RECOGNIZING AND OVERCOMING THE GREEDY NATURE OF LEARNING IN MULTI-MODAL DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We hypothesize that due to the greedy nature of learning in multi-modal deep neural networks (DNNs), these models tend to rely on just one modality while under-utilizing the other modalities. We observe empirically that such behavior hurts its overall generalization. We validate our hypothesis by estimating the gain on the accuracy when the model has access to an additional modality. We refer to this gain as the conditional utilization rate of the modality. In the experiments, we consistently observe an imbalance in conditional utilization rate between modalities, across multiple tasks and architectures. Since conditional utilization rate cannot be computed efficiently during training, we introduce an efficient proxy based on the pace at which a DNN learns from each modality, which we refer to as conditional learning speed. We thus propose a training algorithm, *balanced multi-modal learning*, and demonstrate that it indeed addresses the issue of greedy learning. The proposed algorithm is found to improve the model's generalization on three datasets: Colored MNIST (Kim et al., 2019), Princeton ModelNet40 (Wu et al., 2015), and NVIDIA Dynamic Hand Gesture Dataset (Molchanov et al., 2016).

1 INTRODUCTION

In real-world problems, each instance frequently has multiple modalities associated with it. For example, we detect cancer in both X-ray and ultrasound images. We seek clues from images to answer questions given in text. We are naturally interested in training deep neural networks (DNNs) end-to-end to learn from all available input modalities. We refer to such a training regimen as a *multi-modal learning process* and DNNs resulting from this process as *multi-modal DNNs*.

Several recent studies have reported unsatisfactory performance of multimodal DNN's in various tasks (Wang et al., 2020a; Wu et al., 2020; Gat et al., 2020; Cadene et al., 2019; Agrawal et al., 2016; Hessel & Lee, 2020; Han et al., 2021). Such as in visual question answering, multi-modal DNNs were found to ignore the visual modality (presented as an image) and exploit statistical regularities shared between the textual modality (presented as a question) and the answer alone, resulting in poor generalization (Cadene et al., 2019; Gat et al., 2020; Agrawal et al., 2016). In human action detection task, multi-modal DNNs trained on images and audio were observed to under-perform uni-modal DNNs trained on images only (Wang et al., 2020a).

These earlier negative findings compel us to ask, *what prevents multi-modal DNNs from achieving better performance?* In order to answer this question, we first introduce a metric, *conditional utilization rate*. For a multi-modal DNN trained with two modalities, m_0 and m_1, the conditional utilization rate of m_1 given m_0, denoted by $u(m_1|m_0)$, measures how important it is to use m_1, given the presence of m_0. We experiment on several multi-modal learning tasks and we consistently observe a significant difference of conditional utilization rates between modalities. For example, we observe $u_{RGB|depth} = 0.01$ and $u_{depth|RGB} = 0.63$ for a model trained to identify gestures in videos using the NVIDIA Dynamic Hand Gesture Dataset (NVGesture) (Molchanov et al., 2016). It indicates that the model relies on the depth modality to make predictions and does not pay attention to the RGB modality. This observation leads to a conjecture that the multi-modal learning process often results in models that under-utilize some of the input modalities.

As the focal point of this work, we put forward the greedy learner hypothesis. Under the greedy learner hypothesis, a multi-modal DNN learns to rely on one of the input modalities, based on which it could learn faster, and does not continue to learn to use the other modalities. This greediness leads to imbalance in conditional utilization rate between modalities. In other words, it prevents a multi-modal DNN from learning to exploit all available modalities and often results in worse generalization. It explains the above observed phenomenon, answers the challenges in training multi-modal DNNs, and motivates us to design a better multi-modal learning algorithm.

According to the greedy learner hypothesis, it is the speed at which a multi-modal DNN learns from each modality that leads to imbalance in conditional utilization rate. If we appropriately intervene in the learning process to adjust these speeds, we may be able to prevent the hurtful imbalance across input modalities. We characterize such learning speed by introducing a metric, called *conditional learning speed*. We empirically show that it is a reasonable proxy for conditional utilization rate. We introduce a training algorithm, *balanced multi-modal learning* which accelerates the model to learn from either modality alternatively according to their conditional learning speeds. We show that it produces models that learn to use all modalities appropriately and achieve stronger generalization on three multi-modal datasets: Colored MNIST dataset (Kim et al., 2019), ModelNet40 dataset of 3D objects (Su et al., 2015) and NVGesture Dataset (Molchanov et al., 2015).

2 Problem Setup

We consider two input modalities, referred to as m_0 and m_1, without loss of generality. We denote a multi-modal dataset by $\mathcal{D} = \left\{(\boldsymbol{x}^i, y^i)\right\}_{i=1}^{N}$, where $\boldsymbol{x} = (\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1})$. We partition the dataset $\mathcal{D}$ into training, validation and test sets, denoted by $\mathcal{D}^{\text{train}}$, $\mathcal{D}^{\text{val}}$, and $\mathcal{D}^{\text{test}}$, respectively. The goal is then to train a multi-modal DNN on this datset.

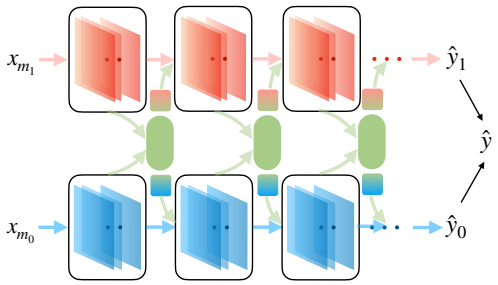


Figure 1: The multi-modal DNN with intermediate fusion that we study in this work. Blue and red blocks represent the two uni-modal networks, learning from modalities m_0 and m_1. The fusion layers are marked with green, green-blue, and green-red. The predictions from the two uni-modal branches are denoted by $\hat{y}_0$ and $\hat{y}_1$.

Multi-modal DNNs We use a multi-modal DNN that has two *uni-modal branches*, each taking one modality as input. The two uni-modal branches are interconnected by layer-wise fusion modules. These fusion modules enable information flow from one branch to another. This network architecture employs intermediate fusion, according the categorization of multi-modal fusion strategy in the deep learning literature (Ngiam et al., 2011; Atrey et al., 2010; Baltrušaitis et al., 2018). It has demonstrated competitive performance against multi-modal DNNs with late fusion in many tasks (Perez et al., 2018; Joze et al., 2020; Anderson et al., 2018; Wang et al., 2020b).

