Doctoral Dissertation

Study on Efficient Multimodal Fusion Strategy with Deep Neural Network using Audio-Visual Data for Fine-Grained Recognition

Graduate School of Engineering, Iwate University
Doctoral Course, Design & Media Technology
Naranchimeg Bold

March, 2020

Abstract

In recent decade, many state-of-the-art algorithms on object recognition as well as audio recognition have achieved noticeable successes with the development of deep convolutional neural network (CNN). However, most of the studies only exploit a single type of data. For that reason, the multimodal research field brings some unique challenges for researchers to make progress in understanding the things (i.e., the world around us) by processing relate information from multiple modalities. Learning from multimodal sources offers the possibility of capturing correspondences between modalities and gaining an in-depth understanding of an object, event, or activity of interest. Furthermore, multimodal fusion is one of the original topics in multimodal research with different fusion approaches: early, late, and hybrid approaches. Multimodal fusion has a very broad range of applications, including image and sentence detection, multimodal emotion recognition, action detection, and especially audio-visual speech recognition (AVSR). The results have shown that one modality can enhance the performance of the other by providing relevant information. However, extracting robust features from different modalities and fusing them in an effective way is crucial for attaining high recognition performance. Moreover, it is practically challenging for a learning model to fuse different

modalities while learning features or to learn joint features while fusing different modalities.

On the other hand, fine-grained recognition (i.e., categorization, classification) is one of the latest and most challenging object recognition task which aims to distinguish similar object categories from another, e.g., species of birds, models of cars, breed of dogs, species of flowers, etc. Fine-grained recognition have been used for widespread applications such as analyzing biodiversity and scene understanding because it has the ability to describe the things in the world in more detail. However, compared to generic object recognition, fine-grained recognition is quite challenging due to the small difference among categories and can be easily overwhelmed by other factors, such as pose, viewpoint, or location of the object in the image. Due to the recent advances in deep learning lead to remarkable progress on fine-grained recognition. Nevertheless, most of the current state-of-the-art methods employee part/pose detection approach, which is also a challenging task to achieve good performances.

This dissertation presents a study on multimodal fusion strategies with deep neural network and audio-visual data for fine-grained recognition. Since sound also provides us important information about the world around us, the goal of this study is to enhance the performance of fine-grained recognition by exploiting the combination of both visual and audio data using CNN, which has been sparsely treated so far. For this purpose, this study aims to answer three research considerations: (1) what to fuse, i.e., what feature representations to use for audio and visual modality, (2) when to fuse, i.e., which fusion strategy performs best when fusing both modalities, (3) how to fuse, i.e., propose or utilize practically efficient methods under

different fusion strategies while learning CNN features with both modalities.

Specifically, the kernel-based fusion approach which fuses audio and visual features at the kernel level is studied. First, audio and visual CNN are trained separately by adapting the weights of the pre-trained model. After the training, deep neural features from both modalities are extracted based on the activation values of an inner layer of the trained CNN and combined by multiple kernel learning (MKL) to perform the final classification. Since generating deep neural features based on visual representations of audio recordings have proven to be very effective, spectrogram representation of audio data is used to extract CNN features. To train Audio and Visual CNN, the most suitable and large enough dataset is required. Such data were not available for audio modality, the audio dataset corresponding to the popular fine-grained image dataset has been collected with proper matching. Unlike support vector machine (SVM) based on a single kernel, MKL uses multiple kernels and learns optimal composite kernels by combining those kernels constructed from different modalities. To automatically determine the kernel weights, an l_p-norm MKL algorithm that produces non-sparse kernel combinations is employed. The experimental results indicate that proposed CNN+MKL method which utilizes the combination of audio-visual data outperforms single-modality kernel methods, some simple kernel combination methods, and the conventional early fusion method.

Furthermore, CNN-based multimodal learning models with three types of fusion strategies (early, middle, late) are proposed to settle the issues of combining training data of both modalities. The advantage of the proposed method lies in the fact that CNN is utilized not only to extract features from visual and audio data but also to combine the features across modalities. Experiments are conducted to

evaluate the multimodal learning models as well as different fusion strategies with respect to the classification accuracy on the integrated dataset. In qualitative and quantitative results, a model that utilizes the combination of both data outperforms models trained with only either type of data, and fusing them at the late stage performs better with a significant margin. It is also shown that transfer learning can significantly increase the classification performance.

Acknowledgment

I would like to express my sincere thanks and appreciation to those who were involved in my study and life in the past three years.

First of all, I would like to express my sincere thanks and appreciation to my supervisor Prof. Takuya Akashi for giving me the opportunity to do my Ph.D. at Smart Computer Laboratory (SmartCV) in a welcoming and positive environment. I have always deeply grateful for his continuous support, guidance, encouragement, motivation, patience in this study and life in Japan. I also express my deep and sincere appreciation to a former student of our lab, currently working as an assistant professor at Fukui University, Prof. Chao Zhang, who is the coauthor of my research papers, for his constructive guidance, rapid response, comments, and effort that greatly contributed to my research papers. Without his generous support, this study would not have been possible.

I am deeply indebted to Prof. Kouichi Konno, Prof. Tadahiro Fujimoto, for their valuable comments and suggestions. I also want to thank all the past and present members of SmartCV laboratory for their help, assistance, and support. Special thanks to Mrs. Hikaru Kaketa for her kindness and assistance in administration works.

My deep appreciation goes to the professors and colleagues at the National University of Mongolia, especially Prof. Lodoiravsal Choimaa, Prof. Oyun-Erdene Namsrai, for their constructive support and encouragement. I greatly appreciate the financial support received towards my Ph.D. from Mongolia-Japan Higher Engineering Education Development (MJEED) project.

Last but not least, I must express my special thanks to my mother, sister, and parents-in-law for their endless patience, encouragement even live far away. I am also very grateful to my friends Dr. Shurentsetseg Erdenebayar, Dr. Zolzaya Kherlenchimeg, for their support, help, and encouragement. In the end, I would like to express my appreciation to my husband Nyamjav and my daughters Nomin and Ninjin for giving me motivation, strength, love, energy to pursue my goal. Thank you very much.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Real-world situations involve multiple modalities and human are able to handle information consist of different information from multiple plural modalities, e.g., we see objects, hear sounds, feel the texture, smell odors, and taste flavors. Generally, modality refers to the way in which something happens or is experienced and a research problem is characterized as multimodal when it includes multiple such modalities [8].

