The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

SMIL: Multimodal Learning with Severely Missing Modality

Mengmeng Ma[1], **Jian Ren**[2], **Long Zhao**[3], **Sergey Tulyakov**[2], **Cathy Wu**[1], **Xi Peng**[1]

[1] University of Delaware, [2] Snap Inc., [3] Rutgers University

{mengma, wuc, xipeng}@udel.edu, {jren, stulyakov}@snap.com, lz311@cs.rutgers.edu

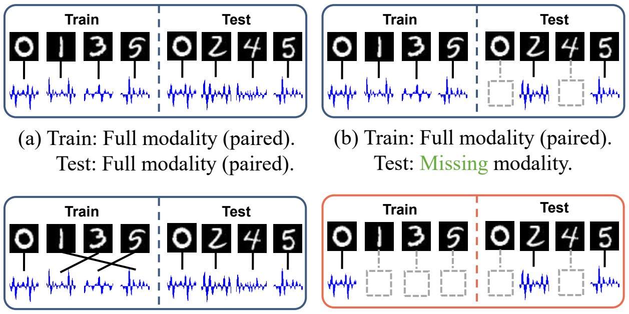
Abstract

A common assumption in multimodal learning is the completeness of training data, i.e., full modalities are available in all training examples. Although there exists research endeavor in developing novel methods to tackle the incompleteness of testing data, e.g., modalities are partially missing in testing examples, few of them can handle incomplete training modalities. The problem becomes even more challenging if considering the case of severely missing, e.g., ninety percent of training examples may have incomplete modalities. For the first time in the literature, this paper formally studies multimodal learning with missing modality in terms of flexibility (missing modalities in training, testing, or both) and efficiency (most training data have incomplete modality). Technically, we propose a new method named SMIL that leverages Bayesian meta-learning in uniformly achieving both objectives. To validate our idea, we conduct a series of experiments on three popular benchmarks: MM-IMDb, CMU-MOSI, and avMNIST. The results prove the state-of-the-art performance of SMIL over existing methods and generative baselines including autoencoders and generative adversarial networks.

Introduction

Multimodal learning attracts intensive research interest because of broad applications such as intelligent tutoring (Petrovica, Anohina-Naumeca, and Ekenel 2017), robotics (Noda et al. 2014), and healthcare (Frantzidis et al. 2010). Generally speaking, existing research efforts mainly focus on how to fuse multimodal data effectively (Liu et al. 2018; Zadeh et al. 2017a) and how to learn a good representation for each modality (Tian, Krishnan, and Isola 2020).

A common assumption underlying multimodal learning is the completeness of modality as illustrated in Figure 1. Existing methods (Ngiam et al. 2011; Zadeh et al. 2017b; Hou et al. 2019) often assume full and paired modalities are available in both training and testing data. However, such an assumption may not always hold in real world due to privacy concerns or budget limitations. For example, in social network, we may not be able to access full-modality data since users would apply various privacy and security constraints. In autonomous driving, we may collect many imaginary data but not as so for 3D point cloud because LiDARs are much less affordable than cameras.



(a) Train: Full modality (paired). Test: Full modality (paired).

(b) Train: Full modality (paired). Test: Missing modality.

(c) Train: Full modality (unpaired). Test: Full modality (paired).

(d) Train: Missing modality. Test: Missing modality.

Figure 1: Multimodal learning configurations. (a) Train and test with full and paired modality (Ngiam et al. 2011); (b) Testing with missing modality (Tsai et al. 2019); (c) Training with unpaired modality (Shi et al. 2020); (d) We study the most challenging configurations of severely missing modality in training, testing, or both.

Although there exist a bunch of research efforts (Tsai et al. 2019; Pham et al. 2019) in developing novel methods to tackle the incompleteness of testing data, few of them can handle incomplete training modalities. An interesting yet challenging research question then arises: Can we learn a multimodal model from an incomplete dataset while its performance should as close as possible to the one that learns from a full-modality dataset?

In this paper, we systematically study this problem by proposing multimodal learning with severely missing modality (SMIL). We consider an even more challenging setting that the missing ratio can be as much as 90%. More specifically, we design two objectives for SMIL: *flexibility* and *efficiency*. The former requires our model to uniformly tackle three different missing patterns in training, testing, or both. The latter enforces our model to effectively learn from incomplete modality as fast as possible.

To jointly achieve both objectives, we leverage Bayesian meta-learning framework in designing a new method. The key idea is to perturb the latent feature space so that embeddings of single modality can approximate ones of full modality. We highlight that our method is better than typical generative designs, such as *Autoencoder (AE)* (Tran et al. 2017), *Variational Autoencoder (VAE)* (Kingma and Welling 2013), or *Generative Adversarial Network (GAN)* (Goodfellow et al. 2014), since they often require a significant amount of full-modality data to learn from, which is usually not available in severely missing modality learning. To summarize, our contribution is three-fold:

- To the best of our knowledge, we are the first work to systematically study the problem of multimodal learning with severely missing modality.
- We propose a Bayesian meta-learning based solution to uniformly achieve the goals of *flexibility* (missing modalities in training, testing, or both) and *efficiency* (most training data have incomplete modality).
- Extensive experiments on MM-IMDb, CMU-MOSI, and avMNIST validate the state-of-the-art performance of SMIL over generative baselines including AE and GAN.

Related Work

Multimodal learning. Multimodal learning utilizes complementary information contained in multimodal data to improve the performance of various computer vision tasks. One important direction in this area is multimodal fusion, which focuses on effective fusion of multimodal data. Early fusion is a common method which fuses different modalities by feature concatenation, and it has been widely adopted in previous studies (Wang et al. 2017; Poria et al. 2016). Instead of concatenating features, Zadeh *et al.* (Zadeh et al. 2017b) proposed a product operation to allow more interactions among different modalities during the fusion process. Liu *et al.* (Liu et al. 2018) utilized modality-specific factors to achieve efficient low-rank fusion.

