Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks

Nan Wu [1] **Stanisław Jastrzębski** [2 1] **Kyunghyun Cho** [1 3 4 5] **Krzysztof J. Geras** [2 1 3]

Abstract

We hypothesize that due to the greedy nature of learning in multi-modal deep neural networks, these models tend to rely on just one modality while under-fitting the other modalities. Such behavior is counter-intuitive and hurts the models' generalization, as we observe empirically. To estimate the model's dependence on each modality, we compute the gain on the accuracy when the model has access to it in addition to another modality. We refer to this gain as the *conditional utilization rate*. In the experiments, we consistently observe an imbalance in conditional utilization rates between modalities, across multiple tasks and architectures. Since conditional utilization rate cannot be computed efficiently during training, we introduce a proxy for it based on the pace at which the model learns from each modality, which we refer to as the *conditional learning speed*. We propose an algorithm to balance the conditional learning speeds between modalities during training and demonstrate that it indeed addresses the issue of greedy learning.[1] The proposed algorithm improves the model's generalization on three datasets: Colored MNIST, ModelNet40, and NVIDIA Dynamic Hand Gesture.

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 regime as a *multi-modal learning process* and DNNs resulting from it as *multi-modal DNNs*.

Several recent studies have reported unsatisfactory performance of multi-modal DNNs 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., 2021b; Hessel & Lee, 2020). For example, in Visual Question Answering (VQA), multi-modal DNNs were found to ignore the visual modality and exploit statistical regularities shared between the text in the question and the text in the answer alone, resulting in poor generalization (Cadene et al., 2019; Gat et al., 2020; Agrawal et al., 2016). Similarly, in Human Action Recognition, multi-modal DNNs trained on images and audio were observed to perform worse than uni-modal DNNs trained on images only (Wang et al., 2020a). 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).

These earlier negative findings compel us to ask, *what prevents multi-modal DNNs from achieving better performance?* In order to answer this question, we put forward the *greedy learner hypothesis*. The greedy learner hypothesis states that 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 prevents a multi-modal DNN from learning to exploit all available modalities and often results in worse generalization. It explains the challenge in training multi-modal DNNs and motivates us to design a better multi-modal learning algorithm.

We first diagnose a multi-modal DNN's ability to utilize all modalities by analyzing its *conditional utilization rates*. On a task with two modalities, m_0 and m_1, we define a model's conditional utilization rates as the relative difference in accuracy between two models derived from the DNN, one using both modalities and the other using only one modality. The conditional utilization rate of m_1 given m_0, denoted by $\boldsymbol{u}(m_1|m_0)$, measures how important it is for the model to use m_1 in order to reach accurate predictions, given the presence of m_0. In several multi-modal learning tasks, we consistently observe a significant imbalance in conditional utilization rate between modalities. For example, we have $\boldsymbol{u}(\text{depth}|\text{RGB}) = 0.63$ and $\boldsymbol{u}(\text{RGB}|\text{depth}) = 0.01$ for

[1]Center for Data Science, New York University [2]NYU Grossman School of Medicine [3]Courant Institute of Mathematical Sciences, New York University [4]Genentech [5]CIFAR LMB. Work was done when Stanisław Jastrzębski was at NYU. Correspondence to: Nan Wu <nan.wu@nyu.edu>.

[1]We provide an implementation of the proposed methods at https://github.com/nyukat/greedy_multimodal_learning

a model trained for Hand Gesture Recognition using the RGB and the depth modalities. Since $0 \leq |\boldsymbol{u}| \leq 1$, these two observed values indicate that the model solely relies on the depth modality, ignoring the RGB modality almost completely. These observations support the conjecture that the multi-modal learning process often results in models that under-utilize some of the input modalities.

According our proposed greedy learner hypothesis, it is the different speeds at which a multi-modal DNN learns from different modalities that leads to an imbalance in conditional utilization rate. If we intervene in the training process to adjust these speeds, we may be able to prevent the hurtful imbalance across input modalities. We analyze the learning dynamics of model components and propose a metric called *conditional learning speed*, defined using the gradient norm and the norm of models' parameters. It measures the speed at which the model learns from one modality relative to the other modalities. We empirically show that it is a reasonable proxy for conditional utilization rate. We introduce a training algorithm called *balanced multi-modal learning* which uses conditional learning speed to guide the model to learn from previously underutilized modalities. We show that models trained with this algorithm 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. Related work

Dataset and model bias inspection in VQA Multi-modal DNNs are expected to leverage cross-modal interactions in VQA, but have been reported to exploit the modality-wise bias in the data (Jabri et al., 2016; Goyal et al., 2017; Winterbottom et al., 2020). Several frameworks were proposed to evaluate the severity of this issue, such as constructing de-biased datasets (Hessel & Lee, 2020; Agrawal et al., 2018), evaluating models' performance when eliminating cross-modal interactions via empirical function projection (Hessel & Lee, 2020) or via permuting the features of a modality across samples (Gat et al., 2021). There are also efforts to overcome this problem. Many of them rely on a question-only branch for capturing spurious relationships between questions and answer candidates and help through cross-entropy re-weighing strategies (Cadene et al., 2019; Han et al., 2021a; Lao et al., 2021). Instead of estimating the bias through a question-only branch, Gat et al. (2020) proposes to supply inputs with Gaussian perturbations to the model and regularize it by maximizing functional entropies, in order to force the model to use multiple sources of information. In this work, we expand the discussion from VQA to multi-modal classification tasks. These tasks are different from VQA but the problem identified above also appears there. We explain this phenomenon by the greedy nature of learning in multi-modal DNNs and design tools to overcome inadequate modality utilization in multi-modal classification.

Video classification How to benefit from multi-modality in video data has been a popular research direction for video understanding. Prior work focuses primarily on improving architectural designs of the DNNs (Ngiam et al., 2011; Neverova et al., 2015; Tran et al., 2015; Lazaridou et al., 2015; Pérez-Rúa et al., 2019; Joze et al., 2020; Arnab et al., 2021; Sun et al., 2021). Our study is related to a recent study by Wang et al. (2020a). They observe that the best uni-modal DNNs often outperforms the multi-modal DNNs, and propose a framework that estimates the uni-modal branches' generalization and overfitting speeds in order to calibrate the learning through loss re-weighing. Their proposed methods rely on estimating models' performance occasionally on a held-out validation set during training, which makes it costly and unstable for small datasets. To serve a similar purpose but not from the perspective of model optimization, Wang et al. (2020b) design a parameter-free multi-modal fusion framework that dynamically exchanges channels between the uni-modal branches. In this work, we provide tools to directly examine the imbalanced modality utilization in jointly-trained multi-modal DNNs, and implement calibration through a computationally efficient method. We focus on verifying our solution with intermediate fusion as it yields better performance than late fusion (Joze et al., 2020), though an investigation such as ours was missing in the literature.