Fueled by the recent advances in deep neural networks, dramatic progress made in multimodal research area. Most of the works in multimodal research involve a certain types of information as modality such as visual (images or videos), audio(voice, sound and music), natural language (both spoken or written), haptics/touch, and other modalities (e.g., medical images such as MRI, EGG or depth images, etc.). Multimodal research needs to be able to interpret such multimodal signals together in order to make progress in understanding the things and events around

us. Since signals from different modalities often carry complementary information about the same structures in the world, learning-based methods that combine information from multiple modalities can improve the recognition performance compared with unimodal recognition by utilizing complementary sources of information. As a consequence, such learning-based model known as multimodal learning (i.e., multimodal deep learning) have been used for tasks such as image and sentence matching [9], RGB-D object recognition [6], action detection [7] and specially speech recognition [9, 10, 11, 12], fusing different modalities. The results have shown that one modality can enhance the performance of the other by providing relevant information. Furthermore, authors of [13, 14] proposed fusion schemes for multimodal learning with considering the architectures of neural networks. However, extracting robust, discriminative features from different modalities and fusing them in an effective and efficient way is critical to achieve high recognition performance. In addition, it is practically challenging for learning-based methods to fuse different modalities while learning feature representations or to learn joint feature representations while fusing different modalities.

Fine-grained recognition is the latest computer vision task that aims to distinguish subordinate categories within a basic-level category. In contrast to general object recognition, fine-grained recognition is quite challenging due to the subtle difference among categories or instances as shown in Fig. 1.1. For example, in fine-grained bird recognition, we want to identify the species of a bird in an image, such as "the red winged blackbird", "the scissor tailed flycatcher", "the evening grosbreak", etc., shown in Fig. 1.1.b. Other examples include distinguishing different species of flowers [15, 16], dogs [17], plants [15, 16], or models of certain products

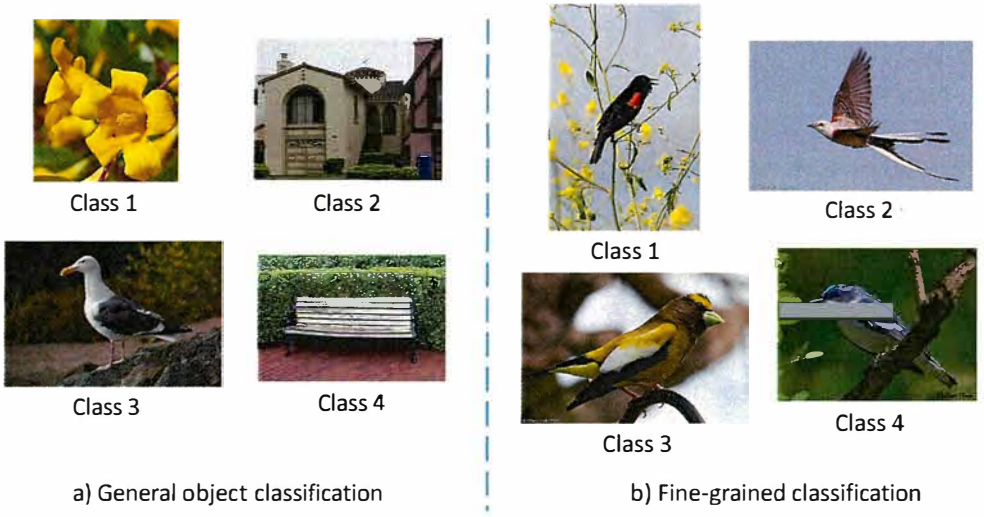


Figure 1.1: General object classification v.s. fine-grained classification. a) The general object classification usually refers to distinguishing visually different object categories such as flowers, houses, birds, etc. b) Fine-grained classification, also known as subcategory classification, refers to distinguishing subordinate categories within a basic category such as distinguishing bird species.

such as cars [18], aircraft [19], shoes [20]. In fine-grained scenarios, it is necessary to learn critical parts of the object in image that can help align objects of same class and discriminate between neighboring classes. Fortunately, fueled by the recent advances in convolutional neural networks (CNN) lead to remarkable progress on fine-grained recognition [5, 1, 21]. However, most of the current state-of-the-art methods employee part/pose detection approach which is also a challenging task, to achieve good performances. For example, methods in [5, 1] heavily rely on the part annotations and train the CNN based on these parts, which is time-consuming and laborious. Despite the success of deep learning, it is still very challenging to learn critical parts, which refer to highly localized features extracted from images are essential to solving fine-grained recognition.

Within fine-grained recognition, bird species recognition is a widely-studied problem to ornithologists, and an important task in ecosystem monitoring and biodi-

versity preservation. Despite of this, bird species recognition is well-suited topic for investigation of multimodal research by integrating audio and visual data. On the other hand, sound also provides us important information about the world around us, especially for birds. Many animals including birds make sounds either for communication or their living activities such as moving, flying, mating etc. Although sound is in some case complementary to visual information, such as when we listen to something out of view, vision and hearing are often informative about the same structures in the world [22]. As a consequence, numerous efforts have been devoted to recognize bird species based on auditory data [23, 24] in recent years. Adapting CNN architectures for the purpose of audio event detection has become a common practice and generating deep features based on visual representations of audio recordings has proven to be very effective [25] such as in bird sounds [26, 24].

In the context of the background presented above, the main objective of this study is to apply multimodal fusion for fine-grained recognition using deep neural network and audio-visual data. To improve the affinity between images and sounds, we utilize the CNN to extract features from spectrogram of audio recordings.

The following research considerations may appear in this study:

- *What to fuse?* What feature representations to use for audio and visual modality. Since feature representations learned by CNN have shown significant improvement than handcrafted features on most learning task, we consider to extract CNN features from both modalities. This is also related to find the data (or dataset) of both modalities as well as matching categories or instances of those datasets. If the most suitable dataset is unavailable, it is related to cre-

ating a dataset with less effort.

- *When to fuse?* i.e., which fusion strategy performs best in our scenario? One of the main consideration is to know what strategy performs best when fusing both modalities. The most widely used strategy is to fuse modalities at the feature level, which is also known as early fusion. The other popular approach is decision level or late fusion strategy, which fuses decisions of scores obtained from each modality. Also, kernel fusion which also called the intermediate fusion by Noble [27], fuses multimodality at the kernel level.

- *How to fuse?* This is related to find or propose practically suitable methods under different fusion strategies while learning CNN features with both modalities.