Recently, there have been a wide range of research interests in handling missing modalities for multimodal learning, such as testing-time modality missing (Tsai et al. 2019) and learning with data from unpaired modalities (Shi et al. 2020). In this paper, we tackle a more challenging and novel multimodal-learning setting where both training and testing data contain samples that have missing modalities. Generative approaches, such as auto-encoders (Tran et al. 2017; Lee et al. 2019), GANs (Goodfellow et al. 2014), and VAEs (Kingma and Welling 2013), offer a straightforward solution to handle this setting, but these methods are neithor flexible nor efficient as SMIL.

Meta-regularization. Meta-learning algorithms focus on designing models that are able to learn new knowledge and adapt to novel environments quickly with only a few training samples. Previous methods studied meta-learning from the perspective of metric learning (Koch 2015; Vinyals et al. 2016; Sung et al. 2018; Snell, Swersky, and Zemel 2017) or probabilistic modeling (Fe-Fei et al. 2003; Lawrence and Platt 2004). Recent advances in optimization-based approaches have evoked more interests in meta-learning. MAML (Finn, Abbeel, and Levine 2017) is a general optimization algorithm designed for few-shot learning and reinforcement learning. It is compatible with models that learn through gradient descent. Nichol *et al.* (Nichol, Achiam, and Schulman 2018) further improved the computation efficiency of MAML. Other works adapted MAML for domain generalization (Li et al. 2018; Qiao, Zhao, and Peng 2020) and knowledge distillation (Zhao et al. 2020). In this work, we extend MAML by learning two auxiliary networks for missing modality reconstruction and feature regularization.

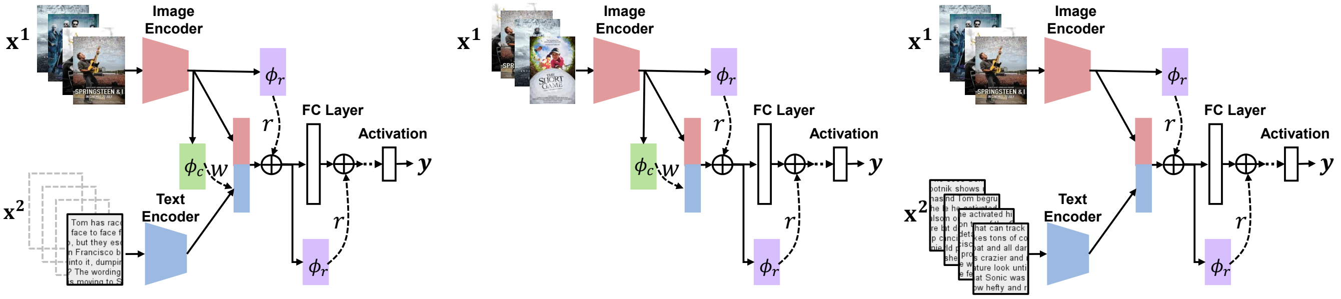
Conventional handcrafted regularization techniques (Hoerl and Kennard 1970; Tibshirani 1996) regularize model parameters to avoid overfitting and increase interpretability. Balaji *et al.* (Balaji, Sankaranarayanan, and Chellappa 2018) modeled the regularization function as an additional network learned through meta-learning to regularize model parameters. Li *et al.* (Li et al. 2019) followed the same idea of (Balaji, Sankaranarayanan, and Chellappa 2018) but learned an additional network to regularize latent features. Lee *et al.* (Lee et al. 2020b) proposed a more general algorithm for latent feature regularization. Other than perturbing features, we propose to learn the regularization function following (Lee et al. 2020b) but regularize the feature to reduce discrepancy between the reconstructed and true modality.

Multimodal generative models. Generative models for multimodal learning fall into two categories: cross-modal generation and joint-model generation. Cross-modal generation methods, such as conditional VAE (CVAE) (Sohn, Lee, and Yan 2015) and conditional multimodal autoencoder (Pandey and Dukkipati 2017), learn a conditional generative model over all modalities. On the other hand, joint-model generation approaches learn the joint distribution of multimodal data. Multimodal variational autoencoder (MVAE) (Wu and Goodman 2018) models the joint posterior as a product-of-expert (PoE). Multimodal VAE (JMVAE) (Suzuki, Nakayama, and Matsuo 2016) learns a shared representation with a joint encoder. With only a few modifications to the original algorithms, we show that multimodal generative models serve as strong baselines for learning with severely missing modalities proposed in this paper.

Proposed Method

We are interested in multimodal learning with severely missing modality, e.g., 90% of the training samples contain incomplete modalities. In this paper, without loss of generality, we consider a multimodal dataset containing two modalities. Formally, we let $\mathcal{D} = \{\mathcal{D}^f, \mathcal{D}^m\}$ denote a multimodal dataset; $\mathcal{D}^f = \{\mathbf{x}_i^1, \mathbf{x}_i^2, y_i\}_i$ is a modality-complete dataset, where $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ represent two different modalities of i-th sample and y_i is the corresponding class label; $\mathcal{D}^m = \{\mathbf{x}_j^1, y_j\}_j$ is a modality-incomplete dataset, where one modality is missing. Our target is to leverage both modality-complete and modality-incomplete data for model training. We propose to address this problem from two perspectives: *1) Flexibility:* how to uniformly handle missing modality in training, testing, or both? *2) Efficiency:* how to improve training efficiency when major data suffers from missing modality?