3. Problem setup

Without loss of generality, we consider two input modalities, referred to as m_0 and m_1. We denote a multi-modal dataset by $\mathcal{D}$. This dataset consists of multiple instances $(\boldsymbol{x}, y)$, 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 to use this data set to train a model that accurately predicts y from $\boldsymbol{x}$.

We use a multi-modal DNN with two uni-modal branches, denoted by ϕ_0 and ϕ_1, taking $\boldsymbol{x}_{m_0}$ and $\boldsymbol{x}_{m_1}$ as input. The two uni-modal branches are interconnected by layer-wise fusion modules. According to the categorization of fusion strategies in the deep learning literature, this is considered "intermediate fusion" (Ngiam et al., 2011; Atrey et al., 2010; Baltrušaitis et al., 2018). It has demonstrated competitive performance in comparison to multi-modal DNNs with late fusion in multiple tasks (Perez et al., 2018; Joze et al., 2020; Anderson et al., 2018; Wang et al., 2020b).

Specifically, we implement each fusion module with a multi-modal transfer module (MMTM) (Joze et al., 2020). It

connects corresponding
uni-modal branches. Le
$\mathbb{R}^{M_1 \times \cdots \times M_J \times C'}$ denote
layers. Here, N_i and M
of each feature map, and
feature maps. We apply
dimensions of the feature
by $h_0 \in \mathbb{R}^C$ and $h_1 \in$
MMTM is then applied t

$$w_0, w$$

where $[\cdot, \cdot]$ represents the
and $w_1 \in \mathbb{R}^{C'}$. The fu
fully-connected layer, Re
and two more fully-cor
activation in the dimensi

Next, we re-scale the fe
obtained activations (w_0

$$\tilde{\boldsymbol{A}}_0 = 2 \times \sigma(w_0) \odot \boldsymbol{A}_0, \quad \boldsymbol{A}_1 = 2 \times \sigma(w_1) \odot \boldsymbol{A}_1 \qquad (2)$$

where $\odot$ is the channel-wise product operation and $\sigma(\cdot)$ is the sigmoid function.

We feed $\tilde{\boldsymbol{A}}_0$ and $\tilde{\boldsymbol{A}}_1$ to the next layer of ϕ_0 and ϕ_1. Thus information from one modality is shared from one uni-modal branch to the other.

We train a multi-modal DNN f on $\mathcal{D}^{\text{train}}$. Let $\hat{y}$ be the prediction of f for an input $\boldsymbol{x} = (\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1})$:

$$\hat{y} = f(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}). \qquad (3)$$

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 f are updated by stochastic gradient descent (SGD) to minimize the loss:

$$L = \text{CE}(y, \hat{y}_0) + \text{CE}(y, \hat{y}_1) \qquad (4)$$

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 it reaches 100% classification accuracy on $\mathcal{D}^{\text{train}}$, and pick the model checkpoint associated with the highest validation accuracy across all epochs.

4. The greedy learner hypothesis

In this section, we first derive conditional utilization rate to help explain the multi-modal DNN and to quantify its imbalance in modality utilization. Then we introduce the greedy learner hypothesis to explain challenges observed in training multi-modal DNNs.

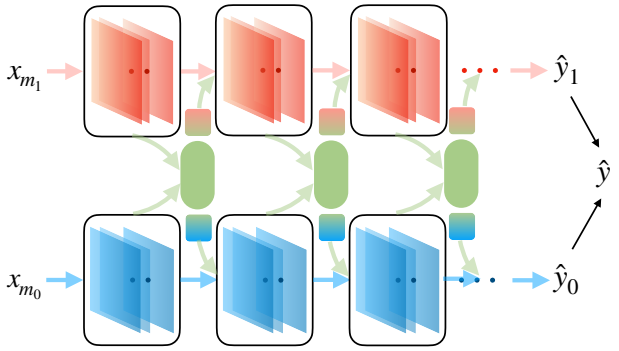


Figure 1. The multi-modal DNN with intermediate fusion that we study in this work. Blue and red blocks represent the two uni-modal networks, ϕ_0 and ϕ_1, learning from modalities m_0 and m_1. The fusion layers are marked with green, green-blue, and green-red. We denote the predictions from ϕ_0 and ϕ_1 by $\hat{y}_0$ and $\hat{y}_1$, and the average of the two by $\hat{y}$ which is the model's prediction.

4.1. Conditional utilization rate

In the multi-modal DNN shown in Figure 1, each uni-modal branch largely focuses on the associated input modality, and the fusion modules generate context information using all modalities, and feed the information to the uni-modal branches. Both $\hat{y}_0$ and $\hat{y}_1$ depend on information from both modalities. We derive f_0 and f_1 from f:

$$\hat{y}_0 = f_0(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}), \quad \hat{y}_1 = f_1(\boldsymbol{x}_{m_0}, \boldsymbol{x}_{m_1}). \qquad (5)$$

Let $\mathcal{D}^{\text{train}} = \{\boldsymbol{x}^i\}_{i=1}^n$ and $h_0(\boldsymbol{x}^i)$ and $h_1(\boldsymbol{x}^i)$ denote the value of h_0 and h_1 when the model takes $\boldsymbol{x}^i$ as input. We compute the average of h_0 and h_1 over $\mathcal{D}^{\text{train}}$:

$$\overline{h}_0 = \frac{1}{n}\sum_{i=1}^{n} h_0(\boldsymbol{x}^i), \quad \overline{h}_1 = \frac{1}{n}\sum_{i=1}^{n} h_1(\boldsymbol{x}^i). \qquad (6)$$

To estimate f's reliance on each modality independently, we modify the operations introduced in §3 for each fusion module as below:

$$w_0, * = g([h_0, \overline{h}_1]), \quad *, w_1 = g([\overline{h}_0, h_1]). \qquad (7)$$

With such modified operation in every fusion module in the multi-modal DNN, we cut off the road to share information between the uni-modal branches and let the output of each branch rely on a single modality. We derive f'_0 and f'_1 and denote their outputs by $\hat{y}'_0$ and $\hat{y}'_1$:

$$\hat{y}'_0 = f'_0(\boldsymbol{x}_{m_0}), \quad \hat{y}'_1 = f'_1(\boldsymbol{x}_{m_1}). \qquad (8)$$

In summary, we derive four models from f and compute the accuracy of each on $\mathcal{D}^{\text{test}}$, denoted by $A(\cdot)$. We group the four accuracies into two pairs: $(A(f_0), A(f'_0))$, and $(A(f_1), A(f'_1))$.

