Towards Balanced Active Learning for Multimodal Classification

Meng Shen
Nanyang Technological University
Singapore
meng005@e.ntu.edu.sg

Yizheng Huang
Institute for Infocomm Research
A*STAR
Singapore
huangyz0918@ieee.org

Jianxiong Yin
NVIDIA AI Tech Center
Singapore
jianxiongy@nvidia.com

Heqing Zou
Nanyang Technological University
Singapore
heqing001@e.ntu.edu.sg

Deepu Rajan
Nanyang Technological University
Singapore
asdrajan@ntu.edu.sg

Simon See
NVIDIA AI Tech Center
Singapore
ssee@nvidia.com

ABSTRACT

Training multimodal networks requires a vast amount of data due to their larger parameter space compared to unimodal networks. Active learning is a widely used technique for reducing data annotation costs by selecting only those samples that could contribute to improving model performance. However, current active learning strategies are mostly designed for unimodal tasks, and when applied to multimodal data, they often result in biased sample selection from the dominant modality. This unfairness hinders balanced multimodal learning, which is crucial for achieving optimal performance. To address this issue, we propose three guidelines for designing a more balanced multimodal active learning strategy. Following these guidelines, a novel approach is proposed to achieve more fair data selection by modulating the gradient embedding with the dominance degree among modalities. Our studies demonstrate that the proposed method achieves more balanced multimodal learning by avoiding greedy sample selection from the dominant modality. Our approach outperforms existing active learning strategies on a variety of multimodal classification tasks. Overall, our work highlights the importance of balancing sample selection in multimodal active learning and provides a practical solution for achieving more balanced active learning for multimodal classification.

CCS CONCEPTS

• **Computing methodologies → Active learning settings**.

KEYWORDS

active learning, multimodal learning

ACM Reference Format:
Meng Shen, Yizheng Huang, Jianxiong Yin, Heqing Zou, Deepu Rajan, and Simon See. 2023. Towards Balanced Active Learning for Multimodal Classification. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3612463

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612463

1 INTRODUCTION

Multimodal classification, as one of the classical multimodal learning tasks, aims to exploit complementary information inherent in multimodal data to achieve better classification performance. To this end, deep learning strategies have been implemented to train large-scale multimodal deep neural networks [4, 15]. However, such networks require an enormous amount of data to learn from, given their huge number of parameters. To reduce data cost, active learning (AL) is used to select a subset of more informative and distinctive unlabeled data samples for label assignment by oracles. Consequently, large networks can maintain performance while utilizing a smaller labeling budget. Most existing active learning algorithms are designed for unimodal tasks such as image classification [5, 31], object detection [21, 45] and language modeling [24, 44]. The objective is to select samples that have high uncertainty in them, carry novel knowledge for model training and those with distinctive features. However, there has been significantly less research reported on the design of effective active learning strategies for multimodal learning [29].

In this paper, we initially examine the performance of existing active learning strategies in selecting multimodal data. Our experiments reveal that these strategies tend to focus more on the dominant modality rather than fairly considering all modalities. For instance, in an image-text classification task, if the text contributes more to model optimization, active learning strategies may exhibit a bias towards the more distinguishable text modality by selecting valuable text samples and disregarding the informativeness of image samples. As a result, the selected multimodal dataset could become unbalanced, with insufficient information from the image modality, potentially leading to a degraded image model backbone. Recent works [18, 28, 40, 43] point out that balancing the training and optimization of all modalities is a key factor for successful multimodal learning. Similarly, it is crucial to design active learning strategies that can select multimodal data with fairness among all modalities to assist balanced multimodal learning.

Based on our findings, we develop a **B**alanced **M**ulti**m**odal **A**ctive **L**earning (**BMMAL**) algorithm that selects multimodal data by fairly considering each modality present in the data. In our approach, we choose the gradient embedding of model parameters, as it reflects the impact on model training and captures the diversity of data samples. However, we examine how the previous gradient embedding method [3] fails to select balanced multimodal data. To ensure fairness, we individually assess the contribution of each modality feature by examining the Shapley value, which attributes its contribution to the final multimodal prediction. We then apply modulation on the gradient embedding to penalize samples with dominant modalities. Lastly, a clustering seed initialization algorithm is employed to select diverse multimodal data with a significant influence on model training.

In summary, our main contributions are as follows:

- We empirically show that most existing active learning strategies fail to select a balanced multimodal dataset. We analyze how to improve the current gradient embedding based active learning strategy to rectify this.
- We propose a method to modulate the gradient embedding on sample-level to select more balanced multimodal candidates.
- We conduct experiments on three multimodal datasets to show that our proposed method treats multimodal data more equally and achieves better performance.

2 RELATED WORKS

2.1 Active Learning

Uncertainty-aware strategies attempt to utilize the data uncertainty or the model uncertainty as a criterion to locate unlabeled data points that the current model has less confidence about. One strategy is to utilize the posterior classification probability distribution by measuring its entropy [32, 42], or the margin between the most confident class and the second most confident class [30]. In addition, uncertainty can be evaluated as the variance of predictions generated by an ensemble of models [5] or by multiple inferences with Monte-Carlo dropout as an alternative Bayesian approximation for static networks [12]. Moreover, ALFA-Mix [27] evaluates unlabeled samples by mixing their features with labeled samples and observing whether there is inconsistency among predictions from mixed features. DFAL [10] incorporates adversarial attack techniques [25] to select unlabeled data samples located close to the classification boundaries.

Diversity-aware strategies tend to select unlabeled data points whose features are as diverse as possible to minimize data redundancy. [26] utilizes K-medoid algorithm [19] to select representative data centroids that minimize the total distance from other data samples to the nearest centroids. CoreSet [31] greedily selects unlabeled data samples that have maximum distances from their nearest neighbors. [6] adopts the determinantal point process (DPP) to evaluate the diversity by calculating the determinant of the similarity matrix. Diversity-aware strategies can also be considered in the context of distribution matching, which aims to reduce the gap between the distributions of labeled and unlabeled samples in latent space or feature space. VAAL [35] trains a variational auto-encoder to construct the latent distribution of labeled samples and an adversarial network to distinguish labeled samples and unlabeled samples in the latent space. Moreover, the maximum mean discrepancy (MMD) [39], the $\mathcal{H}$-divergence [36] and the Wasserstein distance [34] are used to measure the distribution gap.

To achieve a better trade-off between informativeness and diversity, hybrid methods are developed with an awareness of both. Since diversity-aware strategies are orthogonal to most of uncertainty-aware strategies [16], they could be easily combined together. ALFA-Mix [27] adopts K-means clustering to further filter out samples to enhance diversity. BADGE [3] represents unlabeled data samples via gradient embedding of parameters of the last classifier layer and applies K-means++ [2] to form a diverse data selection which still carries high uncertainty.

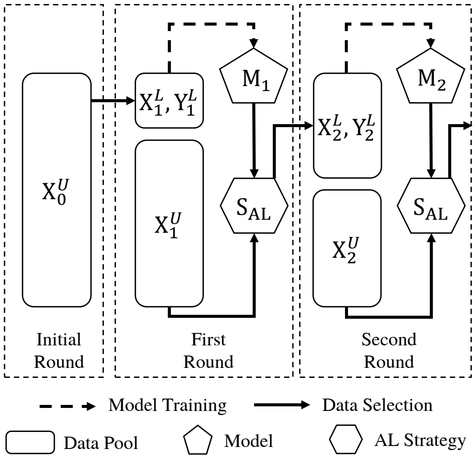


Figure 1: General active learning process. The dashed lines represent model training. The solid lines represent data selection.