Flexibility. We aim to achieve a unified model that can handle missing modality in training, testing, or both. Our idea is to employ a feature reconstruction network to achieve this goal. Instead of following the conventional data reconstruction approaches (Lee et al. 2019; Tran et al. 2017),



(a) Training with severely missing modality (b) Testing with single modality (c) Testing with full modality

Figure 2: SMIL can uniformly learn from severely missing modality and test with either single or full modality. *The reconstruction network* ϕ_c outputs a posterior distribution, from which we sample *weight* ω to reconstruct the missing modality using modality priors. The regularization network ϕ_r also outputs a posterior distribution, from which we sample *regularizer* r to perturb latent features for smooth embedding. The collaboration (ϕ_c and ϕ_r) guarantees flexible and efficient learning.

the feature reconstruction network will leverage the available modality to generate an approximation of the missing-modality feature in a highly efficient way. This will generate complete data in the latent feature space and facilitate the flexibility in two aspects. On the one hand, our model can excavate the full potential of hybrid data by using both modality-complete and -incomplete data for joint training. On the other hand, when testing, by turning on or off the feature reconstruction network, our model can tackle modality-incomplete or -complete inputs in a unified manner.

Efficiency. We intend to train a model on the modality severely missing dataset to achieve comparable performance as the model trained on a full-modality dataset. However, the severely missing modality setting poses significant learning challenges to the feature reconstruction network. The network would be highly bias-prone due to the scarcity of modality-complete data, yielding degraded and low-quality feature generations. Directly train a model with degraded and low-quality features will hinder the efficiency of the training process. We propose a feature regularization approach to address this issue. The idea is to leverage a Bayesian neural network to assess the data uncertainty by performing feature perturbations. The uncertainty assessment is used as feature regularization to overcome model and data bias. Compared with previous deterministic regularization approaches (Balaji, Sankaranarayanan, and Chellappa 2018; Zhao et al. 2020), the proposed uncertainty-guided feature regularization will significantly improve the capacity of the multimodal model for robust generalization behaviors in tackling severely incomplete data.

A meta-learning framework. To effectively organize model training, we integrate the main network f_θ parameterized by $\boldsymbol{\theta}$, the reconstruction network f_{ϕ_c} parameterized by $\boldsymbol{\phi}_c$, and the regularization network f_{ϕ_r} parameterized by $\boldsymbol{\phi}_r$ in a modified *Model-Agnostic Meta-Learning (MAML)* (Finn, Abbeel, and Levine 2017) framework. An overview of our learning framework is shown in Figure 2. In the following sections, we describe the implementation of the feature reconstruction and regularization network.

Missing Modality Reconstruction

We introduce the feature reconstruction network to approximate the missing modality. For a modality-incomplete sample, the missing modality is reconstructed conditioned on the available modality. Given the observed modality $\mathbf{x}^1$, in order to obtain the reconstruction $\hat{\mathbf{x}}^2$ of the missing modality, we optimize the following objective for the reconstruction network:

$$\boldsymbol{\phi}_c^* = \arg\min_{\boldsymbol{\phi}_c} \mathbf{E}_{p(\hat{\mathbf{x}}^1,\mathbf{x}^2)}(-\log p(\hat{\mathbf{x}}^2|\mathbf{x}^1;\boldsymbol{\phi}_c)). \qquad (1)$$

However, under severely missing modality, it is non-trivial to train a reconstruction network from limited modality-complete samples. Inspired by (Kuo et al. 2019), we approximate the missing modality using a weighted sum of modality priors learned from the modality-complete dataset. In this case, the reconstruction network are trained to predict weights of the priors instead of directly generating the missing modality. We achieve this by learning a set of modality priors $\mathcal{M}$ which can be clustered among all modality-complete samples using K-means (MacQueen 1967) or PCA (Pearson 1901).

Specifically, let $\boldsymbol{\omega}$ represent the weights assigned to each modality prior. We model $\boldsymbol{\omega}$ as a multivariate Gaussian with fixed means and changeable variances as $\mathcal{N}(\mathbf{I},\boldsymbol{\sigma})$. The variances are predicted by the feature reconstruction network $\boldsymbol{\sigma} = f_{\phi_c}(\mathbf{x}^1)$. Given the weights $\boldsymbol{\omega}$, we can reconstruct the missing modality $\hat{\mathbf{x}}^2$ by calculating the weighted sum of the modality priors. Then, the reconstructed missing modality can be achieved by:

$$\hat{\mathbf{x}}^2 = \langle \boldsymbol{\omega}, \mathcal{M} \rangle \text{, where } \boldsymbol{\omega} \sim \mathcal{N}(\mathbf{I},\boldsymbol{\sigma}). \qquad (2)$$

We note that modeling $\boldsymbol{\omega}$ as multivariate random variables introduces randomness and uncertainty to the reconstruction process, which has been proved to be beneficial in learning sophisticated distributions (Lee et al. 2020b).

Uncertainty-Guided Feature Regularization

We propose to regularize the latent features by a feature regularization network. In each layer, the regularization network takes the features of the previous layer as input and applies regularization to the features of the current layer. Let $\mathbf{r}$ denote the generated regularization and $\mathbf{h}^l$ be the latent feature of the l-th layer. Instead of generating a deterministic regularization $\mathbf{r} = f_{\phi_r}(\mathbf{h}^{l-1})$, we assume that $\mathbf{r}$ follows a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, where the means and variances are calculated using $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ = $f_{\phi_r}(\mathbf{h}^{l-1})$. Then, we can compute the regularized feature by the following equation:

$$\mathbf{h}^l := \mathbf{h}^l \circ \mathrm{Softplus}(\mathbf{r}), \text{ where } \mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}), \tag{3}$$

where $\circ$ is a predefined operation (either addition or multiplication) for feature regularization. In our experiments, we observe that directly applying regularization to latent features will prevent the feature regularization network from convergence. Hence, we adopt Softplus (Dugas et al. 2000) activation to weaken the regularization.