We implement the fusion module as a multi-modal transfer module (MMTM) (Joze et al., 2020). The first step in MMTM is to squeeze feature maps from each uni-modal branch to vector representations via global average pooling over spatial dimensions. Next we concatenate these representations and apply a linear transformation to obtain cross-modal context representation. We predict channel-wise weights for each modality based this context representation through two independent fully-connected layers. Finally, these weights are used to re-calibrate the channel-wise features per modality.

We train a multi-modal DNN f on $\mathcal{D}^{\text{train}}$. Let $\hat{y}$ be the prediction of f for an input $\boldsymbol{x}$:

$$\hat{y} = f(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}).$$

As shown in Figure 1, $\hat{y} = \frac{1}{2}(\hat{y}_0 + \hat{y}_1)$, where $\hat{y}_0$ and $\hat{y}_1$ are the outputs of the two uni-modal branches. During training, the parameters of the multi-modal DNN are updated by stochastic gradient descent (SGD) to minimize the loss: $L = \text{CE}(y, \hat{y}_0) + \text{CE}(y, \hat{y}_1)$, where CE stands for cross-entropy. We

refer to each of the cross-entropy losses as a modality-specific loss. We train the model until the highest accuracy on $\mathcal{D}^{\text{val}}$ is reached.

During training, the uni-modal branch largely focuses on the associated modality. The fusion modules generate cross-modal context information from the uni-modal branches and pass it back to them. Both $\hat{y}_0$ and $\hat{y}_1$ depend on information from both modalities. We end up with two functions, f_0 and f_1, corresponding to the two uni-modal branches:

$$\hat{y}_0 = f_0(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}), \quad \hat{y}_1 = f_1(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}).$$

By shutting down information sharing between branches, we derive f'_0 and f'_1, each taking only one of the input modalities as input. We describe how we derive them in detail in Appendix A.1. We use $\hat{y}'_0$ and $\hat{y}'_1$ to refer to their outputs:

$$\hat{y}'_0 = f'_0(\boldsymbol{x}_{m_0}), \quad \hat{y}'_1 = f'_1(\boldsymbol{x}_{m_1}).$$

In other words, we derive four models from f. We can now define *conditional utilization rate* using these four derived models.

We first define the ability of each model as its accuracy A on $\mathcal{D}^{\text{test}}$. $A(f)$ is the accuracy reached by the original multi-modal DNN f and is the multi-modal DNN's performance. Based on the symmetry between two modalities in our multi-modal DNN, we group the four accuracies into two pairs: $A(f_0)$ and $A(f'_0)$, and $A(f_1)$ and $A(f'_1)$. The conditional utilization rate is then the relative change of the accuracy between the two within each pair:

Definition 2.1 (Conditional utilization rate) *For a multi-modal DNN, f, taking two modalities m_0 and m_1 as inputs, its conditional utilization rates for m_1 and m_0 are defined as $\boldsymbol{u}(m_1|m_0) = \frac{A(f_0)-A(f'_0)}{A(f_0)}$ and $\boldsymbol{u}(m_0|m_1) = \frac{A(f_1)-A(f'_1)}{A(f_1)}$, respectively.*

The conditional utilization rate of m_0, $u(m_0|m_1)$, measures how important it is for f to use m_0 together with m_1. In other words, it estimates how much m_0 contribute to $\hat{y}_1$.

A multi-modal learning process Given the model architecture and the dataset, described above, a multi-modal learning process includes the following steps: (1) sample a value of learning rate randomly from a given range, (2) initialize the DNN's parameters randomly, (3) update them to fit $\mathcal{D}^{\text{train}}$ with SGD, and (4) select the model checkpoint with the highest accuracy on $\mathcal{D}^{\text{val}}$.

3 THE GREEDY LEARNER HYPOTHESIS

In this section, we introduce the greedy learner hypothesis to explain challenges observed in training multi-modal DNNs. Before describing our hypothesis, we start by discussing some assumptions on the multi-modal data and preliminary observations made in the literature on multi-modal learning.

3.1 ASSUMPTIONS IN MULTI-MODAL LEARNING

First, in multi-modal learning literature (Sridharan & Kakade, 2008), both modalities are assumed to be predictive of the target, i.e., $I(\boldsymbol{Y}, \boldsymbol{X}_{m_0}) > 0$ and $I(\boldsymbol{Y}, \boldsymbol{X}_{m_1}) > 0$, where I denotes the mutual information. In order to minimize one of the modality-specific losses, e.g. $\text{CE}(y, \hat{y}_0)$, one can either update the parameters of the fusion layers that pass information from m_1 to $\hat{y}_0$, or the parameters of the uni-modal branch taking m_0 as input, or both.

Second, the modalities are predictive of the target at different degrees. For example, we have observed that when training DNNs on each modality separately, they usually do not reach the same performance (Wang et al., 2020b; Joze et al., 2020). In addition, multi-modal DNNs exhibit varying performance when being trained on different combinations of modalities present for the task (Weng et al., 2021; Pérez-Rúa et al., 2019; Liu et al., 2018).

Third, DNNs trained with only one of the modalities learn at different speeds. Multi-modal DNNs have been observed to learn from different modalities at different speeds as well (Wang et al., 2020a; Wu et al., 2020).

3.2 MULTI-MODAL LEARNING PROCESS IS GREEDY

Before formalizing our hypothesis and verifying it experimentally, we first demonstrate that we can tell whether the model utilizes all modalities by analyzing its conditional utilization rates.

Given a multi-modal DNN, f, with two input modalities, m_0 and m_1, we compute the difference between conditional utilization rates, $\boldsymbol{u}(m_1|m_0)$ and $\boldsymbol{u}(m_0|m_1)$ as

$$\mathrm{d}_{\mathrm{util}}(f) = \boldsymbol{u}(m_1|m_0) - \boldsymbol{u}(m_0|m_1).$$

If we observe high $|\,\mathrm{d}_{\mathrm{util}}(f)|$, we say that f exhibits *imbalance in utilization between modalities*.

The difference between conditional utilization rates is bounded between -1 and 1. When it is close to -1 or 1, the model benefits only from one of the modalities given the other but not vice versa. This implies that the model's ability to predict $\hat{y}_0$ and $\hat{y}_1$ comes only from one of the modalities.