2.2 Balanced Multimodal Learning

Our work considers joint multimodal learning for classifications. Here, it has been found that the best unimodal networks could potentially outperform multimodal networks regardless of fusion mechanisms or regularization methods [40]. Recent works show that the degradation of multimodal learning could be due to unbalanced optimization among different modalities. In [18], the failure of multimodal learning is attributed to modality competition where only dominant modalities are fully explored by joint training. Similarly, [43] demonstrates that multimodal learning greedily optimizes the dominant modalities and chooses to balance their training speeds. [40] propose to blend gradients with weights that are disproportional to the overfitting and generalization ratio of each modality so that each modality could be optimized in a balanced manner. [28] finds that fusion mechanisms such as concatenation and summation encourage the dominant modality to learn faster and thus develops gradient modulation to adaptively balance the training speed of each modality.

3 METHODOLOGY

3.1 Multimodal Active Learning Framework

The general active learning process is shown in **Figure 1**. Initially, we are given a large unlabeled data pool $X_0^U = \{(x_{m_1}, \ldots, x_{m_M})_{1\ldots n}\}$ of n input data with M modalities and an empty labeled data pool $X_0^L = \emptyset$. The labeling budget of each round is set to B. In the first round of active learning, since there is no trained model to evaluate with, a subset X_1^L containing B multimodal data is randomly selected from X_0^U, and they will be assigned with true labels Y_1^L. After data selection, the unlabeled dataset becomes $X_1^U = X_0^U \setminus X_1^L$. The training dataset for the first round of model training consists of X_1^L and Y_1^L. Starting from the second round, an active learning strategy S_{AL} evaluates the trained model and unlabeled data in the last round using an acquisition function and selects a batch of candidates for label assignment to construct a new training dataset for the current round of model training. The processes of data selection and model training continue until the total labeling budget is run out or the target performance of the trained model is reached.

We then introduce our multimodal learning framework for classification task. x_{m_1} and x_{m_2} represent the input data from two different modalities. They are processed through encoders φ_{m_1} and φ_{m_2} respectively to extract unimodal features $z_{m_1} \in \mathbb{R}^{D_{m1}}$ and $z_{m_2} \in \mathbb{R}^{D_{m_2}}$. We adopt concatenation, a wildly used late-fusion mechanism, to construct multimodal features $z_{mm} = z_{m_1} \oplus z_{m_2}$.[1] The unimodal and multimodal features are fed to unimodal classifiers C_{m_1}, C_{m_2} and multimodal classifier C_{mm} respectively to produce logits f_{m_1}, f_{m_2} and f_{mm} for classification. The final loss is the average cross-entropy loss $\mathcal{L}_{CE}$ of unimodal and multimodal logits with true labels y:

$$\mathcal{L}_{final} = \frac{1}{3}[\mathcal{L}_{CE}(f_{m_1}, y) + \mathcal{L}_{CE}(f_{m_2}, y) + \mathcal{L}_{CE}(f_{mm}, y)]. \quad (1)$$

Once the model is trained, the unlabeled data samples are evaluated using an acquisition function and filtered for labeling.

3.2 Analysis of Imbalance in AL

We introduce one of the state-of-the-art active learning algorithm BADGE [3] and provide analysis of its imbalanced data selection over multimodal data samples. BADGE was the first to propose the replacement of features for embedding with the gradient of the weight of the last FC layer, which acts as the classifier. In our case, the last FC layer for multimodal classification is the multimodal classifier C_{mm}. The weight of classifier W is a 2-dimensional matrix of size $K \times D_{mm}$, where K is the number of classes and D_{mm} is the dimension of concatenated multimodal feature $D_{m_1} + D_{m_2}$. The corresponding multimodal cross-entropy loss can be expanded as

$$\begin{aligned} \mathcal{L}_{mm} &= -\sum_{i=1}^{K} y_i \cdot log\sigma(f_{mm})_i \\ &= -\sum_{i=1}^{K} y_i \cdot log\frac{e^{z_{mm}\cdot W_i^T}}{\sum_{i=1}^{K} e^{z_{mm}\cdot W_i^T}}, \end{aligned} \quad (2)$$

where σ is softmax function and $z_{mm} \cdot W_i^T$ is the i^{th} element of logits f_{mm}. The gradient embedding is defined as $g = \frac{\partial \mathcal{L}_{mm}}{\partial W}$, and it is a 2-D matrix of size $K \times D_{mm}$ where the i^{th} row is

$$g_i = (f_i - 1_{\hat{y}_{mm}=i})z_{mm}, \quad (3)$$

where $\hat{y}_{mm} = \underset{i\in[K]}{\text{argmax}}[(f_{mm})_i]$ is the pseudo label for unlabeled data samples. The gradient embedding is flattened into a vector for sampling. It not only carries the uncertainty of classification from the margin between logits f_i and pseudo labels $\hat{y}_{mm}$, but also is representative enough due to the information present in z_{mm}.

However, in multimodal learning settings, identifying the source of uncertainty can be challenging. Upon examining the calculation of multimodal logits $f_i = z_{mm} \cdot W_i^T = z_{m_1} \cdot (W_i)_{m_1}^T + z_{m_2} \cdot (W_i)_{m_2}^T$, where W_i is divided into two matrices $(W_i)_{m_1}$ and $(W_i)_{m_2}$, it is difficult to determine which modality carries more uncertainty and which carries less. To illustrate, for a visual event such as drawing, the visual modality contains more information and contributes more to multimodal logits by generating a larger output. The multimodal uncertainty calculation is thus skewing the visual uncertainty instead of considering both visual and auditory uncertainties fairly. From **Section 4.4**, we find that BADGE does pay more attention to the dominant modality, which might potentially damage the performance of joint multimodal learning. Another limitation of BADGE is its inability to distinguish modality contributions. For instance, given two data samples with identical logits, we should prioritize the one with a more balanced contribution during data selection to facilitate balanced multimodal learning. However, the current BADGE algorithm cannot achieve this. Similarly, most conventional active learning algorithms lack this capability.

Hence, we develop a balanced multimodal active learning method that could avoid biased data selection towards the dominant modality to mitigate modality competition and assure that the trained multimodal network would not easily degenerate to the dominant modality. While our designed method is encouraged to pay more attention to the weaker modality, it is essential to ensure that it does not overly lean towards the weaker modality, as this may also harm the multimodal classification performance.

3.3 Guidelines to Design Balanced MMAL

To make existing AL strategies more suitable for balanced multimodal learning, it is necessary to inspect the individual modality contribution and reduce the contribution gap among different modalities. We empirically propose three guidelines for designing active learning strategies that treat each modality more equally. Let $\Phi_{m_i}(x)$ represent the contribution of the i^{th} modality of data sample x to the final model outcome , which should satisfy:

$$\sum_{i=1}^{M} \Phi_{m_i}(x) = 1. \quad (4)$$

We introduce the dominance degree $\rho(x)$ to quantify how severely a data sample x is dominated by the strongest modality:

$$\rho(x) = \sum_{i=1}^{M}[max(\Phi_{m_1}(x), ..., \Phi_{m_M}(x)) - \Phi_{m_i}(x)]. \quad (5)$$

We further partition the entire unlabeled dataset into multiple subsets $X = \{X_1, ..., X_M\}$ for the ease of discussion. In each subset X_i, modality m_i contributes the most:

$$\Phi_{m_i}(x) \geq \Phi_{m_j}(x), i \neq j, \forall x \in X_i. \quad (6)$$

Guideline 1: For two multimodal data samples x_i and x_j, if their acquisition scores of conventional active learning (CAL) strategies are equal, the one with more balanced unimodal contributions

[1] Other fusion mechanisms such as summation and NL-gate are implemented in our further experiments.

should have higher acquisition scores of balanced multimodal active learning strategies,

$$\begin{aligned} &a_{BMMAL}(x_i, \rho_i) > a_{BMMAL}(x_j, \rho_j), \rho_i < \rho_j, \\ &\text{where } a_{CAL}(x_i) = a_{CAL}(x_j), i \neq j. \end{aligned} \tag{7}$$

By following Guideline 1, data samples with more equal unimodal contributions are more likely to be selected. However, this does not guarantee that the stronger modality will be suppressed, nor does it ensure that the weaker modality will not be overly encouraged. Therefore, we introduce two additional guidelines.