A Bayesian Meta-Learning Framework

We leverage a Bayesian Meta-Learning framework to jointly optimizing all the networks. Specifically, we *meta-train* the main network f_θ on D^m with the help of reconstruction f_{ϕ_c} network and regularization f_{ϕ_r} network. Then, we *meta-test* the updated main network f_{θ^*} on D^f. Finally, we *meta-update* network parameters $\{\boldsymbol{\theta}, \boldsymbol{\phi}_c, \boldsymbol{\phi}_r\}$ by gradient descent.

For simplicity, we let $\boldsymbol{\psi} = \{\boldsymbol{\phi}_c, \boldsymbol{\phi}_r\}$ denote the combination of the parameters of the reconstruction and regularization network. Our framework aims to optimize the following objective function:

$$\begin{gathered}\min_{\boldsymbol{\theta}, \boldsymbol{\psi}} \mathcal{L}(\mathcal{D}^f; \boldsymbol{\theta}^*, \boldsymbol{\psi}), \\ \text{where } \boldsymbol{\theta}^* = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathcal{D}^m; \boldsymbol{\psi}).\end{gathered} \tag{4}$$

For the above function, $\mathcal{L}$ denotes the empirical loss such as cross entropy, and α is the inner-loop step size.

We use $\mathbf{X}$ and $\mathbf{Y}$ to represent all training samples and their corresponding labels, respectively. Let $\mathbf{z} = \{\boldsymbol{\omega}, \mathbf{r}\}$ be the collection of the generated weights and regularization. Then, inspired by (Finn, Xu, and Levine 2018; Gordon et al. 2019; Lee et al. 2020a), we define the generative process as optimizing the likelihood in a meta-learning framework:

$$p(\mathbf{Y}, \mathbf{z} | \mathbf{X}; \boldsymbol{\theta}) = p(\mathbf{z}) \prod_{i=1}^{N} p(\boldsymbol{y}_i | \mathbf{x}_i^1, \mathbf{x}_i^2, \mathbf{z}; \boldsymbol{\theta}) \prod_{j=1}^{M} p(\boldsymbol{y}_j | \mathbf{x}_j^1, \mathbf{z}; \boldsymbol{\theta}). \tag{5}$$

The goal of Bayesian Meta-Learning is to maximize the conditional likelihood: $\log p(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta})$. However, solving it involves the true posterior $p(\mathbf{z}|\mathbf{X})$, which is intractable. Instead, we approximate the true posterior distribution by an amortized distribution $q(\mathbf{z}|\mathbf{X}; \boldsymbol{\psi})$ (Finn, Xu, and Levine 2018; Gordon et al. 2019; Lee et al. 2020a). The resulting form of approximated lower bound for our meta-learning framework can be defined as:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\theta}, \boldsymbol{\psi}} = \boldsymbol{E}_{q(\mathbf{z}|\mathbf{X}; \boldsymbol{\theta}, \boldsymbol{\psi})}[\log p(\mathbf{Y}|\mathbf{X}, \mathbf{z}; \boldsymbol{\theta})] - \\ \mathrm{KL}[q(\mathbf{z}|\mathbf{X}; \boldsymbol{\psi}) \| p(\mathbf{z}|\mathbf{X})].\end{aligned} \tag{6}$$

We maximize this lower bound by Monte-Carlo (MC) sampling. After combining all these together, we obtain the full training objective of the proposed meta-learning framework for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ which is defined as:

$$\begin{gathered}\min_{\boldsymbol{\theta}, \boldsymbol{\psi}} \frac{1}{L} \sum_{l=1}^{L} -\log p(\boldsymbol{y}_j | \mathbf{x}_j^1, \mathbf{x}_j^2, \mathrm{z}_l; \boldsymbol{\theta}) + \mathrm{KL}[q(\mathbf{z}|\mathbf{X}; \boldsymbol{\psi}) \| p(\mathbf{z}|\mathbf{X})] \\ \text{with } \mathbf{z}_l \sim q(\mathbf{z}|\mathbf{X}; \boldsymbol{\psi}),\end{gathered} \tag{7}$$

where L is the number of MC sampling. We show our detailed algorithm in Algorithm 1.

Algorithm 1: Bayesian Meta-Learning Framework.

```
Input: Multimodal dataset $\mathcal{D} = \{D^f, D^m\}$; # of
       iterations K; inner learning rate $\alpha$; outer
       learning rate $\beta$.
while not converged do
    Sample $\{\mathbf{x}_j^1, y_j\} \sim D^m$; $\{\mathbf{x}_i^1, \mathbf{x}_i^2, y_i\} \sim D^f$
    $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}$
    Meta-train:
    for $k = 0$ to $K - 1$ do
        Sample $\tilde{\mathbf{z}}_j \sim p(\mathbf{z}_j | \mathbf{x}_j^1; \boldsymbol{\psi}, \boldsymbol{\theta}_k)$
        $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - \alpha \nabla_{\boldsymbol{\theta}_k}[-\log p(\boldsymbol{y}_j | \mathbf{x}_j^1, \tilde{\mathbf{z}}_j; \boldsymbol{\theta}_k)]$
    end
    $\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta}_K$
    Meta-test & Meta-update:
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \beta \nabla_{\boldsymbol{\theta}}[-\log p(\boldsymbol{y}_i | \mathbf{x}_i^1, \mathbf{x}_i^2, \tilde{\boldsymbol{z}}_i; \boldsymbol{\theta}^*)]$
    $\boldsymbol{\psi} \leftarrow \boldsymbol{\psi} - \beta \nabla_{\boldsymbol{\psi}}[-\log p(\boldsymbol{y}_i | \mathbf{x}_i^1, \mathbf{x}_i^2, \tilde{\boldsymbol{z}}_i; \boldsymbol{\theta}^*)]$
end
```