1.2 Contributions

Our main contributions are summarized as follows:

- We propose that the combination of image and sound provide richer training signal for fine-grained bird classification under CNN framework, which is the first attempt to the best of our knowledge.

- Different fusion strategies are investigated for fusing audio and image modalities using CNN.

- We collect at least 10 audio recordings for each bird over 178 species, corresponding to the image dataset CUB-200-2011 [28].

Specifically, we adopt CNN to process jointly the two modalities for fine-grained bird classification in an end-to-end manner.

In Chapter 3, the kernel-based fusion strategy is studied for integrating deep neural features of both visual and audio data. In this study, we extract CNN-based deep features from both modalities and integrate these features by a more flexible, efficient kernel-based fusion method known as multiple kernel learning (MKL) [29, 30]. Experimental results indicate that MKL is an effective approach to improve classification performance while fusing different modalities.

In Chapter 4, three strategies are investigated for fusing audio and image modalities using CNN: (1) an early fusion strategy in which the feature vectors related to each modality are concatenated together and input to the CNN. (2) A middle fusion strategy. Features learned by each single modality are combined at the mid-level of the CNN. (3) A late fusion strategy. Outputs of single modality are fused to determine a final classification. Experimental results show that the architecture with late fusion strategy outperforms among the proposed architectures, which indicates that combining decisions of the classifiers from two modalities is superior. In addition, we apply a transfer learning procedure, which is a robust technique to enable leverage knowledge from learned deep learning model to new model, to improve classification accuracy.

The overview of our study on multimodal fusion strategies for fine-grained bird classification is illustrated in Fig. 1.2.

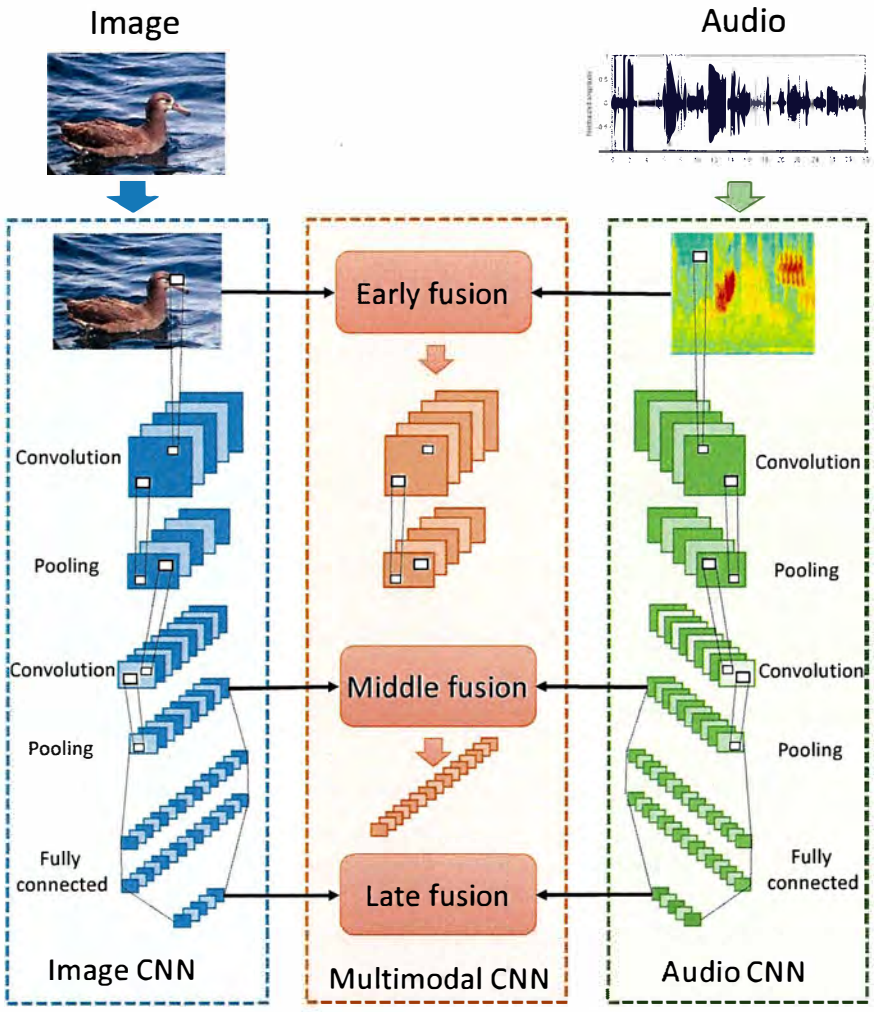


Figure 1.2: Overview of study on multimodal fusion strategies for fine-grained bird classification. a) presents the multimodal learning models with different fusion strategies which described details in Chapter 4. b) presents the kernel-based fusion method that combines deep features that are extracted from Image and Audio CNN in a) and combines them using multiple kernel learning. More details presented in Chapter 3.

1.3 Dissertation outline

The remainder of this dissertation is organized as follows.

Chapter 2 presents an overview of prior works in fine-grained recognition as well as bird species recognition with CNN, deep neural networks, multimodal fusion, and kernel-based fusion methods.

In Chapter 3, we present a study on kernel-based fusion for fine-grained bird classification by combining deep neural features of both modalities at kernel level. First, we introduce an integrated dataset with audio and video modality to conduct experiments for evaluating different fusion strategies on fine-grained bird recognition. Second, we train the audio and visual CNN individually applying the transfer learning, and extract deep neural features based on the activation values of an inner layer of the trained CNNs. Third, we combine these features by multiple kernel learning to perform the final classification. Finally, we discuss the experimental results comparing to single modality model and other kernel-based fusion methods.

In chapter 4, multimodal learning models proposed to combine audio and visual modality using CNN by different fusion (early, middle, late) approaches. Precisely, we focus on the evaluation of feature fusion performance of deep neural networks with different levels (i.e., layers) and investigate how multimodal fusion learning contributes towards fine-grained recognition. To address the integration of the audio and visual modality, quantitative evaluation and analysis are conducted on the integrated dataset by comparing the performance between the single modality model and multimodality models. Furthermore, to analyze the effects of the multimodal learning models, the learned features from each model are qualitatively evaluated.

The experimental results verified that the multimodal learning model with the late fusion strategy outperformed the other models. Subsequently, to confirm that the effectiveness of our late fusion approach, a comparative experiment conducted with other late fusion methods.