Guideline 2: To avoid biased data selection favoring the stronger modality, the gap between the average acquisition scores of data samples dominated by the stronger modality and those dominated by the weaker modality should be reduced. In a two-modality case, where m_1 is the weaker modality and m_2 is the stronger modality (i.e. the average contribution of m_1 over the entire dataset is less than that of m_2, $\frac{1}{|X|}\sum_{x \in X} \Phi_{m_1}(x) < \frac{1}{|X|}\sum_{x \in X} \Phi_{m_2}(x)$), we have

$$\frac{\frac{1}{|X_1|}\sum_{x \subseteq X_1} a_{CAL}(x)}{\frac{1}{|X_2|}\sum_{x \subseteq X_2} a_{CAL}(x)} < \frac{\frac{1}{|X_1|}\sum_{x \subseteq X_1} a_{BMMAL}(x)}{\frac{1}{|X_2|}\sum_{x \subseteq X_2} a_{BMMAL}(x)}. \tag{8}$$

Guideline 3: Lastly, to prevent biased data selection towards the weaker modality, it is necessary to ensure that the contribution of each modality to the acquisition score function a_{BMMAL} is still proportional to its modality contribution to the model outcome on the sample-level. It ensures that the data samples are selected in a way that fairly represents the contributions of each modality to the actual model outcome.

In summary, Guideline 1 prioritizes the samples with more equal unimodal contributions. Guideline 2 and 3 work together to punish the stronger modality on the dataset-level but maintain the relationship between strong and weak modality on the sample-level, avoiding biases towards either the stronger or weaker modalities.

3.4 Estimate Modality Contribution

We show how we compute modality contribution Φ. In the context of multimodal classification, balanced active learning should select data samples that fairly contribute to the performance of all modalities. To achieve this, it is essential to estimate the degree to which each modality of a given data sample contributes to the final multimodal prediction. One approach involves assessing modality importance by computing the disparity in model performance before and after the incorporation of a particular modality. Researchers have proposed various techniques to remove the information of one modality, such as masking [11], permutation [14], and empirical multimodally-additive projection (EMAP) [17]. Nonetheless, these attribution methods are ill-suited for active learning as they require ground truth labels to calculate model performance metrics, such as accuracy. As a result, these methods cannot be employed for estimating modality contribution for unlabeled data due to the absence of ground truth labels.

Therefore, we choose to use the Shapley value to estimate modality contribution without the need for true labels. The Shapley value [33] was proposed to fairly attribute payouts among group of cooperative players based on their contributions to the total payout in game theory. In deep learning, SHapley Additive exPlanations (SHAP) value [23] considers each feature as a player and the model prediction as the total payout to estimate feature contributions. Let $\mathcal{M} = \{z_{m_1}, ..., z_{m_M}\}$ represent the set of all modality features, $\mathcal{S}$ denote the subset, and V symbolize the model outcome. Here, we use features instead of raw data inputs since features are utilized in active learning. To estimate the Shapley value of i^{th} modality feature z_{m_i}, we compute the marginal contribution to the subset $\mathcal{S}$ and average over all possible subset selections:

$$\phi(z_{m_i}) = \sum_{\mathcal{S} \subseteq \mathcal{M} \setminus \{z_{m_i}\}} \frac{|\mathcal{S}|!(|\mathcal{M}| - |\mathcal{S}| - 1)!}{|\mathcal{M}|!} [V(\mathcal{S} \cup \{z_{m_i}\}) - V(\mathcal{S})]. \tag{9}$$

We use the largest predicted class probability $p_{\hat{y}}$ provided by f_{mm} as the model outcome V, where $\hat{y}$ is the pseudo class. For the most common two-modality case, the Shapley values of modality features can be computed as follows ($\emptyset$ represents a zero vector):

$$\begin{aligned} \phi(z_{m_1}) &= \frac{1}{2}[V(z_{m_1}, z_{m_2}) - V(\emptyset, z_{m_2}) + V(z_{m_1}, \emptyset) - V(\emptyset, \emptyset)], \\ \phi(z_{m_2}) &= \frac{1}{2}[V(z_{m_1}, z_{m_2}) - V(z_{m_1}, \emptyset) + V(\emptyset, z_{m_2}) - V(\emptyset, \emptyset)]. \end{aligned} \tag{10}$$

The Shapley value could be positive, negative or zero. While the sign indicates in which direction of each modality contributes, our primary interest lies in the extend of its contribution. Hence, we define modality contribution as follows:

$$\Phi_{m_i} = \frac{|\phi(z_{m_i})|}{\sum_{i=1}^{M} |\phi(z_{m_i})|}. \tag{11}$$

3.5 Proposed Method

Following the proposed guidelines, we redesign the BADGE for multimodal classification scenarios with two modalities, m_1 and m_2, to achieve more balanced data selection. The i^{th} row of gradient embedding in Eq. 3 could be derived as concatenation of two unimodal gradient embeddings:

$$g_i = (f_i - 1_{\hat{y}_{mm}=i}) z_{m_1} \oplus (f_i - 1_{\hat{y}_{mm}=i}) z_{m_2}. \tag{12}$$

We then design two weights w_{m_1} and w_{m_2}, and scale each unimodal gradient embedding by them respectively:

$$\begin{aligned} w_{m_1} &= \begin{cases} 1 & \text{if } \Phi_{m_1} \geq \Phi_{m_2} \\ 1 - \rho & \text{if } \Phi_{m_2} > \Phi_{m_1} \end{cases} \\ w_{m_2} &= \begin{cases} 1 - \rho & \text{if } \Phi_{m_1} \geq \Phi_{m_2} \\ 1 & \text{if } \Phi_{m_2} > \Phi_{m_1} \end{cases} \end{aligned} \tag{13}$$

$$g'_i = w_{m_1}(f_i - 1_{\hat{y}_{mm}=i}) z_{m_1} \oplus w_{m_2}(f_i - 1_{\hat{y}_{mm}=i}) z_{m_2}. \tag{14}$$

Here, $\rho = |\Phi_{m_1} - \Phi_{m_2}|$ is the difference between contributions of two modalities. Note that the gradient embedding of larger l2 norm will be selected more easily by K-Means++ algorithm [3]. Therefore, by multiplying with these weights, the magnitude of gradient embedding will be suppressed more if their unimodal contributions are more unbalanced. It aligns with our Guideline 1 where we want to punish the samples with unbalanced contributions.

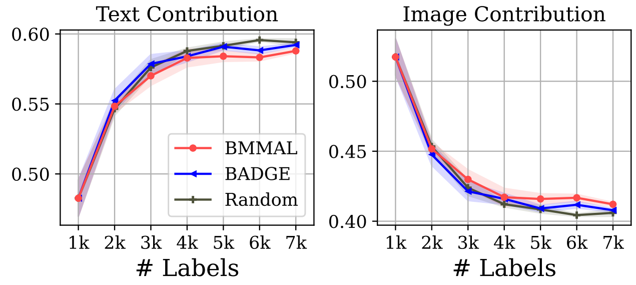


Figure 2: Modality contribution Φ across different AL iterations on the Food101 test set.

