# AdaBoost Based Multimodal Learning

Hongchuan Yu[*,†], Boyuan Cheng[†]

*National Centre for Computer Animation, Bournemouth University*

**Abstract**

This paper focuses on multi-modal learning and introduces an AdaBoost-based approach for multi-modal learning. We address two foundation problems, (1) the difference be-tween AdaBoost with homogeneous and heterogeneous weak learners; (2) generalization metric. By addressing these research questions, this paper enhances our under-standing of AdaBoost in the context of multi-modal learning through comprehensive experiments. The experiment results show that the heterogeneous structure is a trade-off between the performances of different weak learners rather than a clear synergy. The multi-modal learning model's performance depends on how the individual weak learners are composed, and the heterogeneous structure's ad-vantage lies in harnessing the diverse strengths of individual weak learners, even though the improvement achieved is not overwhelmingly pronounced.

**Keywords**

AdaBoost, Homogeneous weak learners, Heterogeneous weak learners, Multimodal learning, Generalization metric

## 1. Introduction

Multi-modal learning refers to the process of extracting at-tributes from one or more data streams, known as modalities, that have different dimensions. The goal is to learn how to combine and project the extracted heterogeneous features into a shared representation space. In various applications, leveraging multiple modalities and sensors can provide valuable contextual information for a given task. Each modality, such as textual, visual, or auditory, has its own structure and encoding mechanisms for handling heterogeneous information harmoniously within a conceptual framework.

While the combination of different modalities or data sources to enhance performance is an ongoing research focus, it is often challenging to distinguish between noise, concepts, and conflicts among the data sources in practice.

Among boosting algorithms, AdaBoost is widely recognized as a prominent member. It converts a set of weak learners into a strong learner. Typically, AdaBoost is formulated using an additive model, where a linear combination of base learners is employed to minimize the exponential loss function. AdaBoost implementation is straightforward and comprehensible, and it is known for its resistance to overfitting [1]. Multi-modal learning aims to tap potentialities of multiple modality data, while AdaBoost is a successful example of ensemble learning. It is natural to apply AdaBoost to multi-modal learning.

However, regardless of whether ensemble learning or multi-modal learning techniques, they all encounter a common challenge: generalization. Initially, these algorithms were developed to tackle the issue of generalization, where a pre-trained model can effectively handle unseen domains. In this paper, we leverage the power of AdaBoost and introduce novel multi-modal learning methods based on it. Unlike the conventional implementation of AdaBoost that assumes homogeneous weak learners, in multi-modal learning scenarios, each modality may have its own individual learners, resulting in heterogeneous learners.

The main challenge we face in our proposed algorithm involves two aspects: (1) Assessing the performance difference of AdaBoost with homogeneous and heterogeneous classifiers, respectively; (2) Establishing a quantifiable metric for generalization, which has been lacking in existing research. Our contributions in this paper are as follows:

1. We demonstrate that AdaBoost performs equally well with homogeneous weak learners as with heterogeneous weak learners.
2. We introduce a new metric for measuring the generalization capability of the proposed algorithm. This metric allows us to assess how well the algorithm generalizes to unseen data.

By addressing these challenges and making these contributions, our paper aims to enhance the understanding and application of multi-modal learning techniques, especially in the context of AdaBoost-based approaches.

## 2. Related Work

Multi-modal learning currently addresses four key challenges as outlined in [2]. First, the challenge of representation involves effectively summarizing and combining

*ENIGMA-23, September 03–04, 2023, Rhodes, Greece*

[*] denotes corresponding author.

[†] denotes equal contribution.

hyu@bournemouth.ac.uk (H. Yu);
s5526309@bournemouth.ac.uk (B. Cheng)

data from diverse modalities while accounting for heterogeneity, noise levels, and missing data. Deep networks have been employed to represent visual, acoustic, and textual data, with recent efforts focusing on fine-tuning these representations for specific tasks [3].

The second challenge is translation, which aims to generate an entity in one modality based on information from a different modality. An example of this is video description generation. Previous work by [4] proposed a system that describes human behavior in videos using detected head and hand positions combined with rule-based natural language generation. Evaluating multimodal translation methods is challenging, as there are often multiple correct answers and subjective judgments involved.