We define the conditional utilization rates as bellow.

Definition 4.1 (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_0 and m_1 are defined as

$$\boldsymbol{u}(m_0|m_1) = \frac{A(f_1) - A(f_1')}{A(f_1)}, \tag{9}$$

and

$$\boldsymbol{u}(m_1|m_0) = \frac{A(f_0) - A(f_0')}{A(f_0)}. \tag{10}$$

The conditional utilization rate is the relative change in accuracy between the two models within each pair. For example, $u(m_0|m_1)$ measures the marginal contribution that m_0 has in increasing the accuracy of $\hat{y}_1$.

Let $\mathrm{d}_{\mathrm{util}}(f)$ denote the difference between conditional utilization rates: $\mathrm{d}_{\mathrm{util}}(f) = \boldsymbol{u}(m_1|m_0) - \boldsymbol{u}(m_0|m_1)$. It is bounded between -1 and 1. When $\mathrm{d}_{\mathrm{util}}(f)$ 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. Thus, if we observe high $|\,\mathrm{d}_{\mathrm{util}}(f)|$, we say that f exhibits imbalance in utilization between modalities.

4.2. Multi-modal learning process is greedy

We propose our hypothesis that builds on the following assumptions regarding multi-modal data, as well as observations previously made in the literature.

First, for most multi-modal learning tasks, both modalities are assumed to be predictive of the target. This can be expressed as $I(\boldsymbol{Y}, \boldsymbol{X}_{m_0}) > 0$ and $I(\boldsymbol{Y}, \boldsymbol{X}_{m_1}) > 0$, where I denotes mutual information (Blum & Mitchell, 1998; Sridharan & Kakade, 2008). In order to minimize one of the modality-specific losses, e.g. $\mathrm{CE}(y, \hat{y}_0)$, one can either update the parameters of the uni-modal branch taking m_0 as input, or the parameters of the fusion layers that pass information from m_1 to $\hat{y}_0$, or both.

Second, as shown in many tasks, different modalities have been said to be predictive of the target at different degrees. For example, it has been observed that when training DNNs on each modality separately, they usually do not reach the same level of performance (Wang et al., 2020b; Joze et al., 2020). In addition, multi-modal DNNs exhibit varying accuracy when trained on different subsets of modalities present for the task (Weng et al., 2021; Pérez-Rúa et al., 2019; Liu et al., 2018). Third, DNNs learn from different modalities at different speeds. This has been observed in both uni-modal DNNs and multi-modal DNNs (Wang et al., 2020a; Wu et al., 2020).

Now we formulate the *greedy learner hypothesis* to explain the behavior of multi-modal DNNs:

> A multi-modal learning process is *greedy* when it produces models that rely on only one of the available modalities. The modality that the multi-modal DNN primarily relies on is the modality that is the fastest to learn from. 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.

To characterize greediness of the multi-modal learning process, we need to repeat it several times. Every time, we sample a learning rate from a given range and initialize the network' parameters randomly. Let $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ denote the expectation of $\widehat{\mathrm{d}}_{\mathrm{util}}$ over the empirical distribution of the models obtained. The absolute value of $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ is associated with the greediness of the process. The higher the $|\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]|$, the greedier the learning process.

To validate our hypothesis empirically, we conduct experiments on several multi-modal datasets using different network architectures (cf. §6.1). We show that the multi-modal learning process cannot avoid producing models that exhibit a high degree of imbalance in utilization between modalities. Their performance can be improved by using a less greedy algorithm which we will introduce in the next section.

5. Making multi-modal learning less greedy

We aim to make multi-modal learning less greedy by controlling the speed at which a multi-modal DNN learns to rely on each modality. In order to achieve this, we first define conditional learning speed to measure the speed at which the DNN learns from one modality. It serves as an efficient proxy for the conditional utilization rate of the corresponding modality, as shown empirically in §6.2 and §6.3. We then propose the balanced multi-modal learning algorithm, which controls the difference in conditional learning speed between modalities that the model exhibits during training.

5.1. Conditional learning speed

As demonstrated in §4.2, 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. However, conditional utilization rates are designed to be measured after training is done and they are derived from the model's accuracy achieved on the held-out test set. This makes conditional utilization rate not the best tool to use to monitor training at the real-time. We instead derive a proxy metric, called *conditional learning speed*, that captures relative learning speed between modalities during training.

Let us revisit the multi-modal DNN presented in Figure 1.

We denote the parameters of uni-modal branches ϕ_0 and ϕ_1 by $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$. Besides $\boldsymbol{\theta}_0$, we consider the parameters of the layers marked with green and green/blue, as important components of the function mapping $\boldsymbol{x}$ to $\hat{y}_0$ (i.e., f_0). These parameters are from the fusion modules and we refer to them as $\boldsymbol{\theta}'_0$. Analogously, parameters in the layers marked with green and green/red are considered as important contributors to the function mapping $\boldsymbol{x}$ to $\hat{y}_1$ (i.e., f_1). We denote them by $\boldsymbol{\theta}'_1$. In this way, we divide the model's parameters into two pairs: $(\boldsymbol{\theta}_0, \boldsymbol{\theta}'_0)$ and $(\boldsymbol{\theta}_1, \boldsymbol{\theta}'_1)$.[2]

We define the model's conditional learning speed as follows.