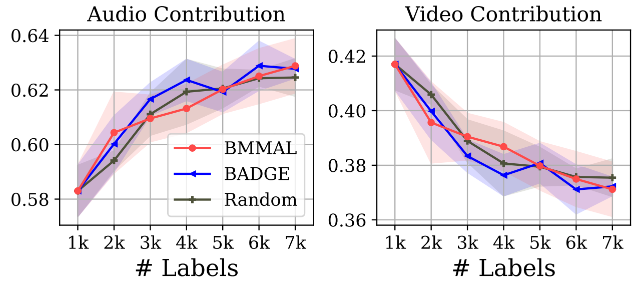


Figure 3: Modality contribution Φ across different AL iterations on the KineticsSound test set.

Moreover, we observe that the average ρ of the subset in which the weaker modality dominates is smaller than that of the subset where the stronger modality dominates. See **Figure 6** and our discussion in **Sec 4.5**. If m_1 is the weaker modality regarding the entire dataset, then we will have $\frac{1}{|X_1|}\sum_{x\in X_1}\rho(x) < \frac{1}{|X_2|}\sum_{x\in X_2}\rho(x)$ for two subsets X_1 and X_2 dominated by m_1 and m_2 respectively. It means that the subset where the stronger modality dominates will be suppressed more, and it follows our Guideline 2 to punish the stronger modality on the dataset-level.

Finally, the Guideline 3 is also adhered to. For each sample, the modality with a higher contribution to the model outcome is always assigned a greater weight, resulting in a higher magnitude of unimodal gradient embedding. This ensures that the contribution to data selection is proportional to the contribution to the model outcome and model optimization if selected.

In the end, we perform K-Means++ over the scaled gradient embedding to select candidates for labeling. As a result, our BMMAL strategy could achieve more balanced active learning on multimodal classification than BADGE. It could prevent biased selection towards either the stronger or weaker modalities, thus benefiting multimodal learning.

4 EXPERIMENT

4.1 Dataset

Food101 [42] is a multi-class food recipe dataset with 101 kinds of food. Each recipe consists of a food image and textual recipe description. The dataset consists of 45,719 samples for training and 15,294 samples for testing.

KineticsSound [1] is a sub-dataset containing 31 action classes selected from Kinetics-400 [20]. These action classes are considered to be correlated to both visual and auditory content. This dataset contains 14,739 clips for training and 2,594 clips for testing.

VGGSound [8] is a large-scale video dataset with 309 classes. Each video clip is 10-second and captures the object making the sound. We are only able to download 180,911 clips for training and 14,843 clips for testing due to the unavailability of YouTube videos.

4.2 Baseline

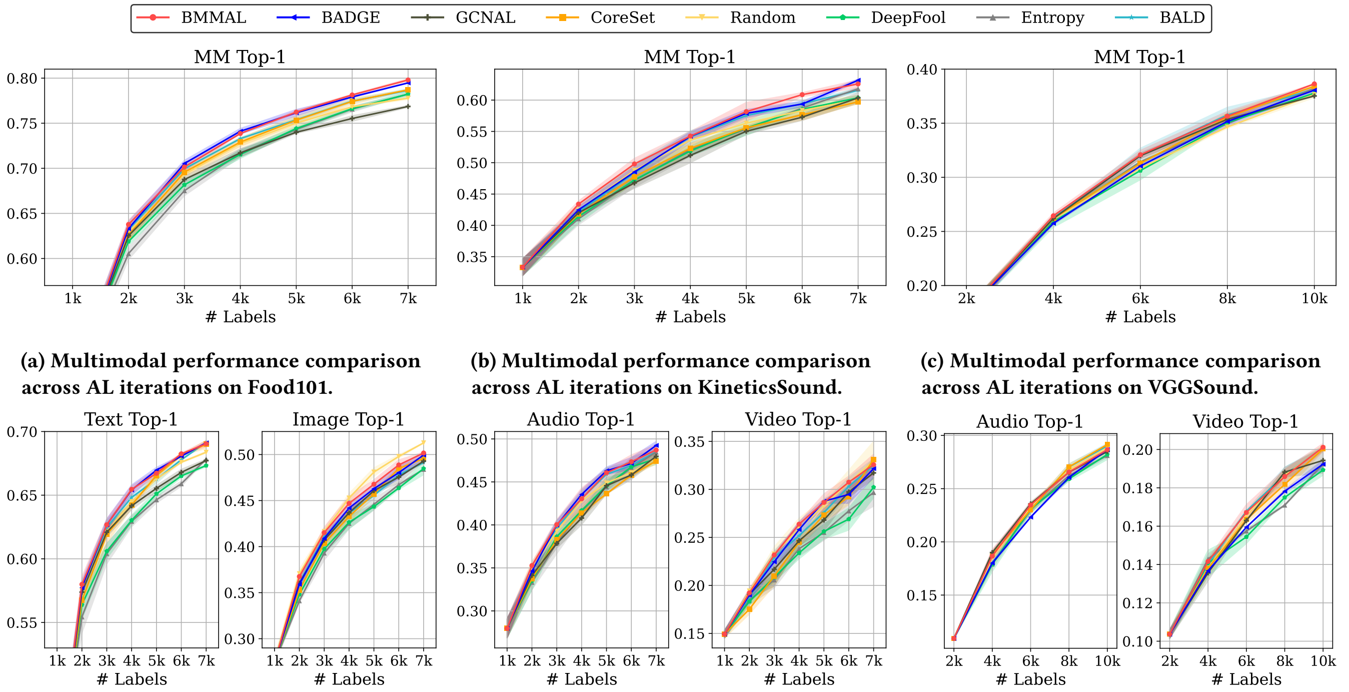
We consider seven existing active learning strategies as baselines. **Random** selects the data samples randomly from the unlabeled data pool. **Entropy** [32] selects data samples with the highest entropy of multimodal classification probabilities. **CoreSet** [31] filters out a subset of unlabeled data with representative multimodal features via K-center greedy algorithm. **BADGE** [3] is a hybrid method that selects diverse data samples by K-means++ sampler over their gradient embedding of multimodal classifier. **BALD** [13] is a Bayesian method to evaluate the mutual information between model predictions and model parameters. Since our model is static, we run five rounds of model forwarding with enabled dropout to obtain the entropy of model parameters. **DeepFool** [10] adopts an adversarial-like approach that adds small perturbations over multimodal features and selects data whose predictions are flipped. **GCNAL** [7] learns an extra graph convolution network to distinguish labelled and unlabelled samples and selects unlabelled samples that are sufficiently different from labelled ones.

4.3 Experiment Setting

Image-text Classification: For the Food101 dataset, we adopt ResNet-101 pre-trained on ImageNet as the image backbone and pre-trained Bert-base model [9] as the text backbone. All unimodal and multimodal classifiers are single FC layers. We use AdamW [22] as the optimizer and train the model for 15 epochs in each AL round and adopt random crop, random horizontal flip and random grey scale for image augmentation.

Video Classification: For VGGSound and KineticsSound, we utilize ResNet2P1D-18 [37] as visual backbone. The difference is that it is pre-trained on Kinetics-400 for VGGSound, while it is randomly initialized for KineticsSound. We use the randomly initialized ResNet-18 as an auditory backbone whose input channel is modified from 3 to 1. The video is uniformly sampled into 10 frames at the rate of one frame per second. The audio clip is transformed into a spectrogram with a window length of 512 and an overlap length of 353. For video augmentation, we randomly sample 5 frames out of 10 frames and apply image augmentation techniques on each frame. For audio augmentation, we randomly extract a 5-second audio fragment from the whole audio clip. We use Adam as optimizer and train the model for 45 epochs in each round.