In Chapter 5, the accomplishments of our study on multimodal fusion strategies with CNN for fine-grained recognition using Audio-Visual modality are summarized.

Chapter 2

Literature Review

Over two decades, considerable efforts have been devoted to studying the relationships between different modalities. Notably, with the advent of deep neural networks in the last decade, a number of groundbreaking improvements have been observed in multimodal research [8, 31].

One of the important category of multimodal research comes from the field of multimedia content analysis, where the goal is to build systems that enable interactivity across various modalities such as text, image, video, audio, animation, etc. For example, in the cross-modal retrieval [32, 33], which aims to take one modality (e.g., text) as the query to retrieve relevant data of another modality (e.g., image).

A second essential research area of multimodal research is multimodal fusion (combined problem-solving approaches), e.g., in audio-visual recognition [10], action [7] and emotion [34] recognition, where improving recognition (e.g., speech) performance compared with single modal (e.g., audio information) recognition by utilizing complementary sources of information (e.g., visual information).

Our study is related to multimodal fusion with early, middle, late and kernel-

based fusion as well as deep neural networks for fine-grained recognition. The chapter gives detailed descriptions of these researches and highlights the connections and differences between our study and existing works.

2.1 Fine-grained recognition

Recently, a variety of approaches have been proposed for fine-grained recognition problem in different domains such as bird species [5, 1, 35, 36], dog breeds [17], plant species [15, 16] and product models [18, 20].

In the case of bird species recognition, current state-of-the-art methods typically adopt CNN-based end-to-end schemes [5, 1, 37, 38, 39], to learns high-level discriminative features for recognition. Since the visual differences between categories are too small and can be easily overwhelmed by other factors such as pose, viewpoint, or location of the object in the image [21]. A common approach for

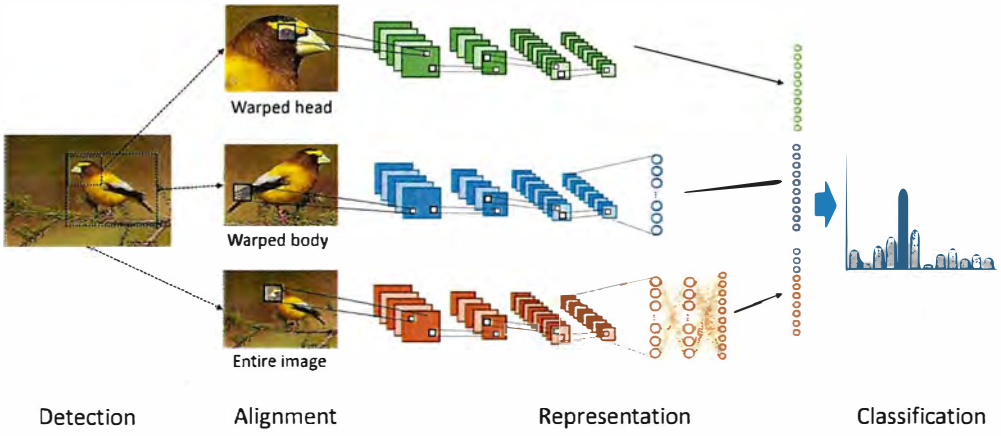


Figure 2.1: Pose normalized part-based deep neural networks [1]. Given a test image, the head and body region is detected and aligned using the learning model that learned pose prototypes from training images with keypoint annotation. Each region (head, body and entire region) is fed through a deep convolutional network, and features are extracted from multiple layers. Finally, features are concatenated into a single feature vector and fed to a classifier.

overcome such factors is to localize key parts of the object and to develop model based on these parts. For example, in fine-grained bird recognition, authors of [5, 1] used head and body as a key parts which are defined manually and the part detectors are trained in a supervised manner as shown in Fig. 2.1. However, annotating parts is a time-consuming and laborious as well as significantly challenging than labeling the image. Therefore, capturing key parts from object in the image is important and challenging part in fine-grained recognition. To overcome this issue, weakly supervised or unsupervised parts detection methods proposed in [40, 21, 41]. Nevertheless, subtle visual differences existed in local regions from similar fine-grained categories are still difficult to learn.

2.2 Deep neural networks

In classical computer vision, for example an image classification task, features are extracted from one class of objects (e.g., cars, birds) and treated these features as a sort of "definition" (known as a bag-of-words) of the object. For example, two simple features that can be extracted from images are edges and corners. Finally, the image is classified using these features. The difficulty with this classical (i.e. traditional) approach is that it is necessary to choose which features are important for that specific object. As the number or classes is increase, feature extraction and/or selection become more complicated. To decide which features are best to describe different classes of objects, a long trial and expert knowledge are necessary for the computer vision engineers.

Consequently, deep learning introduced the concept of **end-to-end** learning

where machine is told to learn what to look for with respect to each specific class of object. Comparing to the traditional computer vision, the deep learning model is trained on a dataset of images which have been annotated with what classes of an object are present in each image, where neural networks discover the underlying patterns in classes of images. With all the state of the art approaches in computer vision employing this methodology, the works of the engineers has changed from extracting hand-crafted features to deep learning architectures. The comparison between two techniques is shown in Fig. 2.2. The detail analysis of two techniques described in [42].

Deep learning approaches which employ deep neural networks (DNN) have successfully applied for single modality such as text [43, 44, 45], images [46, 47, 48] and audio [49, 50] showing their ability to learn representations directly from raw data and can be used to extract a set of discriminative features. Convolutonal neural network (CNN) is one powerful deep architecture of DNN commonly utilized for image classification [47, 48, 3, 51, 52].

One of the first CNN networks is LeNet-5 proposed by Le Cun et al. [53]

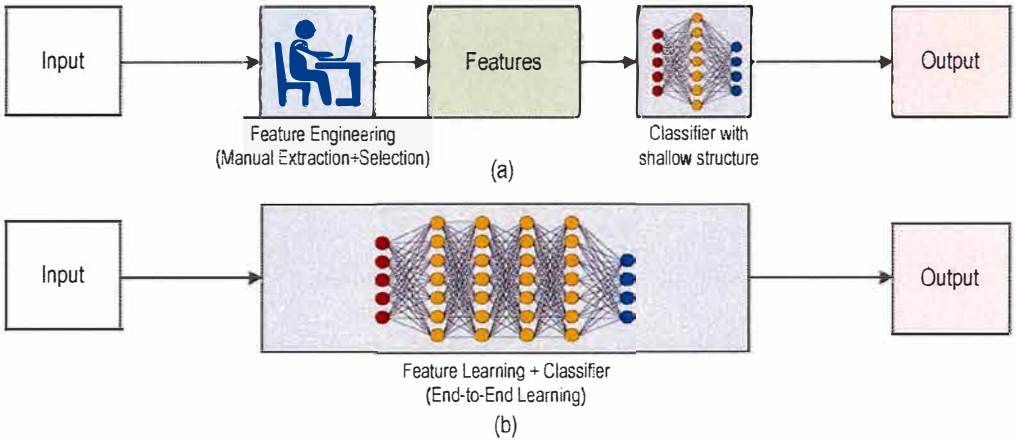


Figure 2.2: Traditional computer vision vs. Deep learning technique. The figure is taken from [2].