Definition 5.1 (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

$$\boldsymbol{s}(m_1|m_0;t) = \log \frac{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}'_0;i)}{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}_0;i)}, \tag{11}$$

and

$$\boldsymbol{s}(m_0|m_1;t) = \log \frac{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}'_1;i)}{\sum_{i=1}^{t} \mu(\boldsymbol{\theta}_1;i)}, \tag{12}$$

where for any parameters $\boldsymbol{\theta}$, let $\boldsymbol{\theta}_{(i-1)}$ and $\boldsymbol{\theta}_{(i)}$ denote its value before and after the gradient descent step i; we have $\boldsymbol{G} = \frac{\partial L}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}_{(i-1)}}$; and $\mu(\boldsymbol{\theta};i) = ||\boldsymbol{G}||_2^2/||\boldsymbol{\theta}_{(i)}||_2^2$ quantifies the change of f that comes from updating $\boldsymbol{\theta}$ at the i^{th} step, which is also interpreted as the effective update on $\boldsymbol{\theta}$.

This definition of $\mu(\boldsymbol{\theta};i)$ is inspired by the discussion on the effective update of parameters (Van Laarhoven, 2017; Zhang et al., 2019; Brock et al., 2021; Hoffer et al., 2018). When normalization techniques, such as batch normalization (Ioffe & Szegedy, 2015), are applied to the DNNs, the key property of the weight vector, $\boldsymbol{\theta}$, is its direction, i.e., $\boldsymbol{\theta}/||\boldsymbol{\theta}||_2^2$. Thus, we measure the update on $\boldsymbol{\theta}$ using the norm of its gradient normalized by its norm.

The conditional learning speed, $\boldsymbol{s}(m_1|m_0;t)$, 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 $\hat{y}_0$ and information from m_0 to $\hat{y}_0$ is carried by $\boldsymbol{\theta}_0$, $\boldsymbol{s}(m_1|m_0;t)$ reflects how fast the model learns from m_1 relative to m_0 after the first t steps.

We 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 $\mathrm{d}_{\mathrm{util}}(f)$. For each model, we report $\mathrm{d}_{\mathrm{speed}}(f;T)$ as $\mathrm{d}_{\mathrm{speed}}(f)$ where the model takes T steps until reaching the highest accuracy on $\mathcal{D}^{\mathrm{val}}$.

We anticipate that the conditional learning speed serves as a proxy for the conditional utilization rate. In §6.2 and §6.3, we show the distributions of $\widehat{\mathrm{d}}_{\mathrm{speed}}$ and $\widehat{\mathrm{d}}_{\mathrm{util}}$ side-by-side to validate this.

[2]Note that parameters in the green block are both in $\boldsymbol{\theta}'_0$ and $\boldsymbol{\theta}'_1$.

5.2. Balanced multi-modal learning

Since we assume $\mathrm{d}_{\mathrm{speed}}(f,t)$ 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 alternately to balance their conditional learning speeds. See Algorithm 1 for an overall description.

Algorithm 1 Balanced multi-modal learning

Input: re-balancing window size Q, imbalance tolerance parameter α, # epochs N, # updating steps per epoch n
Initiate: $M_{\boldsymbol{\theta}_0}, M_{\boldsymbol{\theta}'_0}, M_{\boldsymbol{\theta}_1}, M_{\boldsymbol{\theta}'_1} \leftarrow 0,0,0,0$; $q \leftarrow Q$
for $i = 1$ **to** n **do**
  Take a regular step;
  $M_{\boldsymbol{\theta}_0} \leftarrow M_{\boldsymbol{\theta}_0} + \mu(\boldsymbol{\theta}_0;i)$; $M_{\boldsymbol{\theta}'_0} \leftarrow M_{\boldsymbol{\theta}'_0} + \mu(\boldsymbol{\theta}'_0;i)$;
  $M_{\boldsymbol{\theta}_1} \leftarrow M_{\boldsymbol{\theta}_1} + \mu(\boldsymbol{\theta}_1;i)$; $M_{\boldsymbol{\theta}'_1} \leftarrow M_{\boldsymbol{\theta}_1} + \mu(\boldsymbol{\theta}'_1;i)$;
end for
for $i = 1$ **to** $n \times N$ **do**
  if $q == Q$ **then**
    Take a regular step;
    $M_{\boldsymbol{\theta}_0} \leftarrow M_{\boldsymbol{\theta}_0} + \mu(\boldsymbol{\theta}_0;i)$; $M_{\boldsymbol{\theta}'_0} \leftarrow M_{\boldsymbol{\theta}'_0} + \mu(\boldsymbol{\theta}'_0;i)$;
    $M_{\boldsymbol{\theta}_1} \leftarrow M_{\boldsymbol{\theta}_1} + \mu(\boldsymbol{\theta}_1;i)$; $M_{\boldsymbol{\theta}'_1} \leftarrow M_{\boldsymbol{\theta}_1} + \mu(\boldsymbol{\theta}'_1;i)$;
    $\mathrm{d}_{\mathrm{speed}} = \log(M_{\boldsymbol{\theta}'_0}/M_{\boldsymbol{\theta}_0}) - \log(M_{\boldsymbol{\theta}'_1}/M_{\boldsymbol{\theta}_1})$;
    if $|\,\mathrm{d}_{\mathrm{speed}}| > \alpha$ **then** $q \leftarrow 1$;
  else
    $q \leftarrow q + 1$
    Take a re-balancing step to accelerate learning from m_0 **if** $\mathrm{d}_{\mathrm{speed}} > 0$ **else** from m_1;
  end if
end for

We refer to the training steps at which we perform forward and backward passes normally as the *regular steps*. We introduce the *re-balancing steps* at which we update one of the uni-modal branches intentionally in order to accelerate the model to learn from its input modality.

In a re-balancing step where we aim to accelerate the model's learning from m_0, in every fusion module in the network, we re-scale the feature maps $\boldsymbol{A}_1$ in the same way as in §3 but we re-scale $\boldsymbol{A}_0$ differently:

$$\tilde{\boldsymbol{A}}_0 = 2 \times \sigma(\overline{w}_0) \odot \boldsymbol{A}_0, \quad \tilde{\boldsymbol{A}}_1 = 2 \times \sigma(w_1) \odot \boldsymbol{A}_1, \tag{13}$$

where $\{\boldsymbol{x}^i\}_{i=1}^{n_t}$ are the n_t samples used in the previous regular training steps; $w_0(\boldsymbol{x}^i)$ denotes the activation when the model takes $\boldsymbol{x}^i$ as input; and $\overline{w}_0 = 1/n_t \sum_{i=1}^{n_t} w_0(\boldsymbol{x}^i)$.