Experiments

In this section, we analyze the results of the proposed algorithm for multimodal learning with severely missing modality on three datasets from two perspectives: *efficiency under severely missing modality* (Section 4.2) and *flexibility to various modality missing pattern* (Section 4.3).[1]

Experiment Setting

Datasets. Totally three datasets are used in the experiment:

- *The Multimodal IMDb (MM-IMDb)* (Arevalo et al. 2017) contains two modalities: image and text. We conduct experiments on this dataset to predict a movie genre using image or text modality, which is a multi-label classification task as multiple genres could be assigned to a single movie. The dataset includes $25,956$ movies and 23 classes. We follow the training and validation splits provided in the previous work (Vielzeuf et al. 2018).
- *CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI)* (Zadeh et al. 2016) consists of $2,199$ opinion video clips from YouTube movie reviews. Each clip contains three modalities: the image modality includes the visual gesture, the text modality includes the transcribed speech, and the audio modality includes the automatic audio. We use the feature extraction model from Liu et al.

[1] Our code is available at https://github.com/mengmenm/SMIL

Method	Accuracy (%) ↑			F1 Score ↑		
	10%	20%	100%	10%	20%	100%
Lower-Bound	–	–	44.8	–	–	27.7
Upper-Bound	–	–	71.0	–	–	70.5
MVAE	–	–	58.5	–	–	58.1
AE	56.4	60.4	–	54.4	59.0	–
GAN	56.5	60.6	–	54.6	59.1	–
SMIL	**60.7**	**63.3**	–	**58.0**	**62.5**	–

Table 1: Binary classification accuracy (%) and F1 Score for different methods under three text modality ratios (10%, 20%, and 100%) on the *CMU-MOSI* dataset.

Method	F1 Samples ↑			F1 Micro ↑		
	10%	20%	100%	10%	20%	100%
Lower-Bound	–	–	47.6	–	–	48.2
Upper-Bound	–	–	61.7	–	–	62.0
MVAE	–	–	48.4	–	–	48.6
AE	44.5	50.9	–	44.8	50.7	–
GAN	45.0	51.1	–	44.6	51.0	–
SMIL	**49.2**	**54.1**	–	**49.5**	**54.6**	–

Table 2: Multi-label classification scores (F1 Samples and F1 Micro) for different methods under three text modality ratios (10%, 20%, and 100%) on the *MM-IMDb* dataset.

(2018) for each modality. We conduct experiments on this dataset to predict the sentiment class of the clips, which is a binary classification task as the sentiment of video clips can be either negative or positive. There are $1,284$ segments in the training set, 229 in the validation set, and 686 in the test set. In the experiment section, we only use the image and text modality.

- *Audiovision-MNIST (avMNIST)* (Vielzeuf et al. 2018) consists of an independent image and audio modalities. The images, which are digits from 0 to 9, are collected from the MNIST dataset (LeCun et al. 1998) with a size of 28×28, and the audio modality is collected from Free Spoken Digits Dataset [2] containing raw $1,500$ audios. We use the mel-frequency cepstral coefficients (MFCCs) (Tzanetakis and Cook 2002) as the representation of audio modality. Each raw audio is processed by MFCCs to get a sample with a size of $20 \times 20 \times 1$. The dataset contains $1,500$ samples for both image and audio modalities. We randomly select 70% data for training and use the rest for validation.

Evaluation metrics. For MM-IMDb dataset, we follow previous works (Arevalo et al. 2017; Vielzeuf et al. 2018) by adopting the F1 Samples and F1 Micro to evaluate multi-label classification. For CMU-MOSI, we follow Liu et al. (2018) to compute the binary classification accuracy and F1 Score. For avMNIST dataset, we compute accuracy to measure the performance.

Baseline methods. We compare the proposed approach with the following baseline methods:

- *Lower-Bound* is a model trained using single modality of the data, *i.e.,* 100% image, 100% text, etc. It serves as the lower bound for our method.
- *Upper-Bound* is a model trained leveraging all modalities of the data, *i.e.,* 100% images and 100% text, etc. We regard it as the upper bound.
- *AE* (Autoencoder) (Lee et al. 2019) is a deep model used for efficient data encoding. We can use AE to preprocess the original dataset to tackle the severely missing modality problem. We now describe the procedure for preprocessing. First, we sample a dataset containing only modality-complete samples from the original dataset. Then, we assume one modality is missing and train AE to reconstruct the missing modality. Finally, we impute the missing modality of modality-incomplete data using the trained AE. After finishing the imputation, the dataset is now available for multimodal learning.
- *GAN* (Generative adversarial network) is a deep generative model composed of a generator and a discriminator. We leverage GAN to tackle our problem following the same procedure as described in AE.
- *MVAE* (Wu and Goodman 2018) is proposed for multimodal generative task. We adopt the widely used linear evaluation protocol to adapt MVAE for classification. Specifically, we first train MVAE using all the modalities. We then keep the learned MVAE frozen to train a randomly initialized linear classifier using the latent representation generated by the encoder of MVAE.