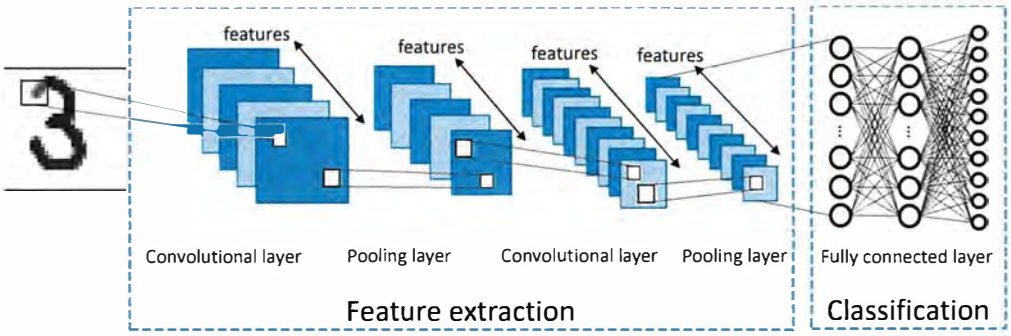


Figure 2.3: Typical CNN architecture with convolution, pooling and fully connected layers.

that recognizes hand written digits. It consists of a chain of convolutional layers, pooling (i.e. subsampling) layers and fully connected layers as shown in Figure 2.3. CNN is a type of feed-forward neural network architecture that uses learnable filters (weights) that slide or convolve across the input-space to analyze distance-pixel relationships. The pooling layer is applied to decrease the input-space so that computing time can be reduced. The last layers of CNN are fully connected and the final layer applies a soft-max function to its input to obtain probabilities for each class (in the case of handwritten digits recognition, the number of classes is 10). The network parameters are trained using back-propagation [48] algorithm as in the case of the usual neural networks.

In 2012, Krizhevsky et al. [3] achieved state-of-the-art performance in the ImageNet Large Scale Recognition Challenge (ILSVRC) by using deep CNN named AlexNet. The architecture of AlexNet demonstrated in Figure 2.4. AlexNet shown that deep or layered compositional architectures can capture salient aspects of given images through the discovery of salient clusters, parts, and mid-level features. An example of different activation layers, feature maps in the CNN is shown in Figure 2.5. It has proven that CNN architectures that learn multiple levels of abstrac-

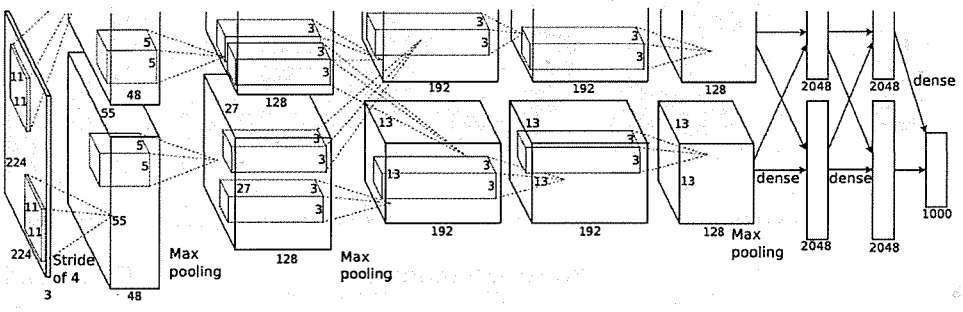


Figure 2.4: The architecture of AlexNet, which composed of 5 Convolution layers with 3 fully connected layers. The figure is taken from [3].

tion, can significantly improve the recognition accuracy in many object recognition tasks. The graph given in Figure 2.6 shows the success of CNN and its variants [3, 54, 51, 52] overtime on ILSVRC challenge.

The use of CNN for distinguishing between fine-grained recognition such as bird species categorization has been proposed in many studies [5, 1, 37, 38, 39], which employ part/pose based approaches to achieve good performance. Besides, CNNs have also been applied for speech processing [55, 56]. In [25, 57], authors utilized CNNs to extract features from spectral representations of audio recordings. As a consequence, numerous efforts have been studied to recognize bird species based on auditory data [23, 26] in recent years. It has become common practice adapting CNN models for the purpose of audio event detection and generating deep features based on the visual representations (such as spectrogram) of audio recordings has proven to be effective such as in bird sounds [23, 26].

In this study, instead of employing part/pose based approaches we will focus on integrating raw image an audio using conventional CNN, in order to know which fusion strategy performs best. On the other hand, since the visual representation from audio recordings proven to be effective when applying CNN, thus we use