We now introduce the *greedy learner hypothesis*:

> We call a multi-modal learning process *greedy*, when it trains models to rely on only one of the available modalities. A greedy multi-modal learning process cannot avoid producing models that exhibit a high degree of imbalance in utilization between modalities. We *hypothesize* that a multi-modal learning process, in which a multi-modal DNN is trained to minimize the sum of modality-specific losses, is greedy.

Let $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ denote the expectation of $\widehat{\mathrm{d}}_{\mathrm{util}}$ over the empirical distribution of models resulting from a multi-modal learning process (as defined in §2). The absolute value of $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ is associated with the greediness of the process. The higher the $|\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]|$ is, the greedier the learning process is.

We propose the following *conjectures* consistent with the greedy learner hypothesis:

1. There exists an $\epsilon > 0$, s.t. $|\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]| > \epsilon$, as long as the multi-modal dataset has the last two properties from §3.1. Otherwise, $\widehat{\mathrm{d}}_{\mathrm{util}}$ is distributed symmetrically around zero and $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}] = 0$.
2. The stronger the regularization we apply in training, the higher the $|\mathbb{E}[\mathrm{d}_{\mathrm{util}}]|$ is.

We expect the first scenario in the first conjecture to be true for all real-world multi-modal learning tasks. We construct datasets by duplicating one of the input modalities to test the second scenario in the first conjecture.

The second conjecture is motivated by the discussion on the implicit regularization of DNNs (Valle-Perez et al., 2019; Zhang et al., 2017; 2021; Smith et al., 2021). It refers to a consensus that DNNs have the tendency to lean towards simplicity during training and it explains their strong generalization. We suspect that the regularization of DNNs is a strong factor impacting the greediness of the multi-modal learning process.

In order to validate the conjectures and the hypothesis above empirically, we conduct experiments on several multi-modal datasets using different network architectures (cf. §5.1). We present our empirical observations in §5.2 and §5.3.

4 MAKING MULTI-MODAL LEARNING LESS GREEDY

According to the greedy learner hypothesis, we can make multi-modal learning less greedy by controlling the speed at which a multi-modal DNN learns to rely on each modality. To this end, we derive *conditional learning speed* to measure the speed at which the DNN learns from one modality. It serves as an efficient proxy to the conditional utilization rate of the corresponding modality. We then propose the *balanced multi-modal learning* algorithm, which controls model's conditional learning speed between modalities in order to prevent it from being greedy.

4.1 CONDITIONAL LEARNING SPEED

As demonstrated in §3.2, the imbalance in conditional utilization rates is a sign of the model exploiting the connection between the target and only one of the input modalities, ignoring cross-modal information. Conditional utilization rate was however can be measured once training was done,

making it difficult to use it in real-time during training. We instead derive a proxy metric, called *conditional learning speed*, that captures relative learning speed between modalities during training.

Let us revisit the architecture of the multi-modal DNN. We denote the parameters of each uni-modal branch taking m_0 and m_1 as inputs, by $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Layers of the fusion modules marked with green and green/blue in Figure 1, are part of the function mapping $\boldsymbol{x}$ to $\hat{y}_0$. Let $\boldsymbol{\theta}'_0$ refer to their parameters. Accordingly, we have the layers marked with green and green/red in Figure 1 that are part of the function mapping $\boldsymbol{x}$ to $\hat{y}_1$. We denote their parameters by $\boldsymbol{\theta}'_1$. In summary, we divide the model's parameters into two pairs: $(\boldsymbol{\theta}_0, \boldsymbol{\theta}'_0)$ and $(\boldsymbol{\theta}_1, \boldsymbol{\theta}'_1)$.[1]

We define the model's conditional learning speed based on these two pairs of parameters:

Definition 4.1 (Conditional learning speed) *Given a multi-modal DNN, f, with two input modalities, m_0 and m_1, the conditional learning speeds of these modalities after t training steps, are*

$$\boldsymbol{s}(m_1|m_0;t) = \log \frac{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}'_0; i)}{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}_0; i)}, \quad \boldsymbol{s}(m_0|m_1;t) = \log \frac{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}'_1; i)}{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}_1; i)},$$

where $\mu(\boldsymbol{\theta};\cdot) = ||\boldsymbol{G}||_2^2/||\boldsymbol{\theta}||_2^2$ quantifies the impact on the model f that comes from the change of $\boldsymbol{\theta}$ at one update step and $\boldsymbol{G} = \partial L/\partial \boldsymbol{\theta}$.

This definition of $\mu(\boldsymbol{\theta};\cdot)$ is inspired by discussion on the effective update of parameters in the literature (Van Laarhoven, 2017; Zhang et al., 2019; Brock et al., 2021). Previous studies found that when normalization techniques are applied, such as batch normalization (Ioffe & Szegedy, 2015), the update of the direction of $\boldsymbol{\theta}$, i.e., $\boldsymbol{\theta}/||\boldsymbol{\theta}||_2^2$, reflects how much the update on $\boldsymbol{\theta}$ changes the model f in order to fit the batch of samples.

The conditional learning speed, $\boldsymbol{s}(m_1|m_0;t)$ (and analogously $\boldsymbol{s}(m_0|m_1;t)$), is the log-ratio between the learning speed of $\boldsymbol{\theta}'_0$ and that of $\boldsymbol{\theta}_0$. Because $\boldsymbol{\theta}'_0$ carries information from m_1 to $\hat{y}_0$, $\boldsymbol{s}(m_1|m_0;t)$ reflects how fast the model learns from m_1 relative to m_0 learned through the corresponding uni-modal branch, after the first t steps.

We further compute the difference between $\boldsymbol{s}(m_1|m_0;t)$ and $\boldsymbol{s}(m_0|m_1;t)$ as:

$$\mathrm{d}_{\mathrm{speed}}(f;t) = \boldsymbol{s}(m_1|m_0;t) - \boldsymbol{s}(m_0|m_1;t),$$

analogous to $\mathrm{d}_{\mathrm{util}}$. If the greedy learner hypothesis is true, the conditional learning speed would serve as a proxy for the conditional utilization rate. $\mathrm{d}_{\mathrm{speed}}(f;t)$ would predict $\mathrm{d}_{\mathrm{util}}(f)$ at the end of training. In §5.2 and §5.3, we show the distributions of $\widehat{\mathrm{d}}_{\mathrm{speed}}$ and $\widehat{\mathrm{d}}_{\mathrm{util}}$ side-by-side to validate this. For each model, we report $\mathrm{d}_{\mathrm{speed}}(f;T)$ as $\widehat{\mathrm{d}}_{\mathrm{speed}}(f)$ where the model takes T steps until reaching the highest accuracy on $\mathcal{D}^{\mathrm{val}}$.