Efficiency with Severely Missing Modality

Conclusion: *Our method demonstrates consistent efficiency, across different datasets, when training data contains a different ratios of modality missing.*

Setting of missing modality. We evaluate the efficiency of our algorithm on two datasets: MM-IMDb and CMU-MOSI. In both datasets, modalities are incomplete for some samples. We define the text modality ratio as $\eta = \frac{M}{N}$, where M is the number of samples with text modality and N is the size of overall samples. η indicates the severity of modality missing. The smaller of η, the severer the modality is missing. For both datasets, we assume image modality to be complete, and the text modality to be incomplete. We express all available data points in the form of 100% Image + η% Text for both datasets.

Implementation details. CMU-MOSI. We follow Liu et al. (2018) to get features for the image and text modality. We use three fully-connected (FC) layers with dimension 16 to get the embedding of image modality. One layer LSTM (Hochreiter and Schmidhuber 1997) extracts the embedding for text modality. The concatenated feature of two modalities is then fed to FC layers for classification. For training process, we use Adam (Kingma and Ba 2014) optimizer with a batch size of 32 and train the networks for $5,000$ iterations with a learning rate of 10^{-4} for both inner-loop and outer-loop of meta-learning. **MM-IMDB.** For im-

[2] https://github.com/Jakobovski/free-spoken-digit-dataset

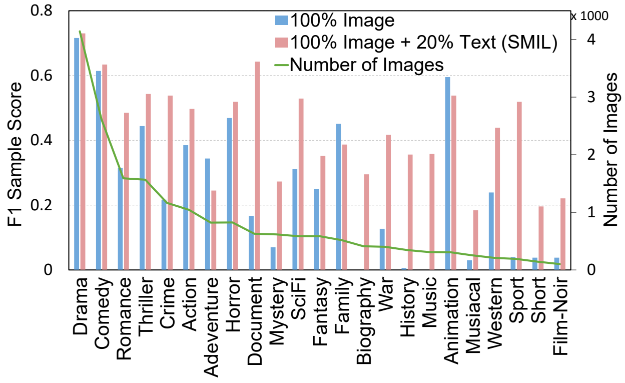


Figure 3: F1 Samples score of each movie genre on the *MM-IMDb* dataset for the lower-bound baseline (*blue*) and SMIL (*red*). The number of image samples for each movie genre is indicated by the green line.

age and text modalities, we adopt the feature extraction models from Arevalo et al. (2017). We feed the feature from each modality to a FC layer to align their output dimension. On top of it, we fuse the feature together and send it to FC layers to conduct multi-label classification. We apply Adam optimizer with a batch size of 128. We train the models for $10,000$ iteration with a learning rate of 10^{-4} for inner-loop and 10^{-3} fro outer-loop. Besides, we follow previous work (Vielzeuf et al. 2018) to add a weight of 2.0 on the positive label to balance the precision and recall since the labels are unbalanced.

Different ratios of modality missing. The results on CMU-MOSI are shown in Table 1. As can be seen, our approach significantly outperforms all baselines among all ratios of modality missing, which showcases the efficiency of our approach in the missing modality problem. The results also show that the severer the missing modality is, the more efficient our approach is. More specifically, when η is 20%, our approach outperforms AE and GAN around 5.0%, while the improvements increase to 7.6% and 7.4%, respectively, when η decreases to 10%. Moreover, our improvements are also consistent on MM-IMDb, as shown in Table 2. The improvement increases as the modality ratio decreasing. From Table 2, we see that our approach performs better than all baseline method under different text ratio. Our method outperforms Lower-Bound and MVAE by a large margin, and quite close to Upper-Bound.

We further show the effect of multimodal learning for different classes of MM-IMDb when $\eta = 20\%$ in Figure 3. First, our method (shown as red bars) can largely improve the model performance even on the tailed genres, such as *Sport* and *Film-Noir*, while the model trained only using images (shown as blue bars) can hardly predict the classes with less training samples. Second, an interesting phenomenon in Figure 3 is that text modality will slightly decrease the performance of movie genres like *Family* and *Animation*. The possible reason is that there is a large overlap between genres of family and animations. As a result, text modality may enforce the model to learn the shared knowledge between these two genres, which reduces the discrepancy and decrease the accuracy.

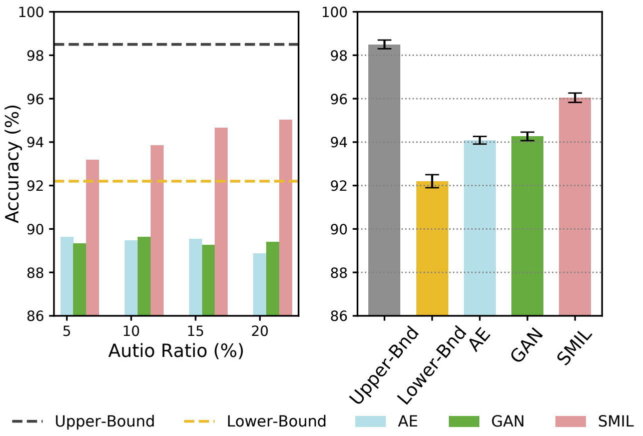


Figure 4: Classification accuracy (%) on *avMNIST* with two missing patterns. *Left*: training with 100% Image + η% Audio and testing with Image Only. *Right*: training with 100% Image + η% Audio and testing with Image + Audio.

Visualization of embedding space. We visualize the embedding space of three genres in MM-IMDb in Figure 5, and observed that our approach can effectively disentangle the latent embedding of the three genres, while the model learned only from image modality cannot. Besides, Our method is efficient when modality is severely missing. Form Figure 5, we see that our model trained using only 10% text modality is comparable to a model trained using 100% text modality.