The experiment is repeated 5 times for image-text classification and 3 times for video classification to remove the randomness of the initial querying. For multimodal fusion, we apply concatenation which is a widely used multimodal fusion mechanism on all tasks.



(a) Multimodal performance comparison across AL iterations on Food101.

(b) Multimodal performance comparison across AL iterations on KineticsSound.

(c) Multimodal performance comparison across AL iterations on VGGSound.

(d) Unimodal performance comparison across AL iterations on Food101.

(e) Unimodal performance comparison across AL iterations on KineticsSound.

(f) Unimodal performance comparison across AL iterations on VGGSound.

Figure 4: Performance comparison between proposed method and other conventional AL strategies with concatenation fusion method. The metric selected is top-1 accuracy (Top-1) on mulitmodal and unimodal classification.

In addition, we implement summation and NL-gate [41] that is similar to multi-head attention [38] in further experiments.

4.4 AL Performance

A fair and good AL strategy ought to select important multimodal data that could contribute to multimodal tasks and, simultaneously, pay fair attention to weaker modalities and strong modalities to prevent the trained multimodal network from degenerating into only a good unimodal network. We run conventional active learning strategies along with our proposed method BMMAL on several multimodal datasets, and compare their multimodal and unimodal classification accuracy.

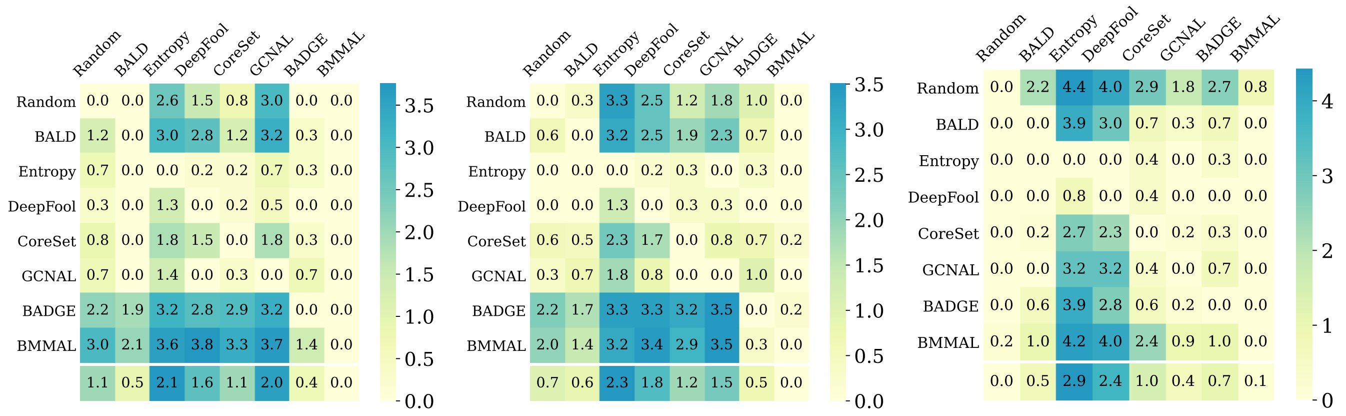
We firstly draw the trend of modality contributions to the predicted probability over the ground truth class on test dataset across different active learning iterations in **Figure 2** and **Figure 3**. As shown in the figures, the textual modal contributes more than the imagery model on the Food101 after second iteration, and the auditory modal contributes more than the visual modal on the KineticsSound. More importantly, the difference between two unimodal contributions of BMMAL is overall smaller than both BADGE and Random. It means that two modalities contribute more equally in the models trained by the data selected by BMMAL.

The performance comparison of each AL iteration on the Food101 dataset is shown in **Figure 10a** and **10c**. Note that textual modality is the stronger modality since iteration 2. Our method outperforms all baselines except BADGE in multimodal classification. In text classification, BMMAL, BADGE and CoreSet achieve good performance. In image classification, our method is superior to most of the baselines except Random. From the above comparison, we can tell that BADGE and CoreSet mainly focus on selecting valuable samples over the stronger text modality and ignore the weaker image modality. Although Random uniformly selects multimodal data without any weighting in image classification, it is considered unfair concerning the text modality.

The performance comparison of each AL iteration on the KineticsSound dataset is shown in **Figure 10b** and **10d**. Note that auditory modality is the stronger modality. Our method outperforms all baselines in multimodal classification. BADGE performs the best on audio classification on many iterations, However, its performance declines on video classification indicating that biased data selection might negatively affect multimodal classification. It shows that BADGE tends to assign more importance to audio modality during data selection and such behavior might negatively affect multimodal joint training.

The performance comparison of each AL iteration on the VGGSound dataset is shown in **Figure 4c** and **4f**. Note that auditory modality is the stronger modality. Our method outperforms BADGE in not only multimodal classification but also in two unimodal classification by an obvious margin.

Findings. Our first finding is that AL methods such as BADGE and BALD which win at classification of the stronger modality could stand a good chance of failing at classification of the weak



(a) Pairwise comparison on multimodal classification.

(b) Pairwise comparison on unimodal classification of stronger modalities.

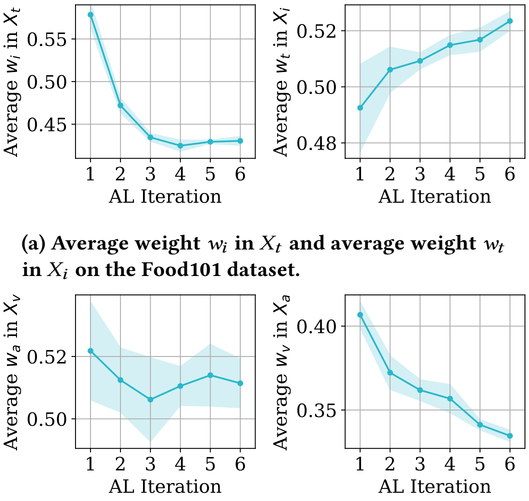
(c) Pairwise comparison on unimodal classification of weaker modalities.

Figure 5: Pairwise comparison of all active learning strategies. Each element in the matrix $P_{i,j}$ represents the number of times strategy i outperforms strategy j. A strategy is considered better if its row-wise value is larger, indicating that it beats other strategies more often. On the other hand, a strategy is better if its column-wise value is smaller, meaning it is rarely beaten by other strategies. The maximum value of each cell is 5, which is the total number of experimental settings. The bottom row displays the column-wise average values (lower is better).

modality. This may be due to biased data selection towards the stronger modality, and it is undesirable for balanced multimodal learning. Our second finding is that Random and CoreSet could perform better in the weaker modality, whereas they are inferior in multimodal classification because random selection treats every sample with absolute fairness and CoreSet focuses too much on the weak modality which are both unfair concerning the stronger modality. Finally, our method achieves a fairer multimodal data selection with a better trade-off between weak and strong modalities.

4.5 Ablation Study

Pairwise Comparison. We illustrate the results across various experimental settings in matrix P in **Figure 5** [3]. We compute the t-score for each repeated experiment and use the two-sided t-test to compare the performance of paired strategies on the test set with a 0.9 confidence interval. If strategy i significantly outperforms strategy j, we add $1/L$ to $P_{i,j}$, where L is the total number of iterations for a single experiment setting. The maximum cell value equals the total number of experiment settings. $P_{i,j}$ indicates the number of times strategy i significantly outperforms strategy j. We compute the matrix for both multimodal and unimodal classification for stronger (text for Food101, audio for KineticsSound and VGGSound) and weaker modalities (image for Food101, video for KineticsSound and VGGSound). The three matrices demonstrate that our proposed method outperforms most baselines across settings. Specifically, BMMAL surpasses BADGE in multimodal classification and unimodal classification on weaker modalities, while performing comparably with BADGE in unimodal classification on stronger modalities. This suggests that the performance improvement of BMMAL in multimodal classification mainly stems from enhancing weaker modalities while maintaining stable performance in stronger modalities.