In a re-balancing step to accelerate learning from m_1, we apply the modification on $\boldsymbol{A}_1$ instead of $\boldsymbol{A}_0$:

$$\tilde{\boldsymbol{A}}_0 = 2 \times \sigma(w_0) \odot \boldsymbol{A}_0, \quad \tilde{\boldsymbol{A}}_1 = 2 \times \sigma(\overline{w}_1) \odot \boldsymbol{A}_1, \tag{14}$$

where $\overline{w}_1 = 1/n_t \sum_{i=1}^{n_t} w_1(\boldsymbol{x}^i)$.

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

6. Experiments and results

6.1. Datasets, tasks and baselines

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. In the validation set of 10,000 examples, each example also has a gray-scale image and a corresponding monochromatic image, although 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 to connect them. The corresponding uni-modal DNNs trained on the monochromatic images and the gray-scale images achieve respective validation accuracies of 41% and 99%. We use this synthetic dataset mainly to demonstrate the proposed greedy learner hypothesis.

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). The task is to classify a 3D object based on the 2D views of its front and back (Su et al., 2015). Each example is a collection of 2D images (224×224 pixels) of a 3D object. 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 89% and 88% accuracy when learning from the front view (m_0) and the rear view (m_1), respectively.

NVGesture (or NVIDIA Dynamic Hand Gesture Dataset (Molchanov et al., 2015)) consists of 1,532 video clips (1,050 training and 482 test ones) of hand gestures divided into 25 classes. We sample 20% of training examples as the validation set and use depth and RGB as the two modalities. We adopt the data preparation steps used in Joze et al. (2020) and use the I3D architecture (Carreira & Zisserman, 2017) as uni-modal branches and MMTMs as fusion modules in the six final inception modules. During training, we perform spatial augmentation on the video, including flipping and random cropping. During inference on the validation or test set, we perform center cropping on the video.

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

6.2. Validating the greedy learner hypothesis

In this section, we run the conventional multi-modal learning process on tasks with different input and output pairs to illustrate our hypothesis experimentally.

For each task introduced in §6.1, in addition to the original dataset, we construct a dataset with two identical input modalities by duplicating one of the modalities. For example, when using the colored-and-gray-MNIST dataset, we predict the digit class using two identical gray-scale images. We train multi-modal DNNs on these datasets as explained below for each task:

- Colored-and-gray-MNIST: we train multi-modal DNNs using SGD with a momentum coefficient of 0.9 and a batch size of 128. We sample 20 learning rates at random from the interval $[10^{-5}, 1]$ on a logarithmic scale. We train the model four times using each of the learning rates and random initialization of the parameters. In total, we train 80 models.

- ModelNet40: we use SGD without momentum and use a batch size of eight. We select nine learning rates from 10^{-3} to 1 and train three copies of the model for each learning rate. This ends up with 27 models.

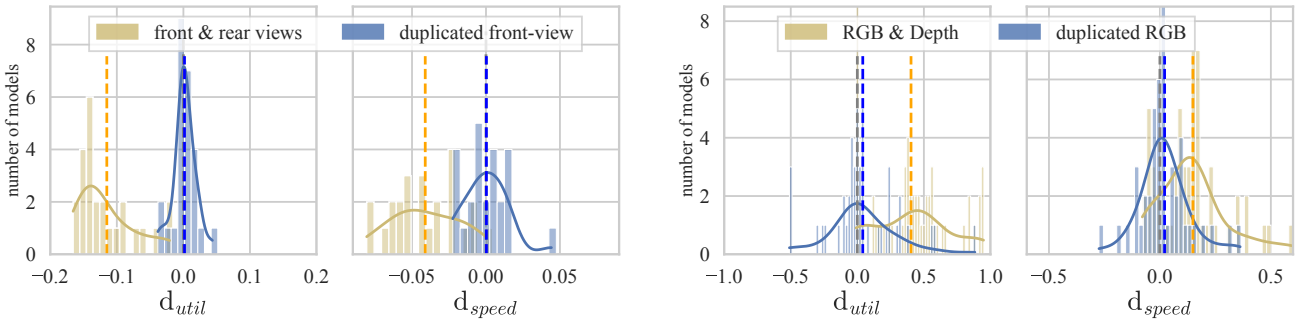
- NVGesture: we use a batch size of four, SGD with a momentum of 0.9, and uniformly sample 20 learning rates from the interval $[10^{-4}, 10^{-1.5}]$ on a logarithmic scale. We train the model three times using each learning rate, resulting in 60 models in total.

We present the results of the experiments on ModelNet40 and NVGesture in Figure 2 and on Colored-and-gray-MNIST in Figure 7 in Appendix B. We discuss three most interesting phenomena that we observe in the results.

First, many models have high $|\,\mathrm{d}_{\mathrm{util}}|$. This confirms that the 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 around 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 $|\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]|$ of approximately 0.3, 0.1 and 0.4 for colored-and-gray-MNIST, ModelNet40 and NVGesture, respectively. We formalize the above observations as the following conjecture.

There exists an $\epsilon > 0$, s.t. $|\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]| > \epsilon$, when modalities



(a) ModelNet40 (b) NVGesture

Figure 2. Histograms and estimated density functions of $\widehat{\mathrm{d}}_{\mathrm{util}}$ and $\widehat{\mathrm{d}}_{\mathrm{speed}}$ of models trained using different modalities or duplicated single modality in (a) ModelNet40 or (b) NVGesture. We mark zero, $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ and $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{speed}}]$ with dashed lines. Many models have high $|\,\mathrm{d}_{\mathrm{util}}|$. 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. We see $\widehat{\mathrm{d}}_{\mathrm{speed}}$ exhibits similar distributions as $\widehat{\mathrm{d}}_{\mathrm{util}}$.

are different. Otherwise, $\widehat{\mathrm{d}}_{\mathrm{util}}$ is distributed symmetrically around zero and $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}] = 0$.

Third, by observing conditional learning speed $\widehat{\mathrm{d}}_{\mathrm{speed}}$, we can draw the same conclusions. In fact, the distributions of $\widehat{\mathrm{d}}_{\mathrm{speed}}$ largely replicate the distributions of $\widehat{\mathrm{d}}_{\mathrm{util}}$. It validates our greedy learner hypothesis which attributes the imbalance in reliance on different modalities to the varying rate at which the learner learns from them. It moreover confirms $\mathrm{d}_{\mathrm{speed}}$ is an appropriate proxy to use to re-balance multi-modal learning.