4.2 Balanced multi-modal learning

Because $\mathrm{d}_{\mathrm{speed}}(f,t)$ is predictive of the imbalanced utilization between modalities, we can take advantage of $\mathrm{d}_{\mathrm{speed}}(f,t)$ to balance conditional utilization on-the-fly. In addition to training the network normally with both modalities, we accelerate the model to learn from either modality alternatively to balance the conditional learning speeds of them. See Algorithm 1 for an overall description of the algorithm.

We refer to the training steps at which we perform forward and backward passes normally as *regular steps*. We introduce *re-balancing steps* at which we force the model to update only one of the uni-modal branches in order to accelerate learning from the associated input modality. See Appendix A.2 for the full explanation of the re-balancing step.

To warm-up the model, we perform regular steps in the first epoch. We switch from regular steps to re-balancing steps if $|\,\mathrm{d}_{\mathrm{speed}}(t)| > \alpha$, where α is the *imbalance parameter*. The training takes Q re-balancing steps before returning to regular mode. We refer to Q as the *re-balancing window size*.

[1] This is not partitioning, as some parameters are both in $\boldsymbol{\theta}'_0$ and $\boldsymbol{\theta}'_1$.

Algorithm 1: Balanced multi-modal learning

Input: Q, re-balancing window size;
α, imbalance parameter
T, # of training steps;
T_1, # of steps in the 1th training epoch

"'Let M denote the accumulated effective update in regular steps"'
$M_{\boldsymbol{\theta}_0}, M_{\boldsymbol{\theta}'_0} = 0, 0$;
$M_{\boldsymbol{\theta}_1}, M_{\boldsymbol{\theta}'_1} = 0, 0$;
for $t \leftarrow 1$ **to** T_1 **do**
  Take a regular step;
  Update $M_{\boldsymbol{\theta}_0}, M_{\boldsymbol{\theta}'_0}, M_{\boldsymbol{\theta}_1}, M_{\boldsymbol{\theta}'_1}$
end
regular_mode = True
for $t \leftarrow T_1$ **to** T **do**
  if *regular_mode* **then**
    Take a regular step;
    Update $M_{\boldsymbol{\theta}_0}, M_{\boldsymbol{\theta}'_0}, M_{\boldsymbol{\theta}_1}, M_{\boldsymbol{\theta}'_1}$;
    Compute $\mathrm{d}_{\mathrm{speed}}$;
    if $|\mathrm{d}_{\mathrm{speed}}| > \alpha$ **then**
      $q = 0$; regular_mode = False
    end
  else
    $q \leftarrow q + 1$;
    if $q = Q$ **then** regular_mode = True;
    Take a re-balancing step to accelerate learning from m_0 **If** $\mathrm{d}_{\mathrm{speed}} > 0$ **Else** m_1;
  end
end

5 Experiments and results

5.1 Datasets and tasks

Colored-and-gray-MNIST (Kim et al., 2019) is a synthetic dataset based on MNIST (LeCun et al., 1998). In the training set of 60,000 examples, each example has two images, a gray-scale image and a monochromatic image, with color strongly correlated with its digit label. For the validation set of 10,000 examples, each example also has a gray-scale and a corresponding monochromatic image however with a low correlation between the color and its label. We consider the monochromatic image as the first modality m_0 and the gray-scale one as the second modality m_1. We use a neural network with four convolutional layers as the uni-modal branch and employ three MMTMs (Joze et al., 2020) to connect them. The corresponding uni-modal DNNs trained on the monochromatic images and the gray-scale images achieve the validation accuracies of 43% and 98%, respectively. We use this synthetic dataset mainly to show that the proposed greedy learner hypothesis happens in practice.

ModelNet40 is one of the Princeton ModelNet datasets (Wu et al., 2015) with 3D objects of 40 categories (9,483 training samples and 2,468 test samples). We use the task in which the goal is to classify a 3D object based on the 2D views of its front and back (Su et al., 2015). Each example is a 3D object rendered as 2D images of 224×224 pixels. For the uni-modal branches, we use ResNet18 (He et al., 2016) and apply MMTMs in the three final residual blocks. The uni-modal DNNs achieve 91.6% and 93.1% accuracy, when learning from the front view (m_0) and the rear view (m_1), respectively.

NVIDIA Dynamic Hand Gesture Dataset, or **NVGesture** in short (Molchanov et al., 2015), consists of 1,532 video clips (1,050 training and 482 test ones) of hand gestures in 25 classes. We sample 20% training samples as the validation set and use depth and RGB as two modalities. We adopt the configuration by Joze et al. (2020) for data preparation and use I3D Carreira & Zisserman (2017) as uni-modal branches and MMTMs as fusion modules in the six final inception modules.

We provide examples of each dataset and details on data preprocessing in Appendix A.4

5.2 Validating the greedy learner hypothesis

In this section, we test the conventional multi-modal learning process on seven tasks with different input and output pairs to demonstrate the first conjecture in §3.2 experimentally and validate the proposed greedy learner hypothesis.

Study design For each task introduced in §5.1, we construct a dataset where we duplicate one of the modalities as the two input modalities, in addition to the original dataset. For example, we predict

the digit class using two identical gray-scale images in the case of MNIST. We train a multi-modal DNN on each dataset of each task as explained below:

- Colored-and-gray-MNIST: we train multi-modal DNNs using SGD with the momentum coefficient of 0.9 and a batch size of 128. We sample uniformly 20 learning rates at random from the interval $[10^{-5}, 1]$ on a logarithmic scale. We train each model four times while varying random seeds. In total, we train 80 models.
- ModelNet40: we use SGD without momentum and use minibatches of size 8. We randomly select nine learning rates from 10^{-3} to 1. We train each model three times and end up with 27 models.
- NVGesture: we use a batch size of four, SGD with momentum of 0.9, and uniformly sample 20 learning rates at random from the interval $[10^{-4}, 10^{-1.5}]$ on a logarithmic scale. We train each model three times, resulting in 60 models in total.