Justification of symbol '-' used in Table 1, 2. We use the '-' symbol for two reasons. First, not applicable. Lower-Bound only requires image modality for training, so it is not applicable to report a Lower-Bound result trained using both image and text. Second, not necessary. For example, in table 1, MVAE trained without missing modality (100% image + 100% text) achieves acc = 58.5%. In comparison, our model trained with severely missing modality (100% image + 10% text) achieves acc = 60.7%. So it is not necessary to train MVAE under severely missing modality.

Flexibility with Different Missing Patterns

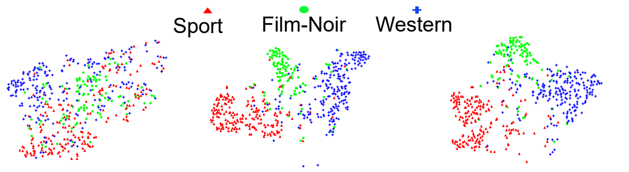
Conclusion: *Our method shows flexibility in handling various missing patterns: (1) full or missing modality at training; and (2) full or missing modality at test time.*

Implementation details. Our network contains two modality-specific feature extractors and a few FC layers. We use LeNet-5 to extract features for image modality, and a modified LeNet-5 to extract audio features. Extracted features are then fused through concatenation and sent into FC layers to perform classification. For the training process, we use Adam optimizer with a batch size of 64 and train the networks for $15,000$ iterations with a learning rate of 10^{-3} for both inner- and outer- loop of meta-learning.

Setting of missing pattern. For the avMNIST dataset, the missing modality problem only happens to audio modality.

Method	F1 Samples ↑		F1 Micro ↑	
	10%	20%	10%	20%
SMIL w/o K-means	0.482	0.535	0.485	0.530
SMIL w/o Regularization	0.469	0.521	0.472	0.530
SMIL w/ Fixed Gaussian	0.475	0.495	0.479	0.502
SMIL w/ Deterministic	0.474	0.527	0.477	0.533
SMIL (Full)	**0.492**	**0.541**	**0.495**	**0.546**

Table 3: Ablation study on the effect of modality reconstruction, feature regularization, and Bayesian inference on *MM-IMDb* under two text modality ratios (10% and 20%).



(a)100%Image (Lower-Bound) (b)100%Image + 20%Text (SMIL) (c)100%Image + 100%Text (Upper-Bound)

Figure 5: t-SNE visualization for embeddings of the lower-bound baseline (a), SMIL (b), and upper-bound baseline (c) on the *MM-IMDb* dataset. Three movie genres, including `Sport`, `Film-Noir`, and `Western` are visualized.

We are interested in two different missing patterns: (1) training with 100% Image + η% Audio and testing with Image Only; (2) training with 100% Image + 20% Audio and testing with Image + Audio. In this section, we show that our approach can flexibly handle these two missing patterns.

Missing pattern 1: testing with image only. Figure 4 (left) shows the classification accuracy under different audio ratio. We see that our approach can successfully handle testing with image modality only, but baseline methods such as AE and GAN fail in this scenario. As can be seen, when $\eta = 20\%$, SMIL is 6.7% higher than the generative-based method, and 3.3% higher the Lower-Bound. We argue that the failure of baseline methods is mainly due to the bias of the reconstructed missing modality. In single modality testing, the method is required to generate the missing modality conditioned on the available modality. The baseline method does not consider the bias of the reconstructed missing modality. In contrast, our method can leverage learned meta-knowledge to generate an unbiased missing modality. Besides, in situations where audio modality is missing severely (*i.e.*, $\eta = 5\%$), The classification accuracy of our method is 1.10% higher than the lower bound. The improvement demonstrates clear advantages of our model under severely missing modality.

Missing pattern 2: testing with image and audio. Figure 4 (right) shows the result of our approach dealing with full modality testing. We observe that our method still performs the best. It outperforms the Lower-Bound by 4.3% and the generative-based method (AE and GAN) by 2.1%. Moreover, under different missing patterns, SMIL is consistently better than AE and GAN. When switching testing patterns from two modalities to a single modality, AE and GAN have a 5.6% performance drop, while SMIL only has a 1.0% performance drop.

Ablation Study

We conduct the ablation analysis on the MM-IMDb dataset to evaluate the effectiveness of the missing modality reconstruction, feature regularization, and Bayesian Inference. We show the results in Table 3.

Effectiveness of missing modality reconstruction. In Section , we use reconstruction network to generate weights for missing modality reconstruction. Here we denote the method that uses the reconstruction network to directly generate the feature of missing modality as *SMIL w/o K-means*, which has worse performance and proves the necessity of K-Means for reconstruction.

Effectiveness of feature regularization. In Section , we introduce feature regularization. Here we denote the method without feature regularization as *SMIL w/o Regularization*. The performance of *SMIL w/o Regularization* is inferior to *SMIL (Full)*, which verifies conducting multimodal learning on D without regularization leads to a sub-optimal model. The superior performance of the regularized model is essential to the explicit objective of reducing discrepancy.

Effectiveness of Bayesian inference. In Section , we introduce the Bayesian Meta-Learning Framework. In this section, we compare it with two variants. *SMIL w/ Fixed Gaussian*: We fix the distribution of feature regularization to a Gaussian distribution, which is $\mathcal{N}(\mathbf{0}, \mathbf{I})$; *SMIL w/ Deterministic*: The missing modality construction and feature regularization is deterministic so the sampling in Eqn. 7 is removed. These two variants are inferior to Bayesian inference.The results show the superiority of Bayesian Meta-Learning framework.

Conclusion

In this paper, we address a challenging and novel problem in multimodal learning: multimodal learning with severely missing modality. We further propose a novel learning strategy based on the meta-learning framework. This framework tackles two important perspectives: missing modality reconstruction (flexibility) and feature regularization (efficiency). We apply the Bayesian meta-learning framework to infer the posterior of them and propose a variational inference framework to estimate the posterior.