(a) Average weight w_i in X_t and average weight w_t in X_i on the Food101 dataset.

(b) Average weight w_a in X_v and average weight w_v in X_a on the KineticsSound dataset.

Figure 6: Average weight for the weaker modality in a sub-dataset dominated by the other stronger modality.

Dominance Degree. As described in **Eq. 6**, we divide the entire unlabeled dataset into multiple sub-datasets in which modality m_i contributes the most. The Food101 dataset is divided into X_t and X_i dominated by text and image modality, respectively. In **Figure 6a**, the average weight values of the weaker modality are showed. As shown before in **Figure 2**, text modality is the stronger one starting from the second iteration. The average value of w_i in X_t accordingly becomes less than that of w_t in X_i from the second

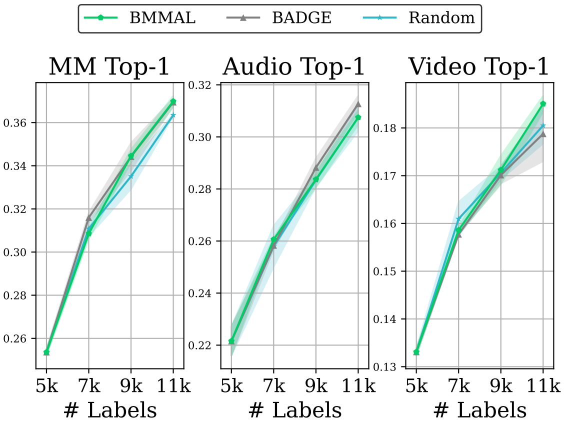


Figure 7: Multimodal and unimodal classification performance comparison with NL-gate fusion method on the VGGSound dataset.

iteration, meaning that the average difference value ρ between two unimodal contributions in X_t is larger than in X_i. The KineticsSound dataset is divided into X_v and X_a dominated by video and audio modality, respectively. In **Figure 6b**, the average weight values of the weaker modality are showed. Similarly, the average difference value ρ between two unimodal contributions in X_a is larger than in X_v. Consequently, on the dataset-level, the sub-dataset dominated by the weaker modality receives less punishment compared to the sub-dataset dominated by the stronger modality.

Different Fusion Mechanisms. We perform experiment by changing the fusion method from concatenation into summation on Food101 and KineticsSound, while keeping other settings unchanged. We include the performance comparison in the pairwise comparison and present the iterative comparison in the supplementary materials. Furthermore, we change concatenation to NL-gate for mixing video and audio features on the VGGSound dataset, setting the initial budget to 5,000 and the AL budget for each round to 2,000, as NL-gate requires more data to demonstrate its efficiency in fusion. We provide the implementation details in the supplementary materials. As shown in **Figure 7**, our method achieves comparable multimodal classification performance to BADGE and becomes worse on auditory classification. However, for the weaker visual classification, our method outperforms the others, demonstrating its effectiveness in balancing weak and strong modalities.

Large-scale Active Learning. We conduct experiment on VGGSound with larger budget size of 5,000 to validate our method on large-scale active learning for multimodal video classification. The results are averaged and shown in **Table 1**. On video classification, the performance of BADGE degrades and becomes worse than random selection, while our method achieves improvement over BADGE and random selection. On audio classification, BADGE and our method are comparable and are both better than random selection. As a result, our method performs better than BADGE and can save around 5k labels compared with random selection if target multimodal classification top-1 accuracy is set to 0.435.

	#Labels	5k	10k	15k	20k	25k
MM-Top-1	Random	0.261	0.340	0.387	0.418	0.435
	BADGE	0.261	**0.355**	0.406	0.433	0.451
	BMMAL	0.261	0.352	**0.407**	**0.437**	**0.458**
Audio-Top-1	Random	0.189	0.251	0.295	0.318	0.334
	BADGE	0.189	**0.262**	0.307	0.332	0.345
	BMMAL	0.189	0.261	**0.308**	**0.333**	**0.350**
Video-Top-1	Random	0.145	0.178	0.203	0.220	0.229
	BADGE	0.145	**0.184**	0.206	0.218	0.225
	BMMAL	0.145	0.180	**0.208**	**0.222**	**0.231**

Table 1: AL performance on VGGSound dataset with budget size of 5,000. The best results are highlight in bold.

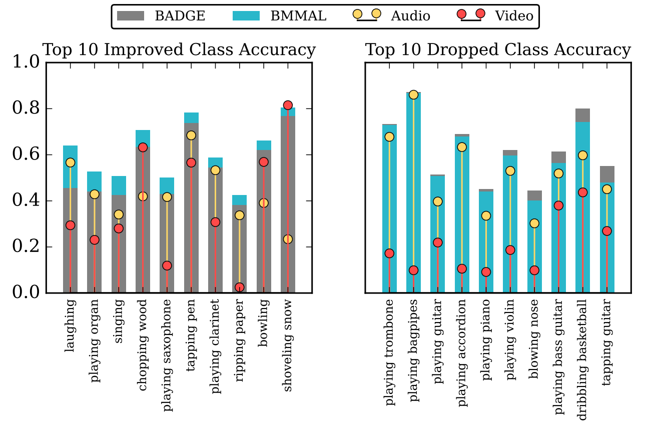


Figure 8: Top 10 improved and dropped classes based on the improvement of BMMAL to BADGE on multimodal classification accuracy on KineticsSound with 5K labeled samples. Bars represent multimodal classification accuracy. Stems represent unimodal classification accuracy.

Classwise Performance Comparison. We show the classwise performance comparison on the KineticsSound dataset. As shown in **Figure 8**, the gain is more significant than the drop. Moreover, improved classes such as 'chopping wood', 'bowling' and 'shoveling snow' carry more visual information, and dropped classes are mostly dominated by the auditory modality. Note that KineticsSound is a dataset where audio contributes more than vision, which means that BMMAL avoids biased selection over auditory modality and focuses more on the weaker visual modality.

5 DISCUSSION

In this paper, we evaluate how existing active learning strategies perform on multimodal classification. Our empirical studies show that they might treat different modalities unfairly, and it could lead to performance degradation for multimodal learning. We propose BMMAL to mitigate this unfairness by separately scaling unimodal gradient embeddings, which avoids mixing all unimodal information and well retain characteristics of each modality. The method performs well on multiple datasets and can be potentially applied on large-scale multimodal active learning.