The third challenge is alignment, which involves finding relationships and correspondences between subcomponents of instances across multiple modalities. For instance, aligning a movie with its corresponding script or book chapters. Dynamic Time Warping and Canonical Cor-relation Analysis are commonly used for multimodal data alignment, and [5] introduced the deep canonical time warping approach, which generalizes deep CCA and DTW.

The fourth challenge is fusion, which aims to integrate in-formation from multiple modalities to improve the robustness of predictions. In the context of continuous multi-modal emotion recognition, [6] demonstrated the advantages of using LSTM models over graphical models and SVMs.

It is worth noting that these challenges and approaches are part of a broader survey on multi-modal learning, and further details can be found in [2].

The main objective of generation is to develop a model from one or multiple distinct yet related domains (i.e., di-verse training datasets) that can generalize effectively on unseen testing domains. Ensemble learning leverages the connections between multiple source domains by employing specific model architecture designs and training strategies to enhance generalization. The underlying assumption is that any sample can be seen as a combination of multiple source domains, resulting in an overall prediction that combines the outputs of various domain-specific models. [7] introduced domain-specific layers cor-responding to different source domains and learned the linear aggregation of these layers to represent a test sample. Similarly, [8] proposed Domain Adaptive Ensemble Learning (DAEL), which comprises a CNN feature extractor shared across domains and multiple domain-specific classifier heads. Each classifier acts as an expert for its own domain but a non-expert for others. The objective of DAEL is to collaboratively train these experts by teaching the non-experts with the expert knowledge, encouraging the ensemble to effectively handle data from previously unseen domains. This approach fosters domain adaptation and allows the model to generalize well across different domains.

In this paper, we aim to tackle these challenges by employing the AdaBoost algorithm since it allows more diversity of models and features.

# 3. Methodology

## 3.1. Problem Description

Consider two modalities generated from the sample set $S$, $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$, where n denotes the index of samples, $x_n \in R^a$ and $y_n \in R^b$ with a and b dimensions respectively. Given the ground truth labels $Z = \{z_1, \ldots, z_n\}$, where $z_n \in \{0, 1\}$ or multiple classes, we aim to train a multi-modal learning model to map both $X$ and $Y$ into the same categorical set of $Z$.

## 3.2. AdaBoost with heterogeneous weak learners

In terms of the additive model in [9], the weak classifier $h(t)$ minimizes the classification error under the distribution $D_t$ over the training data. Its classification error rate should be less than 0.5 for the $D_t$. It can be noted that if any $h(t)$ could satisfy this requirement, the resulting final strong classifier $H(t)$ still satisfy the error bound in [10]. Moreover, with the assumption of the error rate (i.e., loss function) is convex, it's possible to prove that AdaBoosts outperform individual learners according to Jensen's inequality. Note that it is true regardless of where the individual learners come from. These imply that whether homogeneous or heterogeneous weak classifiers do not influence the performance of AdaBoost. Our numerical experiments in Section 4.1 verify this assertion.

## 3.3. Multi-modal learning based on AdaBoost

The basic idea is that the different modalities $X$ and $Y$ are bundled with weak learners together and are viewed as heterogeneous learners. The sample set $S$ and the label set $Z$ are employed to the training dataset. The proposed multi-modal learning model is implemented based on AdaBoost as shown in Figure 1.

The weak classifiers may be either homogeneous or heterogeneous, which is suited to the scenario that modality data has their individual classifiers. Moreover, each sample in $S$ may be a collection of multi-class data. Under the AdaBoost scheme, we update the rule of the sample distribution $D_t$ over the $S$.

Note that different modalities may share the same classifier and their combinations are still regarded as inde-

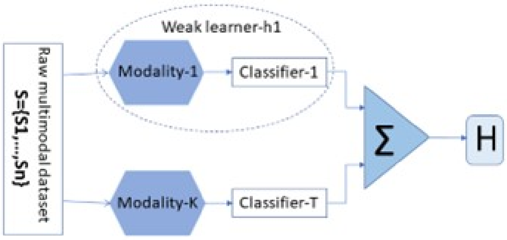

**Figure 1:** Illustration of multi-modal learning model.

pendent heterogeneous learners. This can maximally generalize weak learners.

# 4. Generalization metric

Indeed, the generalization error of an AdaBoost algorithm is influenced by the diversity of its individual learners. This relationship is elucidated through the error-ambiguity decomposition method introduced in [11]. It is sensible to consider the diversity of a classifier as a representation of its generation capabilities. In other words, a more robustly generated classifier exhibits greater diversity, leading to improved performance metrics, such as a lower error rate.