6.3. Strong regularization encourages greediness

We also conjecture that regularization is a factor impacting the greediness of learning in multi-modal DNNs and strong regularization encourages greediness. We provide the following study to validate this.

We investigate L1 regularization's impact on multi-modal DNNs. Precisely, we train the networks to optimize the loss $L' = L + \lambda||\boldsymbol{\theta}||_1$, where L is the classification loss in §3, $\boldsymbol{\theta}$ 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 quantity by $R(f)$. Since L1 regularization encourages sparsity of the network's parameters, as shown in Figure 3, the larger the λ, the higher the $R(f)$.

We conduct this study with ModelNet40, using the front and the rear views. We compute $|\,\mathrm{d}_{\mathrm{util}}(f)|$ for ten values of λ from an interval $[10^{-9}, 10^{-3}]$. We use SGD without momentum, we set the learning rate to 0.1 and batch size to eight. Using each combination of hyperparameters, we repeat training for three times with random initialization and get three models.

As shown in Figure 4, $|\,\mathrm{d}_{\mathrm{util}}(f)|$ is positively correlated with $R(f)$, and when $\lambda \geq 10^{-5}$, $|\,\mathrm{d}_{\mathrm{util}}(f)|$ increases significantly with λ increasing, as shown in Figure 8 in Appendix B. In other words, the stronger the regularization is, the larger the imbalance in utilization between modal-

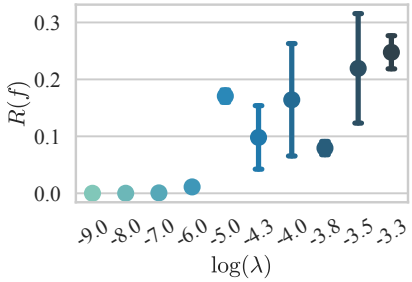
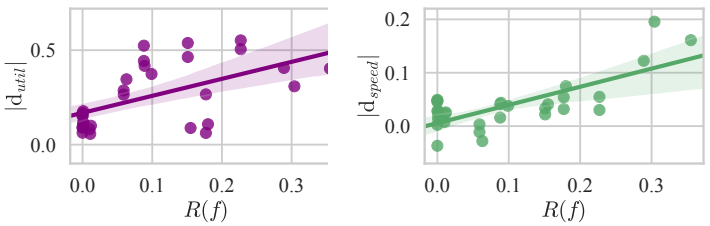


Figure 3. The mean and standard deviation of $R(f)$ for models trained with different weights (λ) on the L1 regularizer.. The larger the λ is, the higher the $R(f)$ is, i.e., the sparser the model's parameters are.



(a) $|\,\mathrm{d}_{\mathrm{util}}|$ vs. $R(f)$ (b) $|\,\mathrm{d}_{\mathrm{speed}}|$ vs. $R(f)$

Figure 4. The observed $|\,\mathrm{d}_{\mathrm{util}}|$ and $|\,\mathrm{d}_{\mathrm{speed}}|$ for models trained with different weights (λ) on the L1 regularizer. In (a) and (b), we see both $|\,\mathrm{d}_{\mathrm{util}}|$ and $|\,\mathrm{d}_{\mathrm{speed}}|$ increases along $R(f)$. It indicates that the multi-modal learning process becomes greedier if we introduce stronger regularization.

Table 1. Test accuracy of best uni-modal DNNs and multi-modal DNNs trained using different strategies.

	Colored-and-gray-MNIST	ModelNet40	NVGesture-scratch	NVGesture-pretrained
uni-modal (best)	99.14±0.11[1]	89.34±0.39	77.59±0.55	78.98±2.02
multi-modal (vanilla)	45.26±0.46	90.09±0.58	79.81±1.14	83.20±0.21
+ RUBi (Cadene et al., 2019)	44.79±0.62	90.45±0.58	79.95±0.12	81.60±1.28
+ random (proposed)	74.07±2.75	91.36±0.10	79.88±0.90	82.64±0.84
+ guided (proposed)	**91.01±1.20**	**91.37±0.28**	**80.22±0.73**	**83.82±1.45**

[1] The monochromatic image cannot help with the prediction and it is hard to avoid it hurting the ensemble performance. The uni-modal branch taking gray-scale image in the multi-modal DNNs (guided) achieves an accuracy of 99.16±0.14.

ities we observe. We also find that $|\,\mathrm{d}_{\mathrm{speed}}|$ 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.

6.4. Balanced multi-modal learning

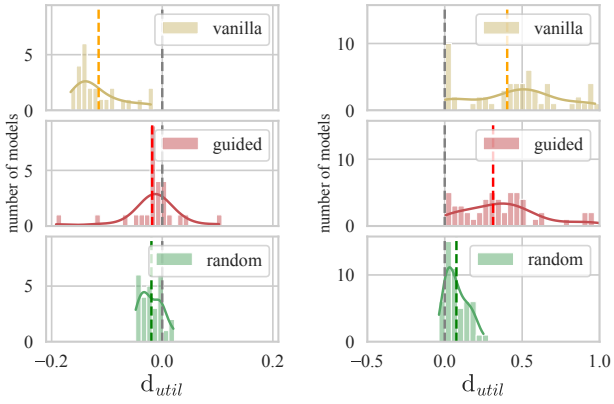
Calibrated modality utilization We train multi-modal DNNs as described in §6.2, using the guided, the random, and the conventional training algorithm (referred to as *vanilla*). For ModelNet40, we set the imbalance tolerance parameter α to 0.01 and the re-balancing window size Q to 5. For NVGesture, we use α of 0.1 and Q of 5.[3] Results are shown in Figure 5. 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 0.3 and 0.4 for models trained with the guided and the vanilla algorithm. On ModelNet40, we obtain $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ of approximately -0.0, and -0.1 for models trained with the guided and the vanilla algorithm.

The random version of the proposed method also calibrates modality utilization effectively, giving $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ of approximately 0.1 and -0.0 for models trained on NVGesture and ModelNet40. We will see in the next section that it helps less on generalization compared to the guided version.

Improved generalization performance We compare the generalization ability of multi-modal DNNs trained by the baseline (vanila), the proposed methods (guided and random) and the RUBi learning strategy (Cadene et al., 2019). We chose RUBi as one of the baselines since it is model-agnostic and fusion-method-agnostic, and it is easy to adapt to different tasks. RUBi is designed to serve a similar purpose, that is, to encourage the model to learn from all modalities and has been shown to be helpful for VQA.