REFERENCES

[1] Relja Arandjelovic and Andrew Zisserman. 2017. Look, Listen and Learn. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017.* IEEE Computer Society, 609–617. https://doi.org/10.1109/ICCV.2017.73

[2] David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007.* SIAM, 1027–1035. http://dl.acm.org/citation.cfm?id=1283383.1283494

[3] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net. https://openreview.net/forum?id=ryghZJBKPS

[4] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[5] William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. 2018. The Power of Ensembles for Active Learning in Image Classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018.* Computer Vision Foundation / IEEE Computer Society, 9368–9377. https://doi.org/10.1109/CVPR.2018.00976

[6] Erdem Biyik, Kenneth Wang, Nima Anari, and Dorsa Sadigh. 2019. Batch Active Learning Using Determinantal Point Processes. *CoRR* abs/1906.07975 (2019). arXiv:1906.07975 http://arxiv.org/abs/1906.07975

[7] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. 2021. Sequential Graph Convolutional Network for Active Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021.* Computer Vision Foundation / IEEE, 9583–9592. https://doi.org/10.1109/CVPR46437.2021.00946

[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A Large-Scale Audio-Visual Dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020.* IEEE, 721–725. https://doi.org/10.1109/ICASSP40776.2020.9053174

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[10] Melanie Ducoffe and Frédéric Precioso. 2018. Adversarial Active Learning for Deep Networks: a Margin Based Approach. *CoRR* abs/1802.09841 (2018). arXiv:1802.09841 http://arxiv.org/abs/1802.09841

[11] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* 9847–9857. https://doi.org/10.18653/v1/2021.emnlp-main.775

[12] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 (JMLR Workshop and Conference Proceedings, Vol. 48).* JMLR.org, 1050–1059. http://proceedings.mlr.press/v48/gal16.html

[13] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70).* PMLR, 1183–1192. http://proceedings.mlr.press/v70/gal17a.html

[14] Itai Gat, Idan Schwartz, and Alexander G. Schwing. 2021. Perceptual Score: What Data Modalities Does Your Model Perceive?. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual,* Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 21630–21643. https://proceedings.neurips.cc/paper/2021/hash/b51a15f382ac914391a58850ab343b00-Abstract.html

[15] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep Multimodal Representation Learning: A Survey. *IEEE Access* 7 (2019), 63373–63394. https://doi.org/10.1109/ACCESS.2019.2916887

[16] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. 2022. Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162).* PMLR, 8175–8195. https://proceedings.mlr.press/v162/hacohen22a.html

[17] Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020,* Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 861–877. https://doi.org/10.18653/v1/2020.emnlp-main.62

[18] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning? (Provably). In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162).* PMLR, 9226–9259. https://proceedings.mlr.press/v162/huang22e.html

[19] Leon Kaufman and Peter Rousseeuw. 1990. Finding Groups in Data. (1990).

[20] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. *CoRR* abs/1705.06950 (2017). arXiv:1705.06950 http://arxiv.org/abs/1705.06950

[21] Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. 2021. Influence Selection for Active Learning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021.* IEEE, 9254–9263. https://doi.org/10.1109/ICCV48922.2021.00914

[22] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019.* OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[23] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA,* Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[24] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active Learning by Acquiring Contrastive Examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021.* Association for Computational Linguistics, 650–663. https://doi.org/10.18653/v1/2021.emnlp-main.51

[25] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016.* IEEE Computer Society, 2574–2582. https://doi.org/10.1109/CVPR.2016.282

[26] Hieu Tat Nguyen and Arnold W. M. Smeulders. 2004. Active learning using pre-clustering. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004 (ACM International Conference Proceeding Series, Vol. 69).* ACM. https://doi.org/10.1145/1015330.1015349

[27] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Reza Haffari, Anton van den Hengel, and Javen Qinfeng Shi. 2022. Active Learning by Feature Mixing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022.* IEEE, 12227–12236. https://doi.org/10.1109/CVPR52688.2022.01192

[28] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022.* IEEE, 8228–8237. https://doi.org/10.1109/CVPR52688.2022.00806

[29] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2022. A Survey of Deep Active Learning. *ACM Comput. Surv.* 54, 9 (2022), 180:1–180:40. https://doi.org/10.1145/3472291

[30] Dan Roth and Kevin Small. 2006. Margin-Based Active Learning for Structured Output Spaces. In *Machine Learning: ECML 2006, 17th European Conference on Machine Learning, Berlin, Germany, September 18-22, 2006, Proceedings (Lecture Notes in Computer Science, Vol. 4212).* Springer, 413–424. https://doi.org/10.1007/11871842_40

[31] Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.* OpenReview.net. https://openreview.net/forum?id=H1aIuk-RW

[32] Burr Settles. 2012. *Active Learning.* Morgan & Claypool Publishers. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

[33] LS SHAPLEY. 1953. A value for n-person games. *Annals of Mathematics Studies* 28 (1953), 307–318.

[34] Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. 2020. Deep Active Learning: Unified and Principled Method for Query and Training. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy] (Proceedings of Machine Learning Research, Vol. 108).* PMLR, 1308–1318. http://proceedings.mlr.press/v108/shui20a.html

[35] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. 2019. Variational Adversarial Active Learning. In *2019 IEEE/CVF International Conference on Computer Vision,*

ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. IEEE, 5971–5980. https://doi.org/10.1109/ICCV.2019.00607

[36] Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, and Manmohan Chandraker. 2020. Active Adversarial Domain Adaptation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. IEEE, 728–737. https://doi.org/10.1109/WACV45572.2020.9093390

[37] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 6450–6459. https://doi.org/10.1109/CVPR.2018.00675

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[39] Tom J. Viering, Jesse H. Krijthe, and Marco Loog. 2019. Nuclear discrepancy for single-shot batch active learning. *Mach. Learn.* 108, 8-9 (2019), 1561–1599. https://doi.org/10.1007/s10994-019-05817-y

[40] Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What Makes Training Multi-Modal Classification Networks Hard?. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 12692–12702. https://doi.org/10.1109/CVPR42600.2020.01271

[41] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-Local Neural Networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 7794–7803. https://doi.org/10.1109/CVPR.2018.00813

[42] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. 2015. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops 2015, Turin, Italy, June 29 - July 3, 2015*. IEEE Computer Society, 1–6. https://doi.org/10.1109/ICMEW.2015.7169757

[43] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. 2022. Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 24043–24055. https://proceedings.mlr.press/v162/wu22d.html

[44] Michelle Yuan, Hsuan-Tien Lin, and Jordan L. Boyd-Graber. 2020. Cold-start Active Learning through Self-supervised Language Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics, 7935–7948. https://doi.org/10.18653/v1/2020.emnlp-main.637

[45] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. 2021. Multiple Instance Active Learning for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 5330–5339. https://doi.org/10.1109/CVPR46437.2021.00529

A COMPUTATIONAL COMPLEXITY

Computing the Shapley values of each unimodal feature requires to perform inference 2^M times in total, where M is the number of modalities. In our two-modality learning case, we need to perform inference four times with different combination of unimodal features to obtain the Shapley values, which is acceptable. Then, given the computed gradient embedding of N unlabeled samples, the sampling time complexity of BMMAL is $O(NBDK)$, where B is the query budget of each AL round, D is the size of weight matrix of the last linear classifier and K is the number of classes.

B IMPLEMENTATION OF NL-GATE

NL-gate [41] is a mid-fusion mechanism that behaves similar to multi-head attention. We implement it in the video classification task, where Resnet-18 is utilized as the audio backbone and Resnet2P1D-18 is utilized as the video backbone. Note that both Resnet-18 and Resnet2P1D-18 have four blocks. We extract the middle 2D audio features from the third block of Resnet-18 and the middle 3D video features from the third block of Resnet2P1D-18 as inputs to the NL-gate.

We show the implementation of NL-gate in **Figure 9**. The 3D video feature is average pooled over the spatial channels into a 1D video feature. It is then tiled over the frequency channel into a 2D video feature that has the same size as the 2D audio feature. The concatenation of the 2D video feature and the 2D audio feature is used as key and value in NL-gate. The original 3D video feature is used as query in NL-gate. After audio and video features are mixed, they will be processed with a random initialized module with the same layout as the fourth Block of Resnet2P1D-18 to produce the final feature. To compute the marginal unimodal contribution, we choose to compute the Shapley values of the features generated by the last shared convolution layers ($Conv_V$ and $Conv_A$) before the NL-gate fusion module.