However, a significant challenge in this context is the absence of a well-defined diversity measurement. While it is intuitive to link diversity to better performance, there is currently a lack of standardized and quantifiable metrics to precisely evaluate and compare the diversity of classifiers. Addressing this gap could potentially enhance our understanding of how diversity impacts generalization and lead to further improvements in ensemble learning algorithms like AdaBoost.

To measure AdaBoost diversity, we apply the Kappa statistic to measuring the pairwise similarity/dissimilarity between two learners, and then average all the pairwise measurements for the overall diversity. This can be simply described in a binary classification application. We have the following contingency table for two learners $h_i$ and $h_j$, where $a+b+c+d=n$ are non-negative variables showing the numbers of examples satisfying the conditions specified by the corresponding rows and columns.

| | $h_j = 0$ | $h_j = 1$ |
|---|---|---|
| $h_i = 0$ | $a$ | $b$ |
| $h_i = 1$ | $c$ | $d$ |

Kappa statistic:

$$k = \frac{p_1 - p_2}{1 - p_2} \tag{1}$$

where,

$$p_1 = \frac{a+d}{n} \tag{2}$$

$$p_2 = \frac{(a+b)(a+c)+(c+d)(b+d)}{n^2} \tag{3}$$

The Kappa statistic is computed using the weak learner's performance data (e.g. error rate) when it is involved in AdaBoost.

# 5. Experiments and analysis

## 5.1. Data Collection

### 5.1.1. Synthetic dataset

To evaluate the performance of AdaBoost, we conducted experiments using homogeneous weak learners and heterogeneous weak learners respectively. For this purpose, we generated a synthetic dataset consisting of 1000 samples, 10 features, and 2 classes using the Gaussian function with zero-mean and variance of 1.

### 5.1.2. CIFAR-10 dataset

The CIFAR-10 dataset [12] is a widely-used benchmark for image classification. It comprises 60,000 color images of size 32x32, distributed across 10 classes with 6,000 images per class. The dataset exhibits diverse and relatively low-resolution images. To simulate a multi-modal learning scenario, we extract three types of feature representations: color-based features (HSV histogram), shape-based features (Histogram of Oriented Gradient), and texture-based features (Gabor filter). How-ever, considering that the original AdaBoost algorithm was designed for binary classification, we selected two classes from the CIFAR-10 dataset for experiments.

### 5.1.3. Million Song Dataset

We also design experiments based on AdaBoost for music emotion recognition with Million Song Dataset [13], which refers to recognizing and classifying emotions in music using multiple modalities (such as audio, lyrics). We chose two different emotion categories as labels based on the quadrant distribution in Russel's emotion model, i.e., positive and negative [14]. We extract the lyrics features from the MusiXmatch dataset derived from Million Song Dataset and a series of emotionally representative acoustic features (i.e., Tempo, Beats, Harmonic, Percussive, Root Mean Square, Zero Crossing Rate, Onset Frames, Chroma short-time Fourier transform, Chroma Energy Normalized, Chroma Constant-Q chromagram, Mel-spectrogram, MFCC, Poly, Tonnetz, Spectral bandwidth, Spectral roll-off, Spectral contrast, Spectral centroid) by the librosa python library [15].

**Table 1**
Performance of AdaBoost with homogeneous weak learners on the synthetic dataset

| Weak Learner | Accuracy (%) |
|---|---|
| DT | 91.09 |
| NB | 90.37 |
| Per | 84.55 |

**Table 2**
Performance of AdaBoost with heterogeneous weak learners on the synthetic dataset

| Weak Learner Combination | Accuracy (%) |
|---|---|
| DT + NB | 92.31 |
| DT + Per | 90.96 |
| NB + Per | 90.49 |
| DT + NB + Per | 92.38 |

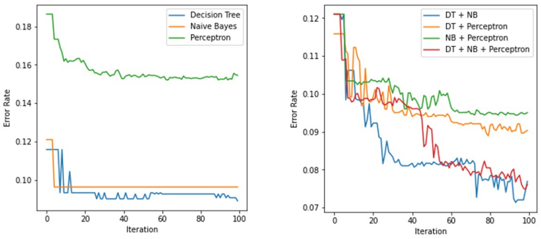

**Figure 2:** Performance of AdaBoost with homogeneous (left) and heterogeneous (right) weak learners on synthetic dataset.