For each algorithm, we train each model three times with the same learning rate. We use 0.01, 0.1 and 0.01 as learning rate for Colored-and-gray-MNIST, ModelNet40 and NVGesture respectively. We use α of 0.1 for Colored-and-gray-MNIST and NVGesture and 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 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".

Besides the multi-modal DNNs trained using different strategies, we implement Squeeze-and-Excitation blocks (**?**) to obtain uni-modal DNNs that are comparable to the multi-modal DNNs with MMTMs (Joze et al., 2020). We train uni-modal DNNs with each modality and present the best performance they achieve for each task. We report means and standard deviations of the models' test accuracies in Table 1. The guided algorithm improves the models' generalization performance over all other three methods in all four cases.[4] RUBi does not show consistent improvement



(a) ModelNet40 (b) NVGesture

Figure 5. Histograms and estimated density functions of $\widehat{\mathrm{d}}_{\mathrm{util}}$ of models trained using the vanilla, the guided and the random algorithm on (a) ModelNet40 and (b) NVGesture. We use dashed lines to mark zero and $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$ over the set of models trained with each of the algorithms. Both versions of the proposed algorithm is less greedy than the vanilla one.

[3] We provide studies on the model's sensitivity to Q and α in Figure 9a and Figure 9b in Appendix B.

[4] Results on NVGesture are not directly comparable to numbers in other works since we use 20% training samples as the validation data.

across tasks compared to the vanilla algorithm. We did not find this strategy as helpful as reported for VQA.

Our idea can be extended naturally to tasks with more than two modalities. The conditional learning speed of the i^{th} modality can be derived using the weight norm and gradient norm of the i^{th} uni-modal network's parameters and the fusion module's parameters impacting $\hat{y}_i$. The conditional utilization rate would be computed analogously to the bi-modal case, by passing the averaged feature maps of all other modalities to the fusion modules. We define $\mathrm{d}_{\mathrm{speed}}$ by the two most different conditional learning speeds and accelerate the learning of one modality per time. We experiment with three views in ModelNet40 and find that the guided algorithm outperforms the vanilla one as it improves the accuracy from 91.21% to 92.45%.

In summary, we present that the greediness of the current training algorithm is an issue preventing multi-modal DNNs from achieving better performance. The proposed method can help to overcome this issue on multiple datasets.

7. Discussion

Our work demonstrates 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. For many multi-modal classification tasks, such as video classification, the harm of this behavior may not be as obvious as in VQA which strongly relies on cross-modal reasoning. However, the tools we propose in this work enable us to have a concrete look of what the model has learned, which has been a missing component in multi-modal learning systems. We validated our statements experimentally on three multi-modal datasets and 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. In addition to the previous discussion on this problem, our greedy learner hypothesis provides a complementary explanation and the methods inspired by it enrich the spectrum of available tools for multi-modal learning.

Acknowledgements

This work was supported in part by grants from the National Institutes of Health (P41EB017183 and R21CA225175), the National Science Foundation (HDR-1922658), the Gordon and Betty Moore Foundation (9683), and Samsung Advanced Institute of Technology (under the project *Next Generation Deep Learning: From Pattern Recognition to AI*). Nan Wu is supported by Google PhD Fellowship. We thank Catriona Geras and Taro Makino for their comments on writing. We appreciate the suggestions from ICML reviewers.

References

Agrawal, A., Batra, D., and Parikh, D. Analyzing the behavior of visual question answering models. *EMNLP*, 2016.

Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. *CVPR*, 2018.

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. Vivit: A video vision transformer. *arXiv:2103.15691*, 2021.

Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 2010.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Blum, A. and Mitchell, T. Combining labeled and unlabeled data with co-training. *COLT*, 1998.

Brock, A., De, S., Smith, S. L., and Simonyan, K. High-performance large-scale image recognition without normalization. *ICML*, 2021.

Cadene, R., Dancette, C., Ben-Younes, H., Cord, M., and Parikh, D. Rubi: Reducing unimodal biases in visual question answering. *NeurIPS*, 2019.

Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.-C., and Liu, J. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *ECCV*, 2020.

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

Gat, I., Schwartz, I., Schwing, A., and Hazan, T. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *NeurIPS*, 2020.

Gat, I., Schwartz, I., and Schwing, A. Perceptual score: What data modalities does your model perceive? *NeurIPS*, 34, 2021.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. *CVPR*, 2017.

Han, X., Wang, S., Su, C., Huang, Q., and Tian, Q. Greedy gradient ensemble for robust visual question answering. *ICCV*, 2021a.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. Trusted multi-view classification. *ICLR*, 2021b.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CVPR*, 2016.

Hessel, J. and Lee, L. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! *EMNLP*, 2020.

Hoffer, E., Banner, R., Golan, I., and Soudry, D. Norm matters: efficient and accurate normalization schemes in deep networks. *NeurIPS*, 2018.

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

Jabri, A., Joulin, A., and Van Der Maaten, L. Revisiting visual question answering baselines. *ECCV*, 2016.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L., and Koishida, K. Mmtm: multimodal transfer module for cnn fusion. *CVPR*, 2020.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. *CVPR*, 2019.

Lao, M., Guo, Y., Liu, Y., Chen, W., Pu, N., and Lew, M. S. From superficial to deep: Language bias driven curriculum learning for visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3370–3379, 2021.

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

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

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

Liu, K., Li, Y., Xu, N., and Natarajan, P. Learn to combine modalities in multimodal deep learning. *arXiv:1805.11730*, 2018.

Molchanov, P., Gupta, S., Kim, K., and Kautz, J. Hand gesture recognition with 3d convolutional neural networks. *CVPR*, 2015.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

Neverova, N., Wolf, C., Taylor, G., and Nebout, F. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. Multimodal deep learning. *ICML*, 2011.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. *AAAI*, 2018.

Pérez-Rúa, J.-M., Vielzeuf, V., Pateux, S., Baccouche, M., and Jurie, F. Mfas: Multimodal fusion architecture search. *CVPR*, 2019.

Sridharan, K. and Kakade, S. M. An information theoretic framework for multi-view learning. *COLT*, 2008.

Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. G. Multi-view convolutional neural networks for 3d shape recognition. *ICCV*, 2015.