C SPLIT THE LARGE UNLABELED DATA POOL

In large-scale AL experiments, the gradient embedding produced by all unlabeled data samples could be too large to be stored in the memory. To address this issue, we split the unlabeled data pool into S smaller pools to save memory space, where S is the split size. After splitting, we query $\frac{N}{S}$ unlabeled samples from each smaller pool and aggregate them to form the final query set. The space complexity of BMMAL is correspondingly reduced by S times. Moreover, the sampling time complexity becomes $O(\frac{N}{S}\frac{B}{S}DKS) = O(\frac{1}{S}NBDK)$, which is also reduced by S times compared with original time complexity. We use split size of eight in the large-scale AL experiment with the VGGSound-full dataset. Although splitting might affect the AL performance, we observe that both BMMAL and BADGE still perform better than random data selection. It indicates that splitting the unlabeled data pool is acceptable in large-scale AL.

D AL PERFORMANCE WITH SUMMATION

We visualize the performance comparison of all baselines with our proposed method in all AL rounds on Food101 and KineticsSound with fusion mechanism of summation in **Figure 10**, and the unimodal contribution among BMMAL, BADGE and Random in **Figure 11**. As shown in the figures, our proposed method outperforms BADGE on Food101 and achieves more balanced unimodal contribution than BADGE. While on KineticsSound, our proposed method is comparable with BADGE, and it may be due to the weak fusion ability of summation.

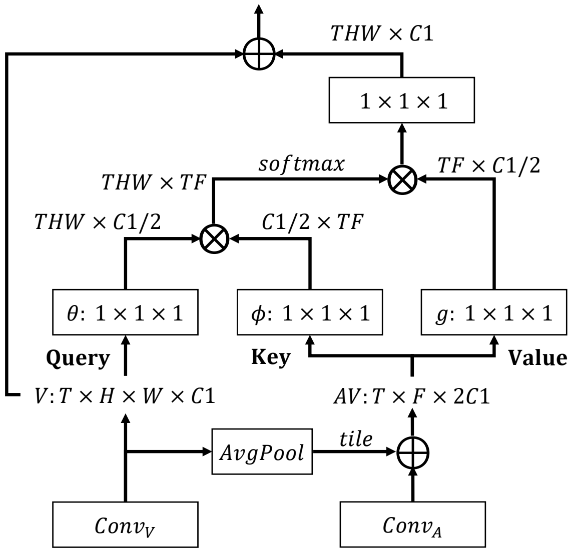
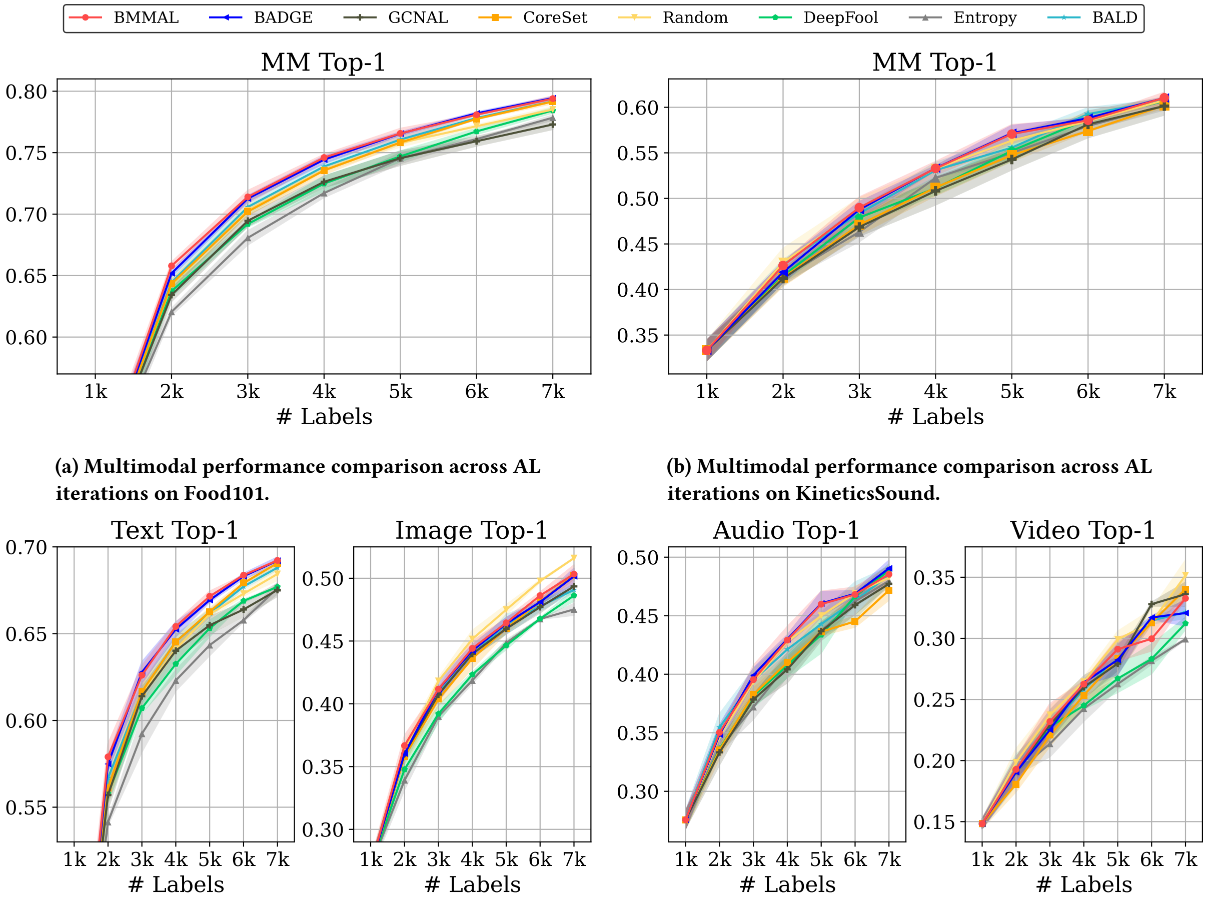


Figure 9: The implementation of NL-gate. We use the 3D video feature as query and the 2D concatenated audio and video feature as key and value.



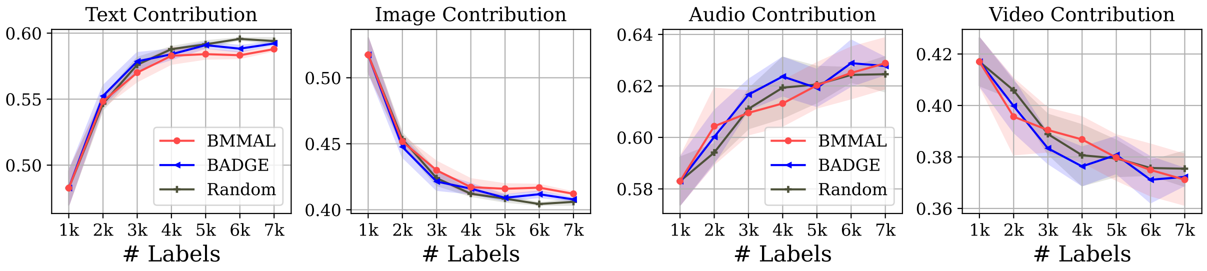
(a) Multimodal performance comparison across AL iterations on Food101.

(b) Multimodal performance comparison across AL iterations on KineticsSound.

(c) Unimodal performance comparison across AL iterations on Food101.

(d) Unimodal performance comparison across AL iterations on KineticsSound.

Figure 10: Performance comparison between proposed method and other conventional AL strategies with Summation fusion method. The metric selected is top-1 accuracy (Top-1) on mulitmodal and unimodal classification.



(a) Modality contribution Φ across different AL iterations on the Food101 test set.

(b) Modality contribution Φ across different AL iterations on the KineticsSound test set.

Figure 11: Unimodal contribution comparison among proposed method, BADGE and random selection with Summation fusion.