## 5.2. Results and analysis

### 5.2.1. Experiment 1: Comparison of AdaBoost with homogeneous and heterogeneous weak learners

We firstly performed AdaBoost on the synthetic dataset and apply three weak learners to homogeneous and heterogeneous scenarios, i.e., Decision Tree (DT), Naive Bayes (NB), Perceptron (Per). The results are shown in Table 1,2 and Figure 2. It can be noted that whether homogeneous or heterogeneous weak learners do not affect the AdaBoost performance.

We further performed AdaBoost on the CIFA-10 with homogeneous and heterogeneous weak learners respectively. The results are shown in Tables 3,4 and Figure 3. It can be noted that (1) the AdaBoost performance is not influenced by homogeneous or heterogeneous weak learners; (2) architecture of heterogeneous weak learners usually does not make the AdaBoost performance improved. This is reasonable since different weak learners in the heterogeneous architecture have the individual performances. This finally results in a trade-off of the performance of different weak learners rather than synergy.

**Table 3**
Performance of AdaBoost with homogeneous weak learners on CIFAR-10 dataset

| Weak Learner | Accuracy (%) |
|---|---|
| DT | 80.13 |
| NB | 71.43 |
| Per | 79.55 |

**Table 4**
Performance of AdaBoost with homogeneous weak learners on CIFAR-10 dataset

| Weak Learner Combination | Accuracy (%) |
|---|---|
| DT + NB | 74.53 |
| DT + Per | 81.31 |
| NB + Per | 79.90 |
| DT + NB + Per | 80.18 |

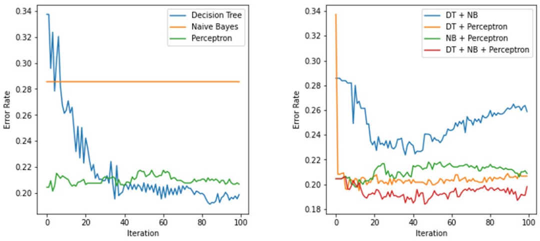

**Figure 3:** Performance of AdaBoost with homogeneous (left) and heterogeneous (right) weak learners on CIFAR-10 dataset.

**Table 5**
Performance of uni-modal learning based on AdaBoost

| Feature | Accuracy (DT) (%) | Accuracy (NB) (%) |
|---|---|---|
| HSV | 80.13 | 71.43 |
| Gabor | 84.68 | 63.24 |
| HOG | 67.48 | 65.16 |

### 5.2.2. Experiment 2: Multi-modal learning tests

We firstly per-formed AdaBoost with the homogeneous weak learners (Decision Tree, Naive Bayes) on each uni-modal feature in the CIFAR-10 and the Million Song Dataset, respectively. The results are shown in Tables 5,6 and Figure 4.

We further applied AdaBoost with homogeneous weak learners (DT, NB) to multi-modal dataset. To mock multi-modal learning, we chose 4 combinations of the features (HSV, Gabor, HOG) as multi-modal data. For the music emotion recognition, there are two kinds of real modality data available. The results are shown in Tables 7,8 and

**Table 6**
Performance of uni-modal music emotion recognition

| Feature | Accuracy (DT) (%) | Accuracy (NB) (%) |
|---|---|---|
| Audio | 65.79 | 72.56 |
| Lyrics | 61.41 | 62.27 |

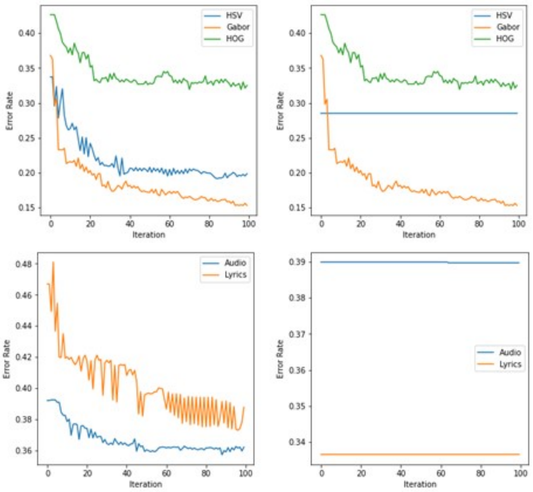

**Figure 4:** Performance of uni-modal learning on CIFAR-10 dataset (above) and uni-modal music emotion recognition (below) based on AdaBoost. The left uses DT while the right using NB.