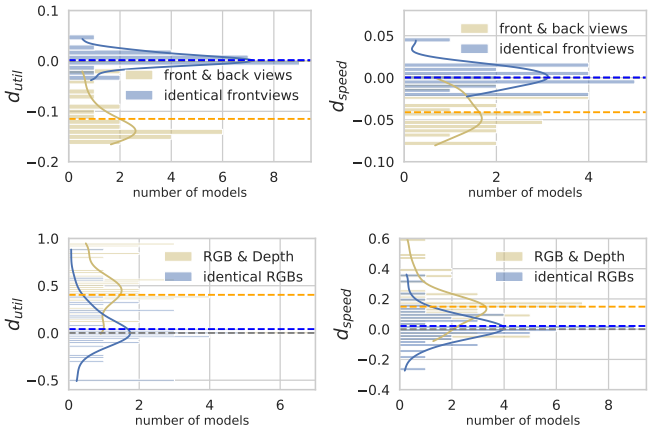


Figure 2: Histograms and estimated density functions of $\widehat{\mathrm{d}}_{\mathrm{util}}$ and $\widehat{\mathrm{d}}_{\mathrm{speed}}$ of models trained using ModelNet40 (top) or NVGesture (bottom). We mark zero and $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ with dashed lines. Many models have large $\widehat{\mathrm{d}}_{\mathrm{util}}$. We see $\widehat{\mathrm{d}}_{\mathrm{util}}$ is distributed asymmetrically around zero when using two different input modalities. When using two identical input modalities, it is distributed symmetrically around zero. $\widehat{\mathrm{d}}_{\mathrm{speed}}$ exhibits similar distributions as $\widehat{\mathrm{d}}_{\mathrm{util}}$.

Results We present the results for the experiments on ModelNet40 and NVGesture in Figure 2 and on Colored-and-gray-MNIST in the Appendix Figure 2.

First, many models have high $|\,\mathrm{d}_{\mathrm{util}}|$, especially when trained to use two distinct modalities. This confirms that the standard multi-modal learning process encourages the model to rely on one modality and ignore the other one, which is consistent with our hypothesis. We make this observation across all tasks, confirming that the conventional multi-modal learning process is greedy regardless of network architectures and tasks.

Second, $\widehat{\mathrm{d}}_{\mathrm{util}}$ is distributed symmetrically centered at zero, and $|\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]|$ is approximately 0.0, for all the experiments using two identical input modalities. On the other hand, if we use two distinct modalities, $\widehat{\mathrm{d}}_{\mathrm{util}}$ is distributed asymmetrically, and we observe positive $|\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]|$'s of approximately 0.3, 0.1 and 0.4 for colored-and-gray-MNIST, ModelNet40 and NVGesture, respectively. This validates first conjecture in §3.2.

Third, we draw the same set of conclusions even when we consider conditional learning speed $\widehat{\mathrm{d}}_{\mathrm{speed}}$ instead of conditional utilization rate. In fact, the distributions of $\widehat{\mathrm{d}}_{\mathrm{speed}}$ largely replicate the observed distributions of $\widehat{\mathrm{d}}_{\mathrm{util}}$. It validates our greedy learner hypothesis which blames the varying rate at which the learner learns to rely on each modality. It moreover confirms $\mathrm{d}_{\mathrm{speed}}$ is an appropriate proxy to use to re-balance multi-modal learning.

5.3 Strong regularization encourages greediness

Regularization is known to impact the learning speed when training a deep neural net with SGD. Here, we investigate the effect of regularization on multi-modal learning. We demonstrate that, as predicted in the last conjecture 3.2, strong regularization encourages the multi-modal learning process to be greedy.

Study design We apply L1 regularization to the multi-modal DNNs. That is, we train the networks to optimize the loss $L' = L + \lambda||\boldsymbol{\theta}||_1$, where L is the classification loss, $\boldsymbol{\theta}$ stands for all model parameters and λ is the weight on the regularizer.

We measure the effect of λ on the network f by computing the fraction of its parameters smaller than 10^{-7}. We denote this by $R(f)$. Since L1 regularization encourages sparsity of the network's parameters, as shown in Appendix Figure 3, the larger the λ is, the higher the $R(f)$.

We conduct this study with ModelNet40, using the front and rear views. We choose ten values from an interval $[10^{-9}, 10^{-3}]$ as λ. We use SGD without momentum, set the learning rate to 0.1 and batch size to eight. For each combination of hyperparameters, we train three model repetitions.

Results As shown in Figure 3, $|\,\mathrm{d}_{\mathrm{util}}(f)|$ increases along λ, especially when $\log(\lambda) \geq -5$ and $|\,\mathrm{d}_{\mathrm{util}}(f)|$ is positively correlated with $R(f)$. In other words, the stronger the regularization, the larger the imbalance in utilization between modalities we obverse. It confirms the second conjecture in §3.2, which states that the stronger regularization we apply, the greedier the multi-modal learning process becomes. We see $|\,\mathrm{d}_{\mathrm{speed}}|$ follows the same trend as $|\,\mathrm{d}_{\mathrm{util}}|$. Again, it supports our choice of using the conditional learning speed to predict the conditional utilization rate.

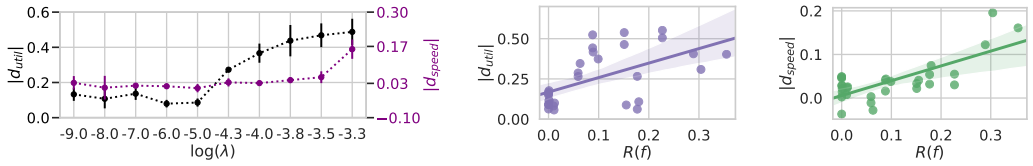


Figure 3: The observed $|\,\mathrm{d}_{\mathrm{util}}|$ and $|\,\mathrm{d}_{\mathrm{speed}}|$ for models trained with varying weights (λ) on the L1 regularizer. *Left*: $|\,\mathrm{d}_{\mathrm{util}}|$ and $|\,\mathrm{d}_{\mathrm{speed}}|$ as a function of $\log(\lambda)$. *Middle*: $|\,\mathrm{d}_{\mathrm{util}}|$ increases along $R(f)$. *Right*: $|\,\mathrm{d}_{\mathrm{speed}}|$ increases along $R(f)$. The multi-modal learning process exhibit higher greed if we introduce stronger regularization on the multi-modal DNNs.