In the experiments, we show that our model outperforms the generative method significantly on three multimodal datasets. Further analysis on the results shows that involving modality reconstruction and feature regularization can effectively handle the missing modality problem and flexible to various missing patterns. We believe that our work makes a meaningful step towards the real-world application of multimodal learning where partial modalities are missing or hard to collect.

Acknowledgements

This work is partially supported by the Data Science Institute (DSI) at University of Delaware and Snap Research.

References

Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; and González, F. A. 2017. Gated Multimodal Units for Information Fusion. In *5th International conference on learning representations 2017 workshop*.

Balaji, Y.; Sankaranarayanan, S.; and Chellappa, R. 2018. Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, 998–1008.

Dugas, C.; Bengio, Y.; Bélisle, F.; Nadeau, C.; and Garcia, R. 2000. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems* 13: 472–478.

Fe-Fei, L.; et al. 2003. A Bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings Ninth IEEE International Conference on Computer Vision*, 1134–1141. IEEE.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*.

Finn, C.; Xu, K.; and Levine, S. 2018. Probabilistic model-agnostic meta-learning. In *Advances in Neural Information Processing Systems*, 9516–9527.

Frantzidis, C. A.; Bratsas, C.; Klados, M. A.; Konstantinidis, E.; Lithari, C. D.; Vivas, A. B.; Papadelis, C. L.; Kaldoudi, E.; Pappas, C.; and Bamidis, P. D. 2010. On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine* 14(2): 309–318.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gordon, J.; Bronskill, J.; Bauer, M.; Nowozin, S.; and Turner, R. 2019. Meta-Learning Probabilistic Inference for Prediction. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=HkxStoC5F7.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.

Hoerl, A. E.; and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1): 55–67.

Hou, M.; Tang, J.; Zhang, J.; Kong, W.; and Zhao, Q. 2019. Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling. In *Advances in Neural Information Processing Systems*, 12113–12122.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Koch, G. 2015. Siamese neural networks for one-shot image recognition. In *International Conference on Machine Learning Deep Learning workshop*.

Kuo, W.; Angelova, A.; Malik, J.; and Lin, T.-Y. 2019. Shapemask: Learning to segment novel objects by refining shape priors. In *Proceedings of the IEEE International Conference on Computer Vision*, 9207–9216.

Lawrence, N. D.; and Platt, J. C. 2004. Learning to learn with the informative vector machine. In *Proceedings of the twenty-first international conference on Machine learning*, 65.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

Lee, H. B.; Lee, H.; Na, D.; Kim, S.; Park, M.; Yang, E.; and Hwang, S. J. 2020a. Learning to Balance: Bayesian Meta-Learning for Imbalanced and Out-of-distribution Tasks. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=rkeZIJBYvr.

Lee, H. B.; Nam, T.; Yang, E.; and Hwang, S. J. 2020b. Meta Dropout: Learning to Perturb Latent Features for Generalization. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=BJgd81SYwr.

Lee, H.-C.; Lin, C.-Y.; Hsu, P.-C.; and Hsu, W. H. 2019. Audio Feature Generation for Missing Modality Problem in Video Action Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3956–3960. IEEE.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Li, Y.; Yang, Y.; Zhou, W.; and Hospedales, T. M. 2019. Feature-critic networks for heterogeneous domain generalization. *arXiv preprint arXiv:1901.11448* .

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064* .

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 281–297. Berkeley, Calif.: University of California Press. URL https://projecteuclid.org/euclid.bsmsp/1200512992.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning*.

Nichol, A.; Achiam, J.; and Schulman, J. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* .

Noda, K.; Arie, H.; Suga, Y.; and Ogata, T. 2014. Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems* 62(6): 721–736.

Pandey, G.; and Dukkipati, A. 2017. Variational methods for conditional multimodal deep learning. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 308–315. IEEE.

Pearson, K. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2(11): 559–572.

Petrovica, S.; Anohina-Naumeca, A.; and Ekenel, H. K. 2017. Emotion recognition in affective tutoring systems: Collection of ground-truth data. *Procedia Computer Science* 104: 437–444.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6892–6899.

Poria, S.; Chaturvedi, I.; Cambria, E.; and Hussain, A. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, 439–448. IEEE.

Qiao, F.; Zhao, L.; and Peng, X. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12556–12565.

Shi, Y.; Paige, B.; Torr, P. H.; and Siddharth, N. 2020. Relating by Contrasting: A Data-efficient Framework for Multimodal Generative Models. *arXiv preprint arXiv:2007.01179* .

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, 4077–4087.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, 3483–3491.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Suzuki, M.; Nakayama, K.; and Matsuo, Y. 2016. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891* .

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=SkgpBJrtvS.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1): 267–288.

Tran, L.; Liu, X.; Zhou, J.; and Jin, R. 2017. Missing Modalities Imputation via Cascaded Residual Autoencoder. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2019. Learning Factorized Multimodal Representations. In *International Conference on Learning Representations*.

Tzanetakis, G.; and Cook, P. 2002. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing* 10(5): 293–302.

Vielzeuf, V.; Lechervy, A.; Pateux, S.; and Jurie, F. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision*.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, 3630–3638.

Wang, H.; Meghawat, A.; Morency, L.-P.; and Xing, E. P. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 949–954. IEEE.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 5575–5585.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017a. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Empirical Methods in Natural Language Processing, EMNLP*.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017b. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* .

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* .

Zhao, L.; Peng, X.; Chen, Y.; Kapadia, M.; and Metaxas, D. N. 2020. Knowledge as Priors: Cross-Modal Knowledge Generalization for Datasets without Superior Knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6528–6537.