**Table 7**
Performance of multi-modal learning based on AdaBoost on CIFAR-10 dataset

| Feature | Acc (DT) (%) | Acc (NB) (%) |
|---|---|---|
| HSV + Gabor | 87.80 | 77.29 |
| HSV + HOG | 81.81 | 74.82 |
| Gabor + HOG | 82.18 | 76.18 |
| HSV + Gabor + HOG | 87.89 | 78.82 |

**Table 8**
Performance of multi-modal music emotion recognition based on AdaBoost

| Feature | Acc (DT) (%) | Acc (NB) (%) |
|---|---|---|
| Audio + Lyrics | 66.29 | 75.11 |

Figure 5. It can be noted that multi-modal learning does not outperform uni-modal learning.

However, an exception can be noted, that is, multi-modal music emotion recognition with the Naive Bayes as the weak learner obviously decreases the error rate. Moreover, we also note that AdaBoost with the homogeneous learner of NB on the feature of lyrics has a good performance in Figure 4. This implies that the features usually have their individual classifiers. To take advantage of the features, it is better to bundle the features with their individual weak learners together as the independent weak learners.

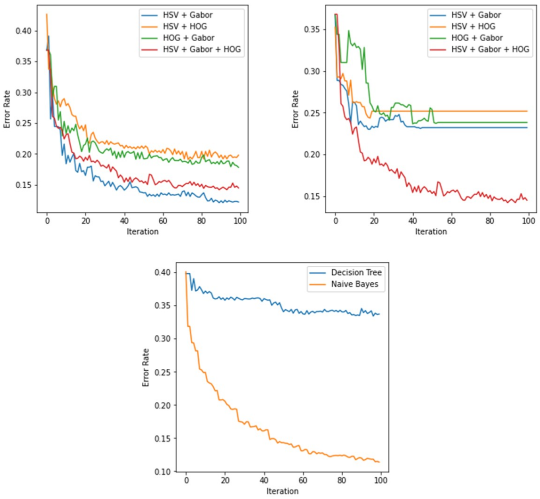

**Figure 5:** Performance of multi-modal learning on CIFAR-10 dataset (above) and multi-modal music emotion recognition (below) based on Ada-Boost. The left uses DT while the right using NB in the above row.

**Table 9**
Performance of multi-modal learning on CIFAR-10 dataset and multi-modal music emotion recognition based on AdaBoost

| Feature | Accuracy (%) |
|---|---|
| HSV + Gabor + HOG | 87.60 |
| Audio + Lyrics | 88.83 |

Therefore, on the CIFAR-10 dataset, we bundled the HSV features with the DT weak learner, the Gabor features with the DT weak learner, and the HOG features with the Stochastic Gradient Descent weak learner (SGD). In multi-modal music emotion recognition, we bundled audio features with the Decision Tree weak learner and lyrics features with the Naive Bayes weak learner. The results are shown in Table 9 and Figure 6.

It can be noted that bundling the features with their individual classifiers together as the independent weak learners can improve performance. For example, in music emotion classification, bundling Audio+DT and lyrics+NB as the weak learners has the accuracy of 88.83% in the Table 9, multi-modal learning with the Naive Bayes as the single learner has the accuracy of 75.11% in the Table 8, and the highest accuracy of uni-modal learning is of 72.56% in the Table 6. Alongside the results of music emotion classification, we can also note that the result

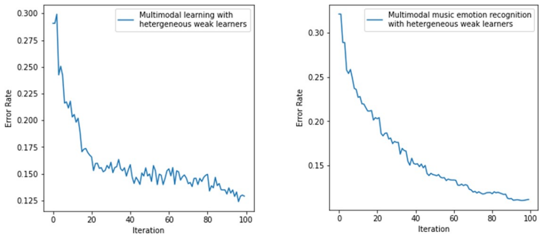

**Figure 6:** Performance of multi-modal learning on CIFAR-10 dataset (left) and multi-modal music emotion recognition (right) based on AdaBoost.

of bundling the features with their individual classifiers (i.e., HSV+DT, Gabor+DT, HOG+SGD) on the CIFAR-10 as the weak learners in Table 9 is only comparable with that of multi-modal learning with the single classifier of DT in Table 7. This is acceptable since these three combinations in Table 9 may have different performance. Experiment 1 justifies that the final result is a trade-off of the performance of different weak learners rather than a synergy.