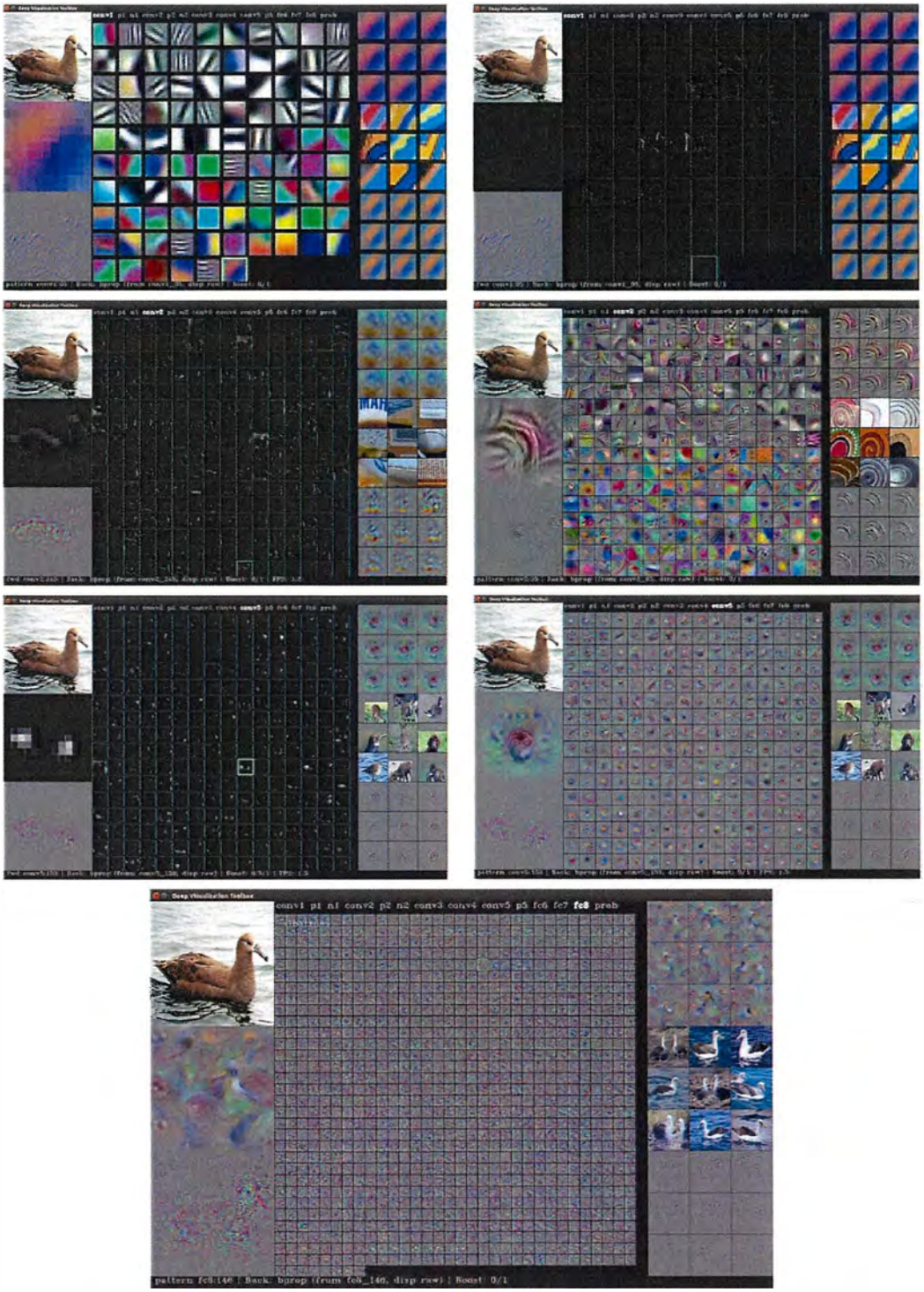


Figure 2.5: Visualization of feature kernels. The results are produced by using deep visualization toolbox by Yosinski et al. [4]

Figure 2.6: Performance of winning entries in the ILSVRC competitions from 2011 to 2017 in the image classification task.

spectrogram representations for audio data.

Transfer learning

One of the key reasons for the success of DNN is a large amount of data, which is required to train DNN in order to converge the cost function at a global minimum and avoid overfitting. However, for some tasks such as fine-grained recognition where the size of the dataset is significantly smaller than a general large-scale dataset (such as ImageNet [58]), a process known as transfer learning (Figure 2.7) can be used as a powerful tool to enable leverage knowledge (e.g., features, weights) from previously trained models for training new models without overfitting. A typical way to perform transfer learning is to train a DNN from a large dataset (such as ImageNet [58]) and fine-tune the network parameters (i.e., transfer the knowledge) on the target dataset. The reason why transfer learning works well on CNN is that

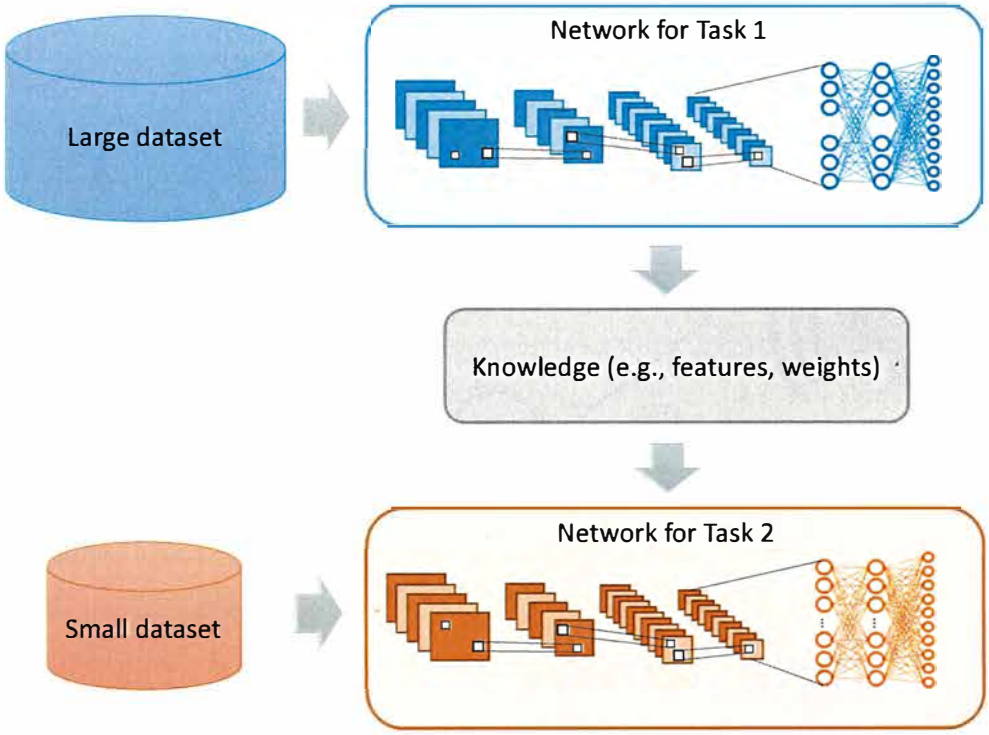


Figure 2.7: Transfer learning is a learning process of a new task relies on the previous learned task. The advantage of the transfer learning process is that a new model can be trained faster, more accurate even there is a small training data. It is often still beneficial to initialize with weights from a pre-trained model. For example, in [5], authors trained their CNN model by fine-tuning the ImageNet model on the fine-grained bird dataset.

CNN uses hierarchical features in its processing pipeline, which means the initial layers are likely learned a general features while late layers are high-level abstract features made from combinations of lower-level features and so these low-level and mid-level features can be used to initialize other CNNs. Recent studies [5, 1] on fine-grained classification have taken advantage of this fact to obtain state-of-the-art results.

Since transfer learning from a general dataset has proven to be an effective and efficient method for fine-grained recognition, we explore different methods for fine-tuning CNN on our multimodal fusion study.

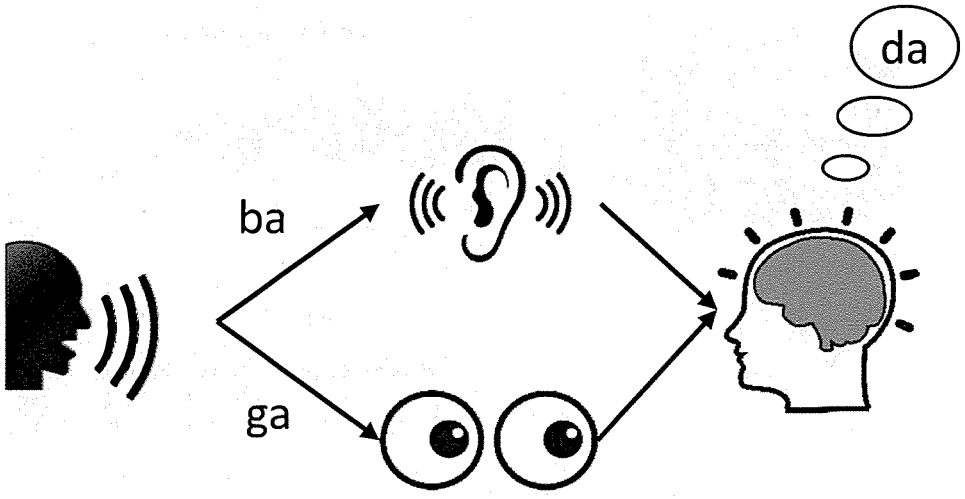


Figure 2.8: McGurk effect. When human hears the syllable "ba-ba" while see the mouth form "ga-ga" and perceives the new sound "da-da".

2.3 Multimodal fusion

Multimodal fusion is one of the original topic of multimodal research area. Multimodal research has a long history from audio-visual speech recognition to a recently renewed interest in computer vision applications which applies deep learning approach. According to [8], there are four eras of multimodal research as follows.

- The "Behavioral" era (1970-1980). One of the earliest works of multimodal research is audio-video speech recognition (AVSR), which is motivated by the McGurk effect [59] (Figure 2.8)-interaction between hearing and vision in speech perception. In the speech recognition community, most of the early studies [60] employed hidden Markov models (HMMs) at that time. Recently, AVSR receives the growth of an interest from the deep learning community [10].

- The "Computational" era (1980-2000). A second era of multimodal research includes the field of multimedia analysis and multimedia retrieval. In that

time, multimodal fusion [31] is a practical approach for improving the retrieval performance.

- The "Interaction" era (2000-2010). In the early 2000s, multimodal interaction with the goal of understanding human multimodal behaviors (such as emotion recognition [61]) during social interactions was an emerging field in multimodal research.
- The "Deep learning" era (2010s-until). Recent advance in deep learning has emerged in multimedia research to make progress in understanding the things or events by processing relate information from multiple modalities. Learning from multimodal data offers the possibility of gaining the deep understanding of things or events. Multimodal learning algorithms [62, 7, 6, 13, 14] have been studied for various recognition tasks by offering how neural networks can be used to construct shared or joint representation, how to train them and what advantages can be gained.

This dissertation relates to the deep learning era of multimodal research, which aims to study multimodal learning for fine-grained bird recognition by combining audio and visual data using DNN.

Multimodal fusion strategies in neural networks

Recently, multimodal learning algorithms performs state-of-the-art performance for tasks such as image sentence matching [62], action recognition [7], RGB-D object recognition [6], and speech recognition (audio-visual speech recognition [10] and visual-only speech recognition [9]). Among the many approaches for multimodal

learning, multimodal fusion is commonly realized by three different categories of approaches.

First, in early fusion approach, feature vectors from multiple modalities are concatenated and transformed to acquire a multimodal feature vector. A typical early fusion scheme is presented in Fig.2.9. For example, Ngiam et al. [13] utilized DNN to extract fused representations directly from multimodal signal inputs.

Likewise, in middle fusion approach, Huang et al. [14] employed deep belief network (DBN) to combine mid-level features learned by a single modality. General middle fusion scheme is presented in Fig.2.10.

Lastly, in late fusion approach, outputs of unimodal classifiers are merged to determine a final classification. A typical late fusion scheme is shown in Fig.2.11.

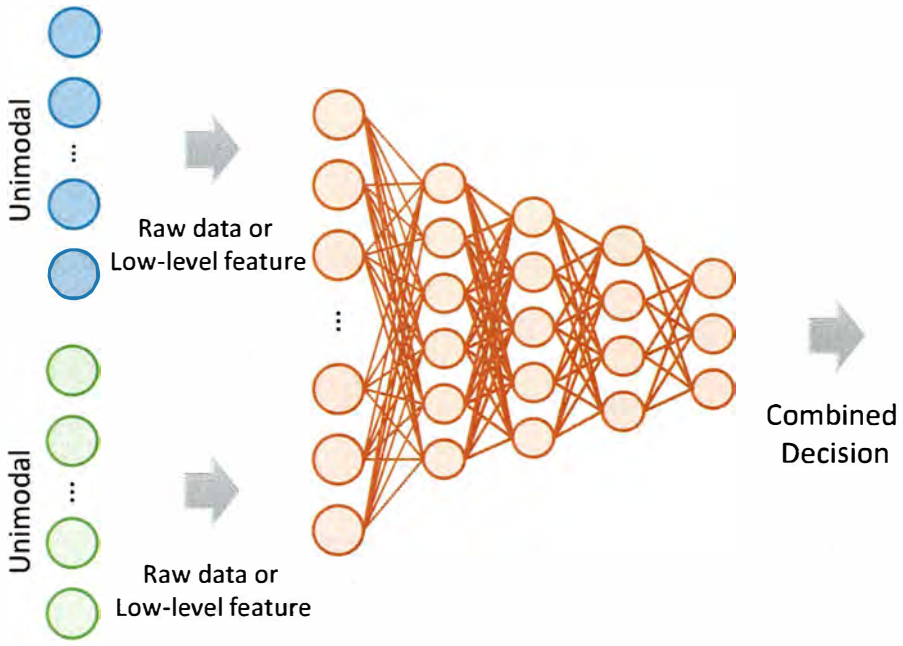


Figure 2.9: General early fusion scheme. Each modality concatenated to acquire a multimodal vector or each modality is learned individually as a first layer and joined into a shared representation as a second layer.

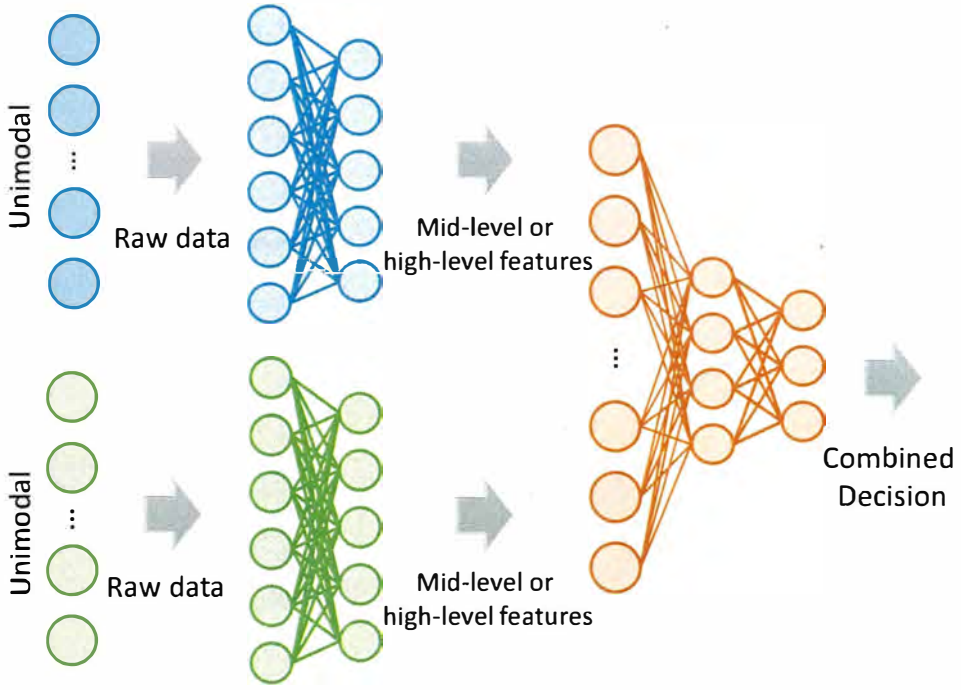


Figure 2.10: General middle fusion scheme. The fusion is performed in the middle of the model.

For example, in RGB-D object recognition, Eitel et al. [6] proposed two separate CNN streams processing RGB and depth data independently are combined with late fusion approach. Therefore, Simonyan et al. [7] proposed two-stream (one stream processing spatial features from RGB image inputs, while the other stream processing temporal features from optical flow inputs) network architecture designed to recognize an action for videos. They combined two streams by concatenating features and by averaging prediction scores from two CNNs, respectively. In contrast to these works, we propose simple yet effective concatenation, summation or multiplication based fusion methods with respect to different fusion strategies.

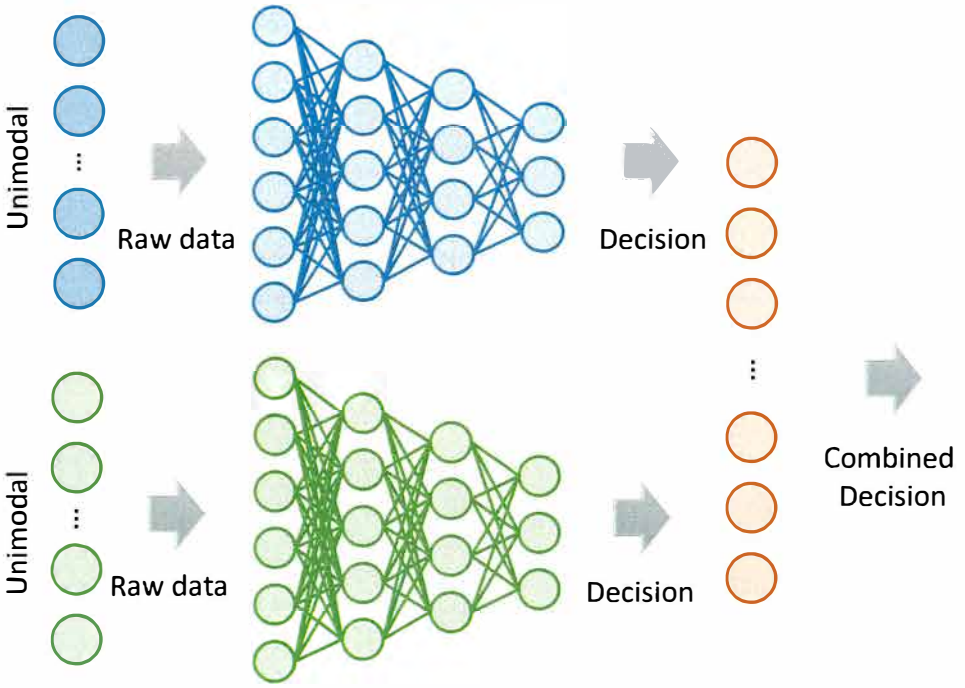


Figure 2.11: General late fusion scheme. The decisions of unimodal networks are joined to determine final decision.

2.3.1 Kernel-based fusion

Recently, kernel-based fusion methods provide a new approach for feature fusion. This subsection gives a brief introduction to a support vector machine (SVM), multiple kernel learning and the related works that employee kernel fusion.

Support vector machine

Kernel methods such as support vector machines (SVM) [63] have become a popular tool in data classification and many kinds of machine learning tasks since its introduction. Given a labeled training set $\{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a d-dimensional input vector and label $y_i \subseteq \{+1, -1\}$. SVM performs classification by finding the hyperplane that maximizes the margin between two classes. The use of kernel al-

lows the SVM to find the optimal hyperplane in the feature space induced by the mapping function $\Phi : \mathbf{R}^d \mapsto \mathcal{H}$.

The resulting linear discriminant function in the feature space is defined as

$$f(x) = \mathbf{w}^T \Phi(x) + b \tag{2.1}$$

The classifier can be trained by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T \