Sun, Y., Mai, S., and Hu, H. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 2021.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3d convolutional networks. *ICCV*, 2015.

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

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? *CVPR*, 2020a.

Wang, Y., Huang, W., Sun, F., Xu, T., Rong, Y., and Huang, J. Deep multimodal fusion by channel exchanging. *NeurIPS*, 2020b.

Weng, Z., Wu, Z., Li, H., and Jiang, Y.-G. Hms: Hierarchical modality selectionfor efficient video recognition. *arXiv:2104.09760*, 2021.

Winterbottom, T., Xiao, S., McLean, A., and Moubayed, N. A. On modality bias in the tvqa dataset. *BMVC*, 2020.

Wu, N., Jastrzębski, S., Park, J., Moy, L., Cho, K., and Geras, K. J. Improving the ability of deep networks to use information from multiple views in breast cancer screening. *MIDL*, 2020.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3d shapenets: A deep representation for volumetric shapes. *CVPR*, 2015.

Zhang, G., Wang, C., Xu, B., and Grosse, R. Three mechanisms of weight decay regularization. *ICLR*, 2019.

A. 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 6.

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

Figure 6. (a) The Colored MNIST dataset (Kim et al., 2019). We consider the monochromatic 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.

B. Supplementary figures

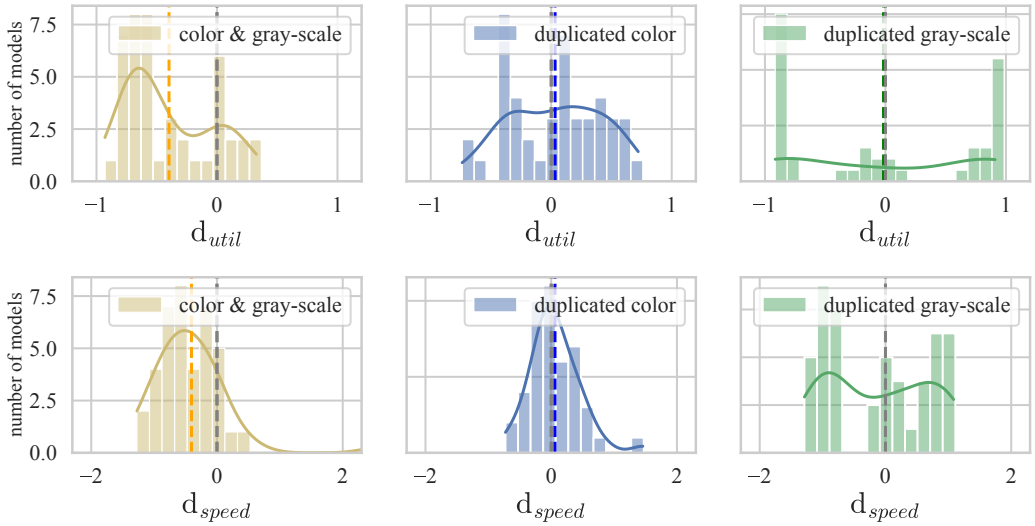


Figure 7. Histograms and estimated density functions of $\mathrm{d_{util}}$ and $\mathrm{d_{speed}}$ of models trained for colored-and-gray-MNIST, using monochromatic and gray-scale images as two modalities, using identical monochromatic images as two modalities and using identical gray-scale images as two modalities. Note that $\mathrm{d_{speed}}$ is not bounded by 1 as $\mathrm{d_{util}}$ is. Thus when learning from the two modalities is happening at very different paces, we can observe patterns shown in this study.

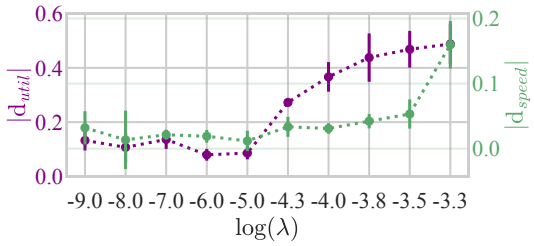
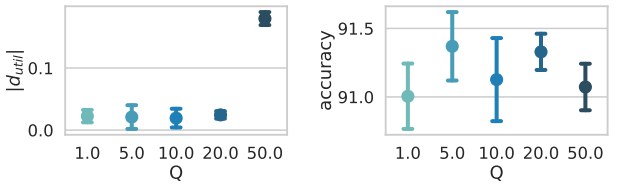
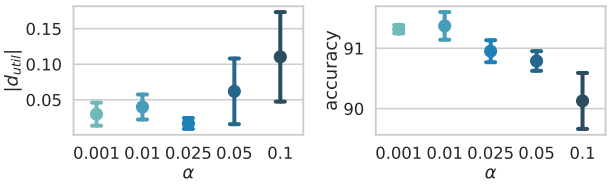


Figure 8. The observed $|\,\mathrm{d_{util}}|$ and $|\,\mathrm{d_{speed}}|$ for models trained with different weights (λ) on the L1 regularizer. We show $|\,\mathrm{d_{util}}|$ and $|\,\mathrm{d_{speed}}|$ as a function of $\log(\lambda)$. As we can see, when $\log(\lambda) \geq -5$, both $|\,\mathrm{d_{util}}(f)|$ and $|\,\mathrm{d_{speed}}|$ increases with λ increasing, .



(a) We train models using $Q \in \{1, 5, 10, 20, 50\}$ and $\alpha = 0.01$. Except $Q = 50$, models trained using other values of Q present relatively balanced conditional utilization rates between modalities (left panel). According to the models' generalization performance (right panel), we choose $Q = 5$ for the final experiment.



(b) We train models using $Q = 5$ and $\alpha \in \{0.01, 0.1, 0.25, 0.5, 1\} \times \mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$, where $\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}] = 0.1$ as shown in §6.2. When setting $\alpha \leq 0.25\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}]$, we can control $\mathrm{d_{util}}$ effectively (left panel). We choose to use $\alpha = 0.1\mathbb{E}[\widehat{\mathrm{d}}_{\mathrm{util}}] = 0.01$ based on the models' accuracy (right panel) for the final experiment.

Figure 9. Models' behavior when using different values for the imbalance tolerance parameter α and the re-balancing window size Q in the balanced multi-modal training algorithm. We use ModelNet40 (front and rear views) and fix the learning rate at 0.1 for the studies on both hyperparameters.