### 5.2.3. Experiment 3: AdaBoost based MLs' diversities

In the proposed multi-modal learning model (refer to Fig. 1), the weak learner can exhibit different compositions, which can be categorized into the following types:

1) The same classifier with different features, resulting in multiple distinct weak learners.

2) The same feature with different classifiers, leading to multiple diverse weak learners.

3) Different features with their individual classifiers, yielding multiple weak learners.

To compare the structures of homogeneous and heterogeneous weak learners, each weak learner is first used in the AdaBoost homogeneous structure. Subsequently, these weak learners are incorporated into the AdaBoost heterogeneous structure. In each AdaBoost iteration, we calculate the pairwise Kappa statistics of weak learners originating from the AdaBoost and their average error rates, which are then represented in a scatter plot. Herein the origin (0,0) denotes error rate=0 and Kappa=0, which is the ideal point. Figure 7 illustrates the results for composition 2, while Figure 8 shows those for composition 1. Overall, the heterogeneous structure broadly encompasses the results obtained from the homogeneous structures. Figure 9 displays the results for composition 3. In the homogeneous structure tests, we experimented with various combinations of features and classifiers for weak learner design, selecting 2 or 3 learners with satisfactory performance for the heterogeneous structure test.

It is noteworthy that the selected weak learners, composed of features and their individual classifiers, exhibited good performance in the homogeneous structure tests. Consequently, the heterogeneous structure demonstrated improved performance compared to the results of the homogeneous tests such as error rate in Figure 9. However, the extent of improvement was not significant, suggesting that the overall outcome represents a

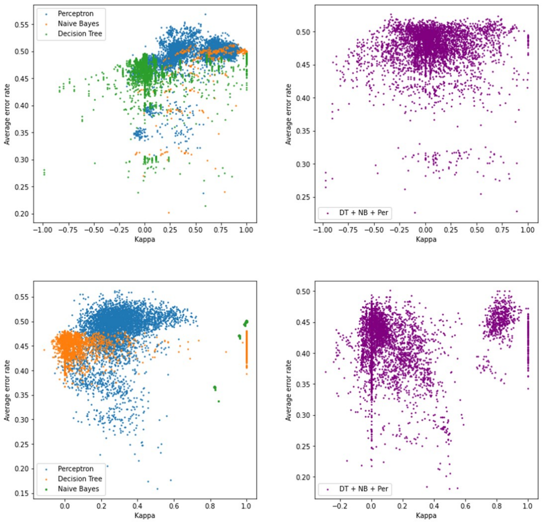

**Figure 7:** Same feature + Multiple classifiers. Synthetic dataset (above) and CIFAR-10 dataset (below). Homogeneous weak learners (left), heterogeneous weak learners (right).

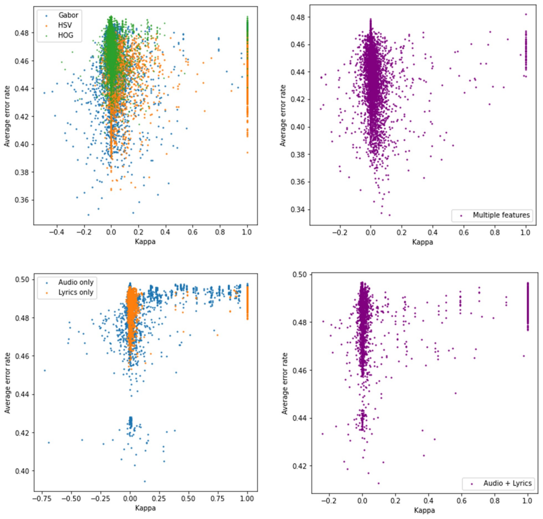

**Figure 8:** Same classifier + Multiple features. CIFAR-10 dataset (above) and multi-modal music emotion recognition (below). Homogeneous weak learners (left), heterogeneous weak learners (right).