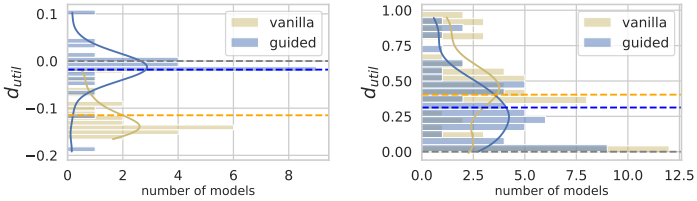
5.4 Balanced multi-modal learning

Besides the proposed training algorithm (cf. §4.2, referred to as *guided*)), we introduce a variant of it in Appendix A.3, referred to as *random*. This algorithm is motivated by Modality Dropout (Neverova et al., 2015) but better suited to the multi-modal DNNs with intermediate fusion. We consider it a stronger baseline which can also balance learning from modalities. In this section we compare the guided, the random and the standard training algorithm (referred to as *vanilla*) on the ModelNet40 dataset (front and rear views), the NVGesture dataset (RGB and depth), and the Colored-and-gray-MNIST dataset (colored and gray-scale images).

Calibrated modality utilization We train multi-modal DNNs as described in §5.2, using each training algorithm. We set the imbalance parameter α to 0.01 and the re-balancing window size Q to 5 for ModelNet40, and α to 0.1 and Q to 5 for NVGesture. Results are shown in Figure 4. Models trained with the guided algorithm have lower $|\widehat{\mathrm{d}}_{\mathrm{util}}|$ compared to the vanilla algorithm. On NVGesture, we obtain $\mathbb{E}(\widehat{\mathrm{d}}_{\mathrm{util}})$ of approximately 0.31 and 0.40 for models trained with the guided and the vanilla algorithm. On ModelNet40, we obtain $\mathbb{E}(\widehat{\mathrm{d}}_{\mathrm{util}})$ of approximately -0.02, and -0.12 for models trained with the guided and the vanilla algorithm. We would expect greater changes from the guided algorithm if we tune α and Q to find the optimal values for each task.

The random algorithm also calibrates modality utilization effectively, as shown in Appendix Figure 4.

Improved generalization performance We compare the generalization ability of the three algorithms (guided, random and vanilla). For each algorithm, we train three repetitions of each model using the same learning rate: 0.01, 0.1 and 0.01 for Colored-and-gray-MNIST, ModelNet40 and NVGesture respectively, given results of the preliminary experiments. We use α of 0.1 for Colored-and-gray-MNIST and NVGesture and α of 0.01 for ModelNet40. We set Q of 5 for all three datasets. For NVGesture, we add one experiment where we initialize the model with parameters pre-trained



(a) ModelNet40 (b) NVGesture

Figure 4: Histograms and estimated density functions of $\widehat{\mathrm{d}}_{\mathrm{util}}$ obtained for models trained using the *guided* and the *vanilla* algorithm. We mark zero and $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ with dashed lines. The guided algorithm is less greedy than the vanilla one.

using the Kinetics dataset (Carreira & Zisserman, 2017) in addition to random initialization. We refer to this setting as "NVGesture-pretrained" and to the other one as "NVGesture-scratch".

We report means and standard deviations of the models' test accuracy in Table 1.[2] The guided algorithm improves the models' generalization performance over the vanilla algorithm in all four cases. It also outperforms the random algorithm, with the exception of ModelNet40, where their performances are very close.

Table 1: Test accuracy of models trained with the vanilla, random and guided algorithms.

	Colored-and-gray-MNIST	ModelNet40	NVGesture-scratch	NVGesture-pretrained
vanilla	45.26±0.46	90.09±0.58	79.81±1.14	83.20±0.21
random	74.07±2.75	**91.36±0.10**	79.81±1.14	82.64±0.84
guided	**91.01±1.20**	91.25±0.36	**80.22±0.73**	**83.82±1.45**

6 Related work

Previous works on multi-modal learning focus more on architectural designs of the DNNs (Ngiam et al., 2011; Tran et al., 2015; Lazaridou et al., 2015; Wang et al., 2020b; Pérez-Rúa et al., 2019). Recently, unbalanced learning dynamics and modality-wise utilization in the end-to-end trained multi-modal classifiers were discussed (Goyal et al., 2017; Han et al., 2021; Wang et al., 2020a; Gat et al., 2020; Hessel & Lee, 2020; Collell & Moens, 2018). For example, Wang et al. (2020b) design a parameter-free multi-modal fusion framework that dynamically exchanges channels between the uni-modal branches. Wang et al. (2020a) estimate the uni-modal branches' generalization and overfitting speeds and calibrate the learning through loss re-weighting. Their weight estimation slows down training and relies on a representative subset of the training data. Finally Gat et al. (2020) point out the bias in the multi-modal data as the cause of the issue and propose to supply inputs with Gaussian perturbations to the model in order to regularize it by maximizing functional entropies. In this work, we explain the phenomenon by the greedy nature of learning in multi-modal DNNs and propose a calibration algorithm based on the training dynamics of model. Our proposed method does not introduce any additional computation and only relies on the norm of the gradients and weights computed during training. It does not require architectural changes of the multi-modal DNNs and can be easily modified to accommodate different fusion modules. In addition to the multi-modal classifiers, the unbalanced modality-wise utilization has been identified in the multi-modal pre-trained models (Li et al., 2020; Cao et al., 2020). We expect that the insights we provide in this work are applicable to the self-supervised multi-modal training framework too.

7 Conclusions

Our work demonstrates the fact that the end-to-end trained multi-modal DNNs rely on one of the input modalities to make predictions while leaving the other modalities underutilized. We explained this through the greedy learner hypothesis, and validated our statements experimentally on three multi-modal datasets. We illustrated that using the proposed algorithm to balance the models' learning from different modalities, enhances generalization. This result emphasizes that the adequate modality utilization is a desired property that a model should achieve in multi-modal learning.

[2]Our results on NVGesture are not directly comparable to the numbers reported in other works since we use 20% samples from the training set as the validation set.

ETHICS STATEMENT

This work focuses on analyzing the challenges of training multi-modal deep neural networks. It does not involve any human subjects. We do not anticipate any potentially harmful consequences for society due to the publication of this work. As we use publicly available datasets in our study, we do not anticipate privacy and security issues. We claim no conflicts of interest.

REPRODUCIBILITY STATEMENT

We provide complete descriptions of the used datasets and the applied processing steps in §5.1 and in Appendix §A.4. We state our experimental configurations for each study in §5.1, §5.2, and §5.3. We plan to publish the source code once the paper is released.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv:1606.07356*, 2016.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *CVPR*, 2018.

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

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.

Andrew Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. *arXiv:2102.06171*, 2021.

Remi Cadene, Corentin Dancette, Hedi Ben-Younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. *arXiv:1906.10169*, 2019.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *ECCV*, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *CVPR*, 2017.

Guillem Collell and Marie-Francine Moens. Do neural network cross-modal mappings really bridge modalities? In *ACL*, July 2018.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *NEURIPS*, 2020.

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. *CVPR*, 2017.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. *ICLR*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

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

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML*, 2015.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: multimodal transfer module for cnn fusion. *CVPR*, 2020.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. *CVPR*, 2019.

Angeliki Lazaridou, Marco Baroni, et al. Combining language and vision with a multimodal skip-gram model. In *HLT-NAACL*, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv:2012.08673*, 2020.

Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv:1805.11730*, 2018.

Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. Hand gesture recognition with 3d convolutional neural networks. *CVPR*, 2015.

Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. *CVPR*, 2016.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 2015.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. *ICML*, 2011.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. *AAAI*, 2018.

Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *CVPR*, 2019.

Samuel L Smith, Benoit Dherin, David Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *ICLR*, 2021.

Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. 2008.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. *ICCV*, 2015.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. *ICCV*, 2015.

Guillermo Valle-Perez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *ICLR*, 2019.

Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv:1706.05350*, 2017.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? *CVPR*, 2020a.

Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *NEURIPS*, 2020b.

Zejia Weng, Zuxuan Wu, Hengduo Li, and Yu-Gang Jiang. Hms: Hierarchical modality selectionfor efficient video recognition. *arXiv:2104.09760*, 2021.

Nan Wu, Stanisław Jastrzębski, Jungkyu Park, Linda Moy, Kyunghyun Cho, and Krzysztof J. Geras. Improving the ability of deep networks to use information from multiple views in breast cancer screening. *MIDL*, 2020.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. *CVPR*, 2015.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 2021.

Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. *ICLR*, 2019.

A APPENDIX

A.1 DERIVING f'_0 AND f'_1 FROM THE MULTI-MODAL DNNS

Given an input $\boldsymbol{x}$, let $h_k^{m_0}(\boldsymbol{x})$ and $h_k^{m_1}(\boldsymbol{x})$ denote the feature vectors from the uni-modal branches. Specifically, these are the feature vectors from the layer which is followed by the kth fusion modules, denoted by g_k.

To derive $f'(\boldsymbol{x}_{m_0})$ and $f'(\boldsymbol{x}_{m_1})$, we force each uni-modal branch to make predictions based on the single modality. Precisely, for each uni-modal branch, we aggregate its feature vectors over the whole training set and compute their average:

$$\overline{h}_k^{m_0} = \frac{1}{N}\sum_{\boldsymbol{x'}} h_k^{m_0}(\boldsymbol{x'}), \qquad \overline{h}_k^{m_1} = \frac{1}{N}\sum_{\boldsymbol{x'}} h_k^{m_1}(\boldsymbol{x'}).$$

Let $w_k^{m_0}(\boldsymbol{x})$ and $w_k^{m_1}(\boldsymbol{x})$ denote the context information g_k gives:

$$w_k^{m_0}(\boldsymbol{x}) = g_k(h_k^{m_0}(\boldsymbol{x}),\, \overline{h}_k^{m_1}), \qquad w_k^{m_1}(\boldsymbol{x}) = g_k(\overline{h}_k^{m_0},\, h_k^{m_1}(\boldsymbol{x})).$$

In this way, we cut off the cross-modality information sharing between the uni-modal branches while limiting the distribution shift for both the fusion networks and the layers in the uni-modal branches.

In our multi-modal DNN, we employ the fusion module as a Multi-Modal Transfer Module (MMTM). We adopt the annotation used by Joze et al. (2020). A MMTM takes two sets of tensors, $\boldsymbol{A} \in \mathbb{R}^{N_1 \times \cdots \times N_K \times C}$ and $\boldsymbol{B} \in \mathbb{R}^{M_1 \times \cdots \times M_L \times C'}$, each as feature maps of a convolutional layer of a uni-modal branch. Here, N_i and M_i represent the spatial dimensions of each feature map, and C and C' represent the number of feature maps.

In order to derive $f'(\boldsymbol{x}_{m_0})$ and $f'(\boldsymbol{x}_{m_1})$, we have operations in a MMTM written as:

$$\begin{aligned}
\overline{S}_A(c) &= \frac{1}{N_{training}} \sum_{i=1}^{N_{training}} S_A^i(c), & \overline{S}_B(c) &= \frac{1}{N_{training}} \sum_{i=1}^{N_{training}} S_B^i(c) \\
Z_A &= \boldsymbol{W}[S_A, \overline{S}_B], & Z_B &= \boldsymbol{W}[\overline{S}_A, S_B] \\
E'_A &= W_A Z_A + b_A, & E'_B &= W_B Z_B + b_B \\
\tilde{\boldsymbol{A}}' &= 2 \times \sigma(E'_A) \odot \boldsymbol{A}, & \tilde{\boldsymbol{B}}' &= 2 \times \sigma(E'_B) \odot \boldsymbol{B},
\end{aligned}$$

where $[\cdot,\cdot]$ represents the concatenation operation; W_i and b_i are parameters of the fully connected layers; $\odot$ is the channel-wise product operation; and $\sigma(\cdot)$ is the sigmoid function.

A.2 THE RE-BALANCING STEP

Suppose that we have K fusion modules and let $w_k^{m_0}(\boldsymbol{x})$ and $w_k^{m_1}(\boldsymbol{x})$ denote the output of the kth fusion module. At the t training step in the balanced multi-modal learning process, we have

$$\overline{w}_k^{m_0} = \frac{1}{N_t}\sum_{\boldsymbol{x'} \in \mathcal{D}^t_{\text{passed}}} w_k^{m_0}(\boldsymbol{x'}), \qquad \overline{w}_k^{m_1} = \frac{1}{N_t}\sum_{\boldsymbol{x'} \in \mathcal{D}^t_{\text{passed}}} w_k^{m_1}(\boldsymbol{x'}),$$

where $\mathcal{D}^t_{\text{passed}}$ denotes all samples appeared in the previous regular training steps until now.

For example, in a re-balancing step on m_0, we aim to accelerate the model's learning from m_0, and we have

$$\hat{y}_0 = \phi_0(\boldsymbol{x}_{m_0}, \overline{w}_1^{m_0}, \cdots, \overline{w}_K^{m_0}), \quad \hat{y}_1 = \phi_1(\boldsymbol{x}_{m_1}, w_1^{m_1}(\boldsymbol{x}), \cdots, \overline{w}_K^{m_1}(\boldsymbol{x})),$$

where $\boldsymbol{x} = (\boldsymbol{x_{m_0}}, \boldsymbol{x_{m_1}})$ is the input at this step, and ϕ_0 and ϕ_1 denote the uni-modal branch take x_{m_0} and x_{m_1} as inputs.

A.3 RANDOM VERSION OF THE BALANCED MULTI-MODAL LEARNING ALGORITHM

We present a *random* version of the balanced multi-modal learning algorithm in Algorithm 1. Similar to the guided version, we perform regular steps in the first epoch. Afterwards, at each time, we let the model take the step that is randomly sampled from: regular step, re-balancing step on m_0, and re-balancing step on m_1.

Algorithm 1: Random version of the balanced multi-modal learning

Input: T, # of training steps;
T_1, # of steps in the 1th training epoch;
$\Omega_{step} = \{$ regular step,
re-balancing step to accelerate learning from m_0,
re-balancing step to accelerate learning from $m_1 \}$

for $t \leftarrow 1$ **to** T_1 **do**
  Take a regular step;
end
for $t \leftarrow T_1$ **to** T **do**
  Randomly sample a step from Ω_{step};
  Take the sampled step;
end

A.4 DATA PREPARATION

We used three datasets in the paper: Colored MNIST dataset (Kim et al., 2019), ModelNet40 dataset (Su et al., 2015) and NVGesture dataset (Molchanov et al., 2015), as illustrated in Figure 1.

For Colored MNIST dataset and ModelNet40 dataset, we did not perform any extra data pre-processing steps on the original datasets.

For the NVGesture dataset, each video has a resolution of 240×320 and a duration of 80 frames from action starting to ending. There are only three videos with unmatched starting indices between RGB and depth and we adopt the starting frame indice of RGB for all modalities. We randomly select 64 consecutive frames from the videos in the dataset and if the video has less than 64 frames, we zero-pad on both sides of it to obtain 64 frames. Frames are resized as 256×256 and are cropped into 224×224 as inputs (we use static cropping where we crop from the same location across times and modalities).

During training, we perform spatial augmentation on the video, including flipping and random cropping. During inference on validation or test, we perform center cropping on the video.

(a) Colored MNIST (b) ModelNet40 (c) NVGesture

Figure 1: (a) The Colored MNIST dataset (Kim et al., 2019). We consider the colored image, and the gray-scale image as the two input modalities (b) The ModelNet40 dataset (Su et al., 2015). The 2D representations are gray-scale images rendered from 12 different viewpoints of the object. (c)The NVGesture dataset (Molchanov et al., 2015). We use depth and RGB channels as the two modalities.

A.5 MORE RESULTS

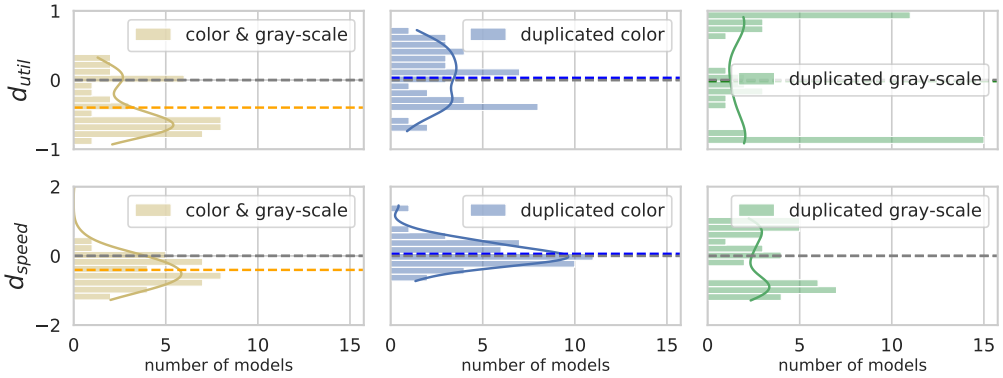


Figure 2: Histograms and estimated density functions of $\mathrm{d_{util}}$ and $\mathrm{d_{speed}}$ of models trained for colored-and-gray-MNIST, using colored and gray-scale images as two modalities, using identical colored images as two modalities and using identical gray-scale images as two modalities.

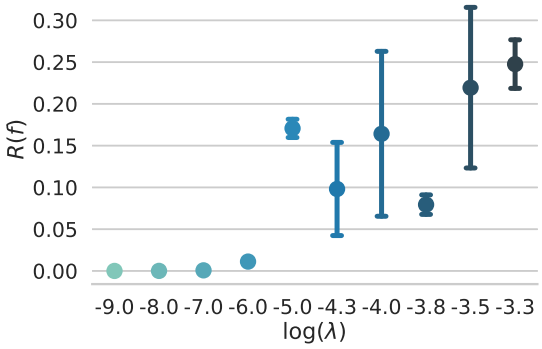
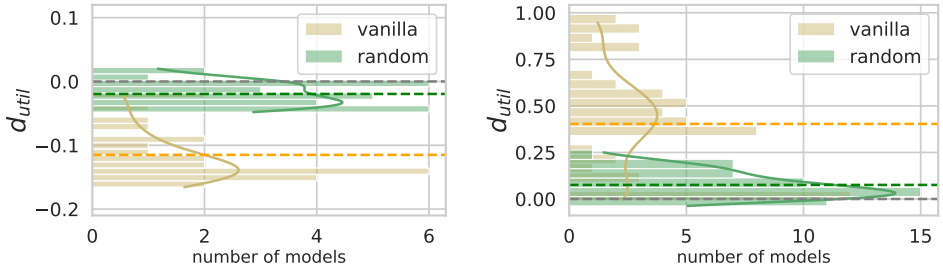


Figure 3: The mean and standard deviation of $R(f)$ for three model repetitions obtained by training the model with λ as the weight on the L1 regularization. Generally, the larger the λ is, the higher the $R(f)$ is, i.e., the sparser the model's parameters are.



(a) ModelNet40 (front and rear views)

(b) NVGesture (RGB and Depth)

Figure 4: Histograms and estimated density functions of $\mathrm{d_{util}}$ for random version of the balanced multi-modal learning process and the vanilla process. Both versions of the balanced multi-modal learning process are less greedy than the vanilla one.