* modalities, we consider a general case, $\mathcal{M}$ *vs* $\mathcal{N}$ modalities, which are distinct subsets of all modalities. We focus on the condition that the latter is a subset of the former. Our second result is a bound for the closeness between $\hat{g}$ and the $g^\star$, from which we provably show that the latent representation $\hat{g}_{\mathcal{M}}$ learning from the $\mathcal{M}$ modalities is closer to the true $g^\star$ than $\hat{g}_{\mathcal{N}}$ learning from $\mathcal{N}$ modalities. As shown in Figure 1, $\hat{g}_{\mathcal{M}}$ has a more sufficient latent space exploration than $\hat{g}_{\mathcal{N}}$. Moreover, in a specific linear regression model, we directly verify that using multiple modalities rather than its subset learns a better latent representation.

The main contributions of this paper are summarized as follows:

- We formalize the multimodal learning problem into a theoretical framework. Firstly, we show that the performance of multimodal learning in terms of population risk can be bounded by the *latent representation quality*, a novel metric we propose to measure the distance from a learned latent representation to the true representation, which reveals that the ability of learning the whole task coincides with the ability of learning the latent representation when we have sufficient training samples.
- We derive an upper bound for the latent representation quality of training over a subset of modalities. This directly implies a principle to guide us in modality selection, i.e., when the number of sample size is large and multiple modalities can efficiently optimize the empirical risk, using multimodal to build a recognition or detection system can have a better performance.
- Restricted to linear latent and task mapping, we provide rigorous theoretical analysis that latent representation quality degrades when the subset of multiple modalities is applied. Experiments are also carried out to empirically validate the theoretical observation that $\hat{g}_{\mathcal{N}}$ is inferior to $\hat{g}_{\mathcal{M}}$.

The rest of the paper is organized as follows. In the next section, we review the related literature. The formulation of multimodal learning problem is described in Section 3. Main results are presented

in Section 4. In Section 5, we show simulation results to support our theoretical claims. Finally, conclusions are drawn in Section 6.

2 Related Work

Multimodal Learning Deep learning makes fusing different signals easier, which enable us to develop many multimodal frameworks. For example, [24, 13, 11, 18, 10] combine RGB and depth images to improve semantic segmentation; [6, 9] fuse audio with video to do scene understanding; researchers also explore audio-visual source separation and localization [34, 8]. In semi-supervised setting, [25] proposed a novel method, Total Correlation Gain Maximization (TCGM), based on information theory, which explores the information intersection by maximizing the total correlation gain among all the modalities. They theoretically proved that TCGM can find the groundtruth Bayesian classifier given each modality.

Multi-view Learning Similar to multimodal learning, multi-view learning also utilizes data from diverse sources. The main distinction is that multi-view generally refers to different manifestations of the same object, like multilingual text while different modalities may represent different objects but with connections, such as text and corresponding audio and video. The problem why multi-view has the advantage over single view has been well understood in theory. [1] derived a generalization bound for classification with multiple artificially generated views, which identify a a trade-off between the number of views, the size of the training set, the number of views, and the quality of the view generating functions. Following the complementary principle [30], information from each single view can complement each other to improve the generalization ability of the whole learner on different tasks [12, 2, 32, 33].

Transfer Learning A line of work closely related to our composite learning framework is transfer learning via representation learning, which firstly learns a shared representation on various tasks and then transfers the learned representation to a new task. [26, 5, 20, 7] have provided the sample complexity bounds in the special case of the linear feature and linear task mapping. [27] introduces a new notion of task diversity and provides a generalization bound with general tasks, features, and losses. Unfortunately, the function class which contains feature mappings is the same across all different tasks while our focus is that the function classes generated by different subsets modalities are usually inconsistent.

Notation: Throughout the paper, we use $\|\cdot\|$ to denote the ℓ_2 norm. We also denote the set of positive integer numbers less or equal than n by $[n]$, i.e. $[n] \triangleq \{1, 2, \cdots, n\}$.

3 The Multimodal Learning Formulation

In this section, we present the Multimodal Learning problem formulation. Specifically, we assume that a given data $\mathbf{x} := \left(x^{(1)}, \cdots, x^{(K)}\right)$ consists of K modalities, where $x^{(k)} \in \mathcal{X}^{(k)}$ the domain set of the k-th modality. Denote $\mathcal{X} = \mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(K)}$. We use $\mathcal{Y}$ to denote the target domain and use $\mathcal{Z}$ to denote a latent space. Then, we denote $g^\star : \mathcal{X} \mapsto \mathcal{Z}$ the true mapping from the input space (using all of K modalities) to the latent space, and $h^\star : \mathcal{Z} \mapsto \mathcal{Y}$ is the true task mapping. For instance, in aggregation-based multimodal fusion, $g^\star$ is an aggregation function compounding on K seperate sub-networks and $h^\star$ is a multi-layer neural network [29].

In the learning task, a data pair $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$ is generated from an unknown distribution $\mathcal{D}$, such that

$$\mathbb{P}_{\mathcal{D}}(\mathbf{x}, y) \triangleq \mathbb{P}_{y|\mathbf{x}}\left(y \mid h^\star \circ g^\star(\mathbf{x})\right) \mathbb{P}_{\mathbf{x}}(\mathbf{x}) \tag{1}$$

Here $h^\star \circ g^\star(\mathbf{x}) = h^\star(g^\star(\mathbf{x}))$ represents the composite function of $h^\star$ and $g^\star$.

In real-world settings, we often face incomplete multimodal data, i.e., some modalities are not observed. To take into account this situation, we let $\mathcal{M}$ be a subset of $[K]$, and without loss of generality, focus on the learning problem only using the modalities in $\mathcal{M}$. Specifically, define $\mathcal{X}' := \left(\mathcal{X}^{(1)} \cup \{\perp\}\right) \times \ldots \times \left(\mathcal{X}^{(K)} \cup \{\perp\}\right)$ as the extension of $\mathcal{X}$, where $\mathbf{x}' \in \mathcal{X}'$, $\mathbf{x}'_k = \perp$ means

that the k-th modality is not used (collected). Then we define a mapping $p_{\mathcal{M}}$ from $\mathcal{X}$ to $\mathcal{X}'$ induced by $\mathcal{M}$:

$$p_{\mathcal{M}}(\mathbf{x})^{(k)} = \begin{cases} \mathbf{x}^{(k)} & \text{if } k \in \mathcal{M} \\ \perp & \text{else} \end{cases}$$

Also define $p'_{\mathcal{M}} : \mathcal{X}' \mapsto \mathcal{X}'$ as the extension of $p_{\mathcal{M}}$. Let $\mathcal{G}'$ denote a function class, which contains the mapping from $\mathcal{X}'$ to the latent space $\mathcal{Z}$, and define a function class $\mathcal{G}_{\mathcal{M}}$ as follows:

$$\mathcal{G}_{\mathcal{M}} \triangleq \{g_{\mathcal{M}} : \mathcal{X} \mapsto \mathcal{Z} \mid g_{\mathcal{M}}(\mathbf{x}) := g'(p_{\mathcal{M}}(\mathbf{x})), g' \in \mathcal{G}'\} \tag{2}$$

Given a data set $\mathcal{S} = ((\mathbf{x}_i, y_i))_{i=1}^m$, where $(\mathbf{x}_i, y_i)$ is drawn i.i.d. from $\mathcal{D}$, the learning objective is, following the Empirical Risk Minimization (ERM) principle [17], to find $h \in \mathcal{H}$ and $g_{\mathcal{M}} \in \mathcal{G}_{\mathcal{M}}$ to jointly minimize the empirical risk, i.e.,

$$\min \qquad \hat{r}\left(h \circ g_{\mathcal{M}}\right) \triangleq \frac{1}{m} \sum_{i=1}^{m} \ell\left(h \circ g_{\mathcal{M}}(\mathbf{x}_i), y_i\right) \tag{3}$$

$$\text{s.t.} \qquad h \in \mathcal{H}, g_{\mathcal{M}} \in \mathcal{G}_{\mathcal{M}}. \tag{4}$$

where $\ell(\cdot, \cdot)$ is the loss function. Given $\hat{r}\left(h \circ g_{\mathcal{M}}\right)$, we similarly define its corresponding population risk as

$$r\left(h \circ g_{\mathcal{M}}\right) = \mathbb{E}_{(\mathbf{x}_i, y_i) \sim \mathcal{D}}\left[\hat{r}\left(h \circ g_{\mathcal{M}}\right)\right] \tag{5}$$

Similar to [1, 27], we use the population risk as to measure the performance of learning.

Example. As a concrete example of our model, consider the video classification problem under the late-fusion model in [28]. In this case, each modality k, e.g. RGB frames, audio or optical flows, is encoded by a deep network φ_k, and their features are fused and passed to a classifier $\mathcal{C}$. If we train on the first M modalities, we can let $\mathcal{M} = [M]$. Then $g_{\mathcal{M}}$ has the form: $\varphi_1 \oplus \varphi_2 \oplus \cdots \oplus \varphi_M$, where $\oplus$ denotes a fusion operation, e.g. self-attention ($\mathcal{Z}$ is the output of $g_{\mathcal{M}}$), and h is the classifier $\mathcal{C}$.

4 Main Results

In this section, we provide main theoretical results to rigorously establish various aspects of the folklore claim that multimodal is better than single. We first detail several assumptions throughout this section.

Assumption 1. *The loss function $\ell(\cdot, \cdot)$ is L-smooth with respect to the first coordinate, and is bounded by a constant C.*

Assumption 2. *The true latent representation $g^\star$ is contained in $\mathcal{G}$, and the task mapping $h^\star$ is contained in $\mathcal{H}$.*

Assumption 1 is a classical regularity condition for loss function in theoretical analysis works [17, 27, 26]. Assumption 2 is also be known as realizability condition in representation learning [27, 7, 26], which ensures that the function classes that we optimize over contains the true latent representation and the task mapping.

Assumption 3. *For any $g' \in \mathcal{G}'$ and $\mathcal{M} \subset [K]$, $g' \circ p'_{\mathcal{M}} \in \mathcal{G}'$.*

To understand Assumption 3, note that for any $\mathcal{N} \subset \mathcal{M} \subset [K]$, by definition, for any $g_{\mathcal{N}} \in \mathcal{G}_{\mathcal{N}}$, there exists $g' \in \mathcal{G}'$, s.t.

$$g_{\mathcal{N}}(\mathbf{x}) = g'(p_{\mathcal{N}}(\mathbf{x})) = g'(p'_{\mathcal{N}}(p_{\mathcal{M}}(\mathbf{x})))$$

Therefore, Assumption 3 directly implies $g_{\mathcal{N}} \in \mathcal{G}_{\mathcal{M}}$. Moreover, we have $\mathcal{G}_{\mathcal{N}} \subset \mathcal{G}_{\mathcal{M}} \subset \mathcal{G}$, which means that the inclusion relationship of modality subsets remains unchanged on the latent function class induced by them. As an example, if $\mathcal{G}'$ is linear, represented as matrices $\mathbf{G} \in \mathbb{R}^{Q \times K}$. Also $p'_{\mathcal{M}}$ can be represented as a diagonal matrix $\mathbf{P} \in \mathbb{R}^{K \times K}$ with the i-th diagonal entry being 1 for $i \in \mathcal{M}$ and 0 otherwise. In this case, Assumption 3 holds, i.e. $\mathbf{G} \times \mathbf{P} \in \mathcal{G}'$. Moreover, $\mathbf{G} \times \mathbf{P}$ is a matrix with i-th column all be zero for $i \notin \mathcal{M}$, which is commonly used in the underfitting analysis in linear regression [23].

4.1 Connection to Latent Representation Quality

Latent space is employed to better exploit the correlation among different modalities. Therefore, we will naturally conjecture that the performance of training with different modalities is related to their ability to learn latent space representation. In this section, we will formally characterize this relationship.

In order to measure the goodness of a learned latent representation g, we introduce the following definition of *latent representation quality*.

Definition 1. Given a data distribution with the form in (1), for any learned latent representation mapping $g \in \mathcal{G}$, the **latent representation quality** is defined as

$$\eta(g) = \inf_{h \in \mathcal{H}} \left[r\left(h \circ g\right) - r(h^* \circ g^*) \right] \tag{6}$$

Here $\inf_{h \in \mathcal{H}} r\left(h \circ g\right)$ is the best achievable population risk with the fixed latent representation g. Thus, to a certain extent, $\eta(g)$ measures the loss incurred by the distance between g and $g^\star$.

Next, we recap the Rademacher complexity measure for model complexity. It will be used in quantifying the the population risk performance based on different modalities. Specifically, let $\mathcal{F}$ be a class of vector-valued function $\mathbb{R}^d \mapsto \mathbb{R}^n$. Let $Z_1, \ldots, Z_m$ be i.i.d. random variables on $\mathbb{R}^d$ following some distribution P. Denote the sample $S = (Z_1, \ldots, Z_m)$. The empirical Rademacher complexity of $\mathcal{F}$ with respect to the sample S is given by [3]

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g\left(Z_i\right) \right]$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)^\top$ with $\sigma_i \overset{iid}{\sim}$ unif $\{-1, 1\}$. The Rademacher complexity of $\mathcal{F}$ is

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_S \left[\widehat{\mathfrak{R}}_S(\mathcal{F}) \right]$$

Now we present our first main result regarding multimodal learning.

Theorem 1. *Let $\mathcal{S} = ((\mathbf{x}_i, y_i))_{i=1}^m$ be a dataset of m examples drawn i.i.d. according to $\mathcal{D}$. Let $\mathcal{M}, \mathcal{N}$ be two distinct subsets of $[K]$. Assuming we have produced the empirical risk minimizers $(\hat{h}_\mathcal{M}, \hat{g}_\mathcal{M})$ and $(\hat{h}_\mathcal{N}, \hat{g}_\mathcal{N})$, training with the $\mathcal{M}$ and $\mathcal{N}$ modalities separately. Then, for all $1 > \delta > 0$, with probability at least $1 - \frac{\delta}{2}$:*

$$\begin{aligned} & r\left(\hat{h}_\mathcal{M} \circ \hat{g}_\mathcal{M}\right) - r\left(\hat{h}_\mathcal{N} \circ \hat{g}_\mathcal{N}\right) \\ & \leq \gamma_\mathcal{S}(\mathcal{M}, \mathcal{N}) + 8L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_\mathcal{M}) + \frac{4C}{\sqrt{m}} + 2C\sqrt{\frac{2\ln(2/\delta)}{m}} \end{aligned} \tag{7}$$

where

$$\gamma_\mathcal{S}(\mathcal{M}, \mathcal{N}) \triangleq \eta(\hat{g}_\mathcal{M}) - \eta(\hat{g}_\mathcal{N}) \qquad \square \tag{8}$$

Remark. A few remarks are in place. First of all, $\gamma_\mathcal{S}(\mathcal{M}, \mathcal{N})$ defined in (8) compares the quality between latent representations learning from $\mathcal{M}$ and $\mathcal{N}$ modalities with respect to the given dataset $\mathcal{S}$. Theorem 1 bounds the difference of population risk training with two different subsets of modalities by $\gamma_\mathcal{S}(\mathcal{M}, \mathcal{N})$, which validates our conjecture that including more modalities is advantageous in learning. Second, for the commonly used function classes in the field of machine learning, Radamacher complexity for a sample of size m, $\mathfrak{R}_m(\mathcal{F})$ is usually bounded by $\sqrt{C(\mathcal{F})/m}$, where $C(\mathcal{F})$ represents the intrinsic property of function class $\mathcal{F}$. Third, (7) can be written as $\gamma_\mathcal{S}(\mathcal{M}, \mathcal{N}) + \mathcal{O}(\sqrt{\frac{1}{m}})$ in order terms. This shows that as the number of sample size grows, the performance of using different modalities mainly depends on its latent representation quality.

4.2 Upper Bound for Latent Space Exploration

Having establish the connection between the population risk difference with latent representation quality, our next goal is to estimate how close the learned latent representation $\hat{g}_\mathcal{M}$ is to the true latent representation $g^\star$. The following theorem shows how the latent representation quality can be controlled in the training process.

Theorem 2. *Let $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ be a dataset of m examples drawn i.i.d. according to $\mathcal{D}$. Let $\mathcal{M}$ be a subset of $[K]$. Assuming we have produced the empirical risk minimizers $(\hat{h}_{\mathcal{M}}, \hat{g}_{\mathcal{M}})$ training with the $\mathcal{M}$ modalities. Then, for all $1 > \delta > 0$, with probability at least $1 - \delta$:*

$$\eta(\hat{g}_{\mathcal{M}}) \leq 4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_{\mathcal{M}}) + 4L\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}) + 6C\sqrt{\frac{2\ln(2/\delta)}{m}} + \hat{L}(\hat{h}_{\mathcal{M}} \circ \hat{g}_{\mathcal{M}}, \mathcal{S}) \tag{9}$$

where $\hat{L}(\hat{h}_{\mathcal{M}} \circ \hat{g}_{\mathcal{M}}, \mathcal{S}) \triangleq \hat{r}\left(\hat{h}_{\mathcal{M}} \circ \hat{g}_{\mathcal{M}}\right) - \hat{r}\left(h^\star \circ g^\star\right)$ is the centered empirical loss. $\square$

Remark. Consider sets $\mathcal{N} \subset \mathcal{M} \subset [K]$. Under Assumption 3, $\mathcal{G}_{\mathcal{N}} \subset \mathcal{G}_{\mathcal{M}} \subset \mathcal{G}$, optimizing over a larger function class results in a smaller empirical risk. Therefore

$$\hat{L}(\hat{h}_{\mathcal{M}} \circ \hat{g}_{\mathcal{M}}, \mathcal{S}) \leq \hat{L}(\hat{h}_{\mathcal{N}} \circ \hat{g}_{\mathcal{N}}, \mathcal{S}) \tag{10}$$

Similar to the analysis in Theorem 1, the first term on the Right-hand Side (RHS), $\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_{\mathcal{M}}) \sim \sqrt{C(\mathcal{H} \circ \mathcal{G}_{\mathcal{M}})/m}$ and $\mathfrak{R}_m(\mathcal{H} \circ \mathcal{G}_{\mathcal{N}}) \sim \sqrt{C(\mathcal{H} \circ \mathcal{G}_{\mathcal{N}})/m}$. Following the basic structural property of Radamacher complexity [3], we have $C(\mathcal{H} \circ \mathcal{G}_{\mathcal{N}}) \leq C(\mathcal{H} \circ \mathcal{G}_{\mathcal{M}})$. Therefore, Theorem 2 offers the following principle for choosing modalities to improve the latent representation quality.

Principle: *choose to learn with more modalities if:*

$$\hat{L}(\hat{h}_{\mathcal{N}} \circ \hat{g}_{\mathcal{N}}, \mathcal{S}) - \hat{L}(\hat{h}_{\mathcal{M}} \circ \hat{g}_{\mathcal{M}}, \mathcal{S}) \geq \sqrt{\frac{C(\mathcal{H} \circ \mathcal{G}_{\mathcal{M}})}{m}} - \sqrt{\frac{C(\mathcal{H} \circ \mathcal{G}_{\mathcal{N}})}{m}}$$

What this principle implies are twofold. (i) When the number of simple size m is large, the impact of intrinsic complexity of function classes will be reduced. (ii) Using more modalities can efficiently optimize the empirical risk, hence improve the latent representation quality.

Through the trade-off illustrated in the above principle, we provide theoretical evidence that when $\mathcal{N} \subset \mathcal{M}$ and training samples are sufficient, $\eta(\hat{g}_{\mathcal{M}})$ may be less than $\eta(\hat{g}_{\mathcal{N}})$, i.e.$\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N}) \leq 0$. Moreover, combining with the conclusion from Theorem 1, if the sample size m is large enough, $\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N}) \leq 0$ guarantees $r\left(\hat{h}_{\mathcal{M}} \circ \hat{g}_{\mathcal{M}}\right) \leq r\left(\hat{h}_{\mathcal{N}} \circ \hat{g}_{\mathcal{N}}\right)$, which indicates learning with the $\mathcal{M}$ modalities outperforms only using its subset $\mathcal{N}$ modalities.

4.3 Non-Positivity Guarantee

In this section, we focus on a composite linear data generating model to theoretically verify that the $\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N})$ is indeed non-positive in this special case.[1]Specifically, we consider the case where the mapping to the latent space and the task mapping are both linear. Formally, let the function class $\mathcal{G}$ and $\mathcal{H}$ be:

$$\begin{aligned} \mathcal{G} &= \left\{g \mid g(\mathbf{x}) = \mathbf{A}^\top \mathbf{x}, \mathbf{A} \in \mathbb{R}^{d \times n}, \mathbf{A}\right\} \\ \mathcal{H} &= \left\{h \mid h(\mathbf{z}) = \boldsymbol{\beta}^\top \mathbf{z}, \boldsymbol{\beta} \in \mathbb{R}^n, \|\boldsymbol{\beta}\| \leq C_b\right\} \end{aligned} \tag{11}$$

where $\mathbf{x} = \left(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(K)}\right)$ is a d-dimensional vector, $\mathbf{x}^{(k)} \in \mathbb{R}^{d_k}$ denotes the feature vector for the k-th modality and $\sum_{k=1}^K d_k = d$. Here, the distribution $\mathbb{P}_{\mathbf{x}}(\cdot)$ satisfies that its covariance matrix is positive definite. The data is generated by:

$$y = \left(\boldsymbol{\beta}^\star\right)^\top \mathbf{A}^{\star\top} \mathbf{x} + \epsilon \tag{12}$$

where r.v. ϵ is independent of $\mathbf{x}$ and has zero-mean and bounded second moment. Note that in practical multimodal learning, usually only one layer is linear and the other is a neural network. For instance, [35] employs the linear matrix to project the feature matrix from different modalities into a common latent space for early dementia diagnosis, i.e., $\mathcal{G}$ is linear. Another example is in pedestrian detection [31], where a linear task mapping is adopted, i.e., $\mathcal{H}$ is linear. Thus, our composite linear model can be viewed as an approximation to such popular models, and our results can offer insights into the performance of these models.

We consider a special case that $\mathcal{M} = [K]$ and $\mathcal{N} = [K-1]$. Thus $\mathcal{G}_{\mathcal{M}} = \mathcal{G}$ and we have the following result.

$$\mathcal{G}_{\mathcal{N}} = \left\{g \mid g(\mathbf{x}) = \begin{bmatrix} \mathbf{A}_{1:\sum_{k=1}^{K-1} d_k} \\ \mathbf{0} \end{bmatrix}^\top \mathbf{x}, \mathbf{A} \in \mathbb{R}^{d \times n} \text{ with orthonormal columns} \right\} \tag{13}$$

[1]Proving that $\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N}) \leq 0$ holds in general is open and will be an interesting future work.

Proposition 1. *Consider the dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ generating from the linear model defined in (12) with ℓ_2 loss. Let $\mathcal{M} = [K]$ and $\mathcal{N} = [K-1]$. Let $\hat{\mathbf{A}}_{\mathcal{M}}$, $\hat{\mathbf{A}}_{\mathcal{N}}$ denote the projection matrix estimated by $\mathcal{M}$, $\mathcal{N}$ modalities. Assume that $\hat{\mathbf{A}}_{\mathcal{M}}$, $\mathbf{A}^\star$ has orthonormal columns. If $n = d$, for sufficiently large constant C_b, we have:*

$$\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N}) \leq 0 \tag{14}$$

In this special case, Proposition 1 directly guarantees that training with incomplete modalities weakens the ability to learn a optimal latent representation. As a result, it also degrades the learning performance.

5 Experiment

We conduct experiments to validate our theoretical results. The source of the data we consider is two-fold, multimodal real-world dataset and dataset generated by well-designed simulators. All experiments are based on open source machine learning library PyTorch [19].

5.1 Real-world dataset

Dataset. The natural dataset we use is the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, which is an acted multimodal and multispeaker database [4]. It contains three modalities, Text, Video and Audio. We follow the data preprocessing method of [21] and obtain 100 dimensions data for audio, 100 dimension for text, and 500 dimension for video. There are six labels here, namely, happy, sad, neutral, angry, excited and frustrated. We use 13200 data for training and 3410 for testing.

Training Setting. For all experiments on IEMOCAP, we use one linear neural network layer and ReLU as the activation function to extract the latent feature, and we set the hidden dimension to be 128. In Unimodal model, we directly map the latent feature to the output space, while in Multimodal model, we concatenate the features first, and then do the mapping. Especially, any modal feature extractor does not share parameters. We use Adam [15] as the optimizer and we set the learning rate to be 0.01, other hyper-parameters default. The batch size is 2048 for the data and does not take up lots of memory. For this classification task, the accuracy is used for performance measurement.

Connection to the Latent Representation Quality. The classification accuracy on IEMOCAP, using different combinations of modalities are summarized in Table 1. All learning strategies using multiple modalities outperform the singlemodal baseline. To validate Theorem 1, we calculate the test accuracy difference between different subsets of modalities using the result in Table 1 and show them in the third column of Table 2.

Moreover, we empirically evaluate the $\eta(\hat{g}_{\mathcal{M}})$ in the following way: freeze the encoder $\hat{g}_M$ obtained through pretraining and then finetune a better classifier h. Having $\eta(\hat{g}_{\mathcal{M}})$ and $\eta(\hat{g}_{\mathcal{N}})$, the values of $\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N})$ between different subsets of modalities are also presented in Table 2. Previous discussions on Theorem 1 implies that the population risk difference between $\mathcal{M}$ and $\mathcal{N}$ modalities has the same sign as $\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N})$ when the sample size is large enough, and negativity implies performance gains. Since we use accuracy as the measure, on the contrary, positivity indicates a better performance in our settings. As shown in Table 2, when more modalities are added for learning, the test accuracy difference and $\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N})$ are both positive, which confirms the important role of latent representation quality characterized in Theorem 1.

Upper Bound for Latent Space Exploration. Table 2 also confirms our theoretical analysis in Theorem 2. In all cases, the $\mathcal{N}$ modalities is a subset of $\mathcal{M}$ modalities, and correspondingly, a positive $\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N})$ is observed. This indicates that $\mathcal{M}$ modalities has a more sufficient latent space exploration than its subset $\mathcal{N}$ modalities.

We also attempt to understand the use of sample size for exploring the latent space. Table 3 presents the latent representation quality η obtained by using different numbers of sample size, which is measured by the test accuracy of the pretrain+finetuned modal. Here, the ratio of sample size is set to the total number of training samples. The corresponding curve is also ploted in Figure 2(a). As

Table 1: Test classification accuracy on IEMOCAP, using different combinations of modalities, only Text, Text + Video, Text + Audio and Text + Video + Audio.

Modalities	Test Acc
Text(T)	49.93±0.57
Text + Video(TV)	51.08±0.66
Text + Audio(TA)	53.03±0.21
Text + Video + Audio(TVA)	**53.89 ± 0.47**

Table 2: Comparison of test accuracy and latent representation quality among different combinations of modalities.

$\mathcal{M}$ Modalities	$\mathcal{N}$ Modalities	Test Acc **Difference**	$\gamma_{\mathcal{S}}(\mathcal{M}, \mathcal{N})$
TV	T	1.15	1.36
TA	T	3.10	3.57
TVA	TA	0.86	0.19
TVA	TV	2.81	2.4

the number of sample size grows, the increase in performance of η is observed, which is in keeping with the $\mathcal{O}(\sqrt{1/m})$ term in our upper bound for η. The phenomenon that the combination of Text, Video and Audio (TVA) modalities underperforms the unimodal when the number of sample size is relatively small, can also be interpreted by the trade-off we discussed in Theorem 2. When there are insufficient training examples, the intrinsic complexity of the function class induced by multiple modalities dominates, thus weakening its latent representation quality.

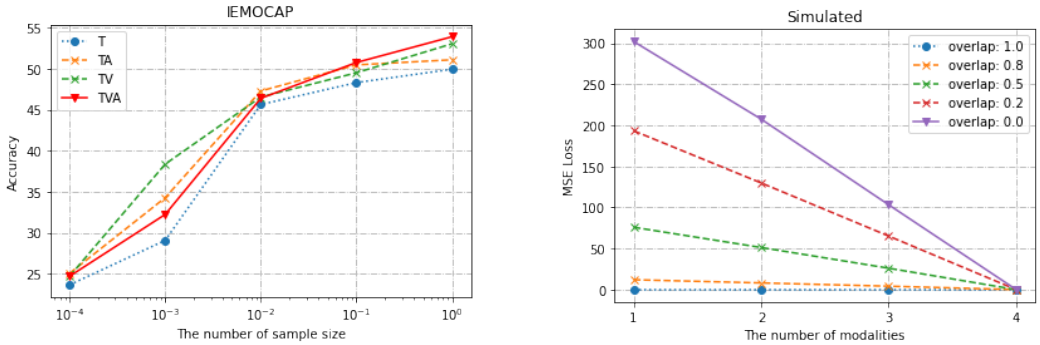
Table 3: Latent representation quality vs. The number of the sample size on IEMOCAP

Modalities	Test Acc (Ratio of Sample Size)				
	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1
T	23.66±1.28	29.08±3.34	45.63±0.29	48.30±1.31	49.93±0.57
TA	**25.06±1.05**	34.28±4.54	**47.28±1.24**	50.46±0.61	51.08±0.66
TV	24.71±0.87	**38.37±3.12**	46.54±1.62	49.50±1.04	53.03±0.21
TVA	24.71±0.76	32.24±1.17	46.39±3.82	**50.75±1.45**	**53.89±0.47**

5.2 Dataset Generated by Simulation

In this subsection, we investigate the effect of modality correlation on latent representation quality. Typically, there are three situations for the correlation across each modality [25]. (i) Each modality does not share information at all, that is, each modality only contains modal-specific information. (ii) The other is that, all modalities only maintain the share information without unique information on their own. (iii) The last is a moderate condition, i.e., each modal not only shares information, but also owns modal-specific information. The reason to utilize simulated data in this section is due to the fact that it is hard in practice to have natural datasets that possess the required degree of modality correlation.

Data Generation. We firstly generate modality one by sampling from a normal distribution with mean 0 and variance 1, where each dimension is unrelated. And then weight modality one and an independent sampling as a new modality, so that every new modality share some same information and have the specific information. The weight of modality one can represent the degree of overlap. We tune it in $\{0.0, 0.2, 0.5, 0.8, 1.0\}$. Here 1.0 means totally overlap and 0.0 is the totally specific scenario. Each modality has 100 dimension and the label has 1 dimension, and we do a regression task in this experiment. The training set contains 10000 data and the testing set contains 3000 data. We use 4 modalities totally, denoted by $\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}$, respectively.



(a) Latent representation quality vs.The ratio of the sample size on IEMOCAP

(b) Latent representation quality vs. The number of modalities on simulated data

Figure 2: Evaluation of the latent representation quality on different data sets

Training Setting. We use the multi-layer perceptron neural network as our model. To be more specific, we first use a Linear Layer to encode the input to a 10-dimension latent space and then we map the latent feature to the output space. The first layer's input dimension depends on the number of modality. For example, if we use two modalities, the input dimension is 200. We use SGD as our optimizer, and we set a learning rate 0.01, momentum 0.9, batch size 10000, and train for 10000 steps. The Mean Square Error (MSE) loss is considered for evaluation in this regression problem.

η vs Modality Correlation Our aim is to discover the influence of modality correlation on the latent representation quality η. To this end, Table 4 shows the η with the varying number of modalities under different correlation conditions, which is measured by the MSE loss of the pretrain+finetuned modal. The trend that the loss decreases as the number of modalities increases is described in Figure 2(b), which also validates our analysis of Theorem 2. Moreover, Figure 2(b) shows that higher correlation among modalities achieves a lower loss for η, which means a better latent representation. This emphasizes the role of latent space to exploit the intrinsic correlations among different modalities.

Table 4: Latent representation quality among different correlation situations on simulated data

Modalities	MSE Loss (Degree of Overlap)				
	1	0.8	0.5	0.2	0.0
$\mathbf{1}$	0	12.04±0.39	75.89±1.28	193.28±1.08	301.92±7.85
$\mathbf{1,2}$	0	8.16±0.17	51.25±1.06	129.81±4.36	207.45±4.68
$\mathbf{1,2,3}$	0	4.18±0.05	26.06±0.69	65.17±1.52	103.23±0.61
$\mathbf{1,2,3,4}$	0	0	0	0	0

6 Conclusions

In this work, we provably show that learning with multiple modalities is superior to employing its subset of modalities, since the former has access to a better latent space representation. The results answer the two questions: when and why multimodal outperforms unimodal jointly. To the best of our knowledge, this is the first theoretical treatment to capture important qualitative phenomena observed in real multimodal applications. Also our theoretical results show that as the sample size increases, the overall ability to mapping the input features from different modalities to a nearly optimal latent state is more crucial, which points to an interesting direction that are worth further investigation: to find which encoder is the bottleneck and focus on improving it. More importantly, our work provides new insights for future multimodal theoretical research.

References

[1] Massih-Reza Amini, Nicolas Usunier, Cyril Goutte, et al. Learning from multiple partially observed views-an application to multilingual text categorization. In *NIPS*, volume 22, pages 28–36, 2009.

[2] Raman Arora, Poorya Mianjy, and Teodor Marinov. Stochastic optimization for multiview representation learning using partial least squares. In *International Conference on Machine Learning*, pages 1786–1794. PMLR, 2016.

[3] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[5] Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11:2901–2934, 2010.

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[7] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[8] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[10] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer, 2016.

[11] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444. IEEE, 2019.

[12] Tomoharu Iwata and Makoto Yamada. Multi-view anomaly detection via robust probabilistic latent variable models. In *NIPS*, pages 1136–1144, 2016.

[13] Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*, 2018.

[14] M Esat Kalfaoglu, Sinan Kalkan, and A Aydin Alatan. Late temporal modeling in 3d cnn architectures with bert for action recognition. In *European Conference on Computer Vision*, pages 731–747. Springer, 2020.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 641–656, 2018.

[17] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[18] Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 4980–4989, 2017.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[20] Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pages 55–76. PMLR, 2013.

[21] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 873–883, 2017.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[23] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

[24] Daniel Seichter, Mona Köhler, Benjamin Lewandowski, Tim Wengefeld, and Horst-Michael Gross. Efficient rgb-d semantic segmentation for indoor scene analysis. *arXiv preprint arXiv:2011.06961*, 2020.

[25] Xinwei Sun, Yilun Xu, Peng Cao, Yuqing Kong, Lingjing Hu, Shanghang Zhang, and Yizhou Wang. Tcgm: An information-theoretic framework for semi-supervised multi-modality learning. In *European Conference on Computer Vision*, pages 171–188. Springer, 2020.

[26] Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

[27] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12692–12702. IEEE, 2020.

[29] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33, 2020.

[30] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[31] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5363–5371, 2017.

[32] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[33] Yang Yang, De-Chuan Zhan, and Yuan Jiang. Learning by actively querying strong modal features. In *IJCAI*, pages 2280–2286, 2016.

[34] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

[35] Tao Zhou, Kim-Han Thung, Mingxia Liu, Feng Shi, Changqing Zhang, and Dinggang Shen. Multi-modal latent space inducing ensemble svm classifier for early dementia diagnosis with neuroimaging data. *Medical image analysis*, 60:101630, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 5.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] The code and the data are proprietary.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]