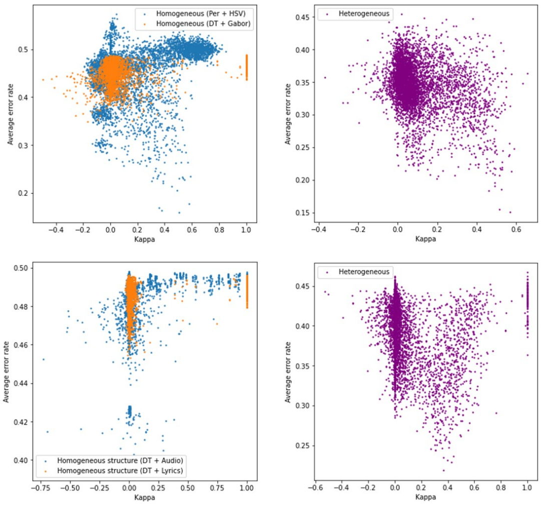

**Figure 9:** Multiple features with individual classifiers. CIFAR-10 dataset (above) and multi-modal music emotion recognition (below). Homogeneous tests (left), Heterogeneous tests (right).

trade-off between the performances of different weak learners rather than a clear synergy. The heterogeneous structure did not lead to a distinct and prominent change in performance

## 6. Conclusion

In this paper, we conducted experiments and analysis to explore AdaBoost-based multi-modal learning methods. Our findings lead to the following conclusions:

(1) The architecture of homogeneous or heterogeneous weak learners does not significantly impact the performance of AdaBoost.

(2) In the architecture of heterogeneous weak learners, each weak learner contributes individual performance, and the ensemble learning result is a trade-off among the performances of different weak learners rather than a synergistic effect.

(3) In multi-modal learning, each modality possesses its own classifiers. To fully maximize the potential of multi-modalities, it is preferable to bundle the modalities with their individual classifiers as independent weak learners for ensemble learning. However, whether homogeneous or heterogeneous architectures do not bring about distinct change.

In future research, we plan to apply AdaBoost-based multi-modal learning to address various challenges in the field, such as representation, alignment, explainability, and more. This will further demonstrate the potential and effectiveness of AdaBoost in the context of multi-modal learning.

## References

[1] Z.-H. Zhou, Large margin distribution learning, in: Artificial Neural Networks in Pattern Recognition: 6th IAPR TC 3 International Workshop, ANNPR 2014, Montreal, QC, Canada, October 6-8, 2014. Proceedings 6, Springer, 2014, pp. 1–11.

[2] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE transactions on pattern analysis and machine intelligence 41 (2018) 423–443.

[3] D. Wang, P. Cui, M. Ou, W. Zhu, Deep multimodal hashing with orthogonal regularization, in: Twenty-fourth international joint conference on artificial intelligence, 2015.

[4] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions, International Journal of Computer Vision 50 (2002) 171–184.

[5] G. Trigeorgis, M. A. Nicolaou, S. Zafeiriou, B. W. Schuller, Deep canonical time warping, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5110–5118.

[6] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, Lstm-modeling of continuous emotions in an audiovisual affect recognition framework, Image and Vision Computing 31 (2013) 153–163.

[7] M. Mancini, S. R. Bulo, B. Caputo, E. Ricci, Best sources forward: domain generalization through source-specific nets, in: 2018 25th IEEE international conference on image processing (ICIP), IEEE, 2018, pp. 1353–1357.

[8] K. Zhou, Y. Yang, Y. Qiao, T. Xiang, Domain adaptive ensemble learning, IEEE Transactions on Image Processing 30 (2021) 8008–8018.

[9] J. Friedman, T. Hastie, R. Tibshirani, Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors), The annals of statistics 28 (2000) 337–407.

[10] P. Bartlett, Y. Freund, W. S. Lee, R. E. Schapire, Boosting the margin: A new explanation for the effectiveness of voting methods, The annals of statistics 26 (1998) 1651–1686.

[11] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, Advances in neural information processing systems 7 (1994).

[12] H. Li, H. Liu, X. Ji, G. Li, L. Shi, Cifar10-dvs: an event-stream dataset for object classification, Frontiers in neuroscience 11 (2017) 309.

[13] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, P. Lamere, The million song dataset (2011).

[14] J. A. Russell, A circumplex model of affect., Journal of personality and social psychology 39 (1980) 1161.

[15] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar,

E. Battenberg, O. Nieto, librosa: Audio and music signal analysis in python, in: Proceedings of the 14th python in science conference, volume 8, 2015, pp. 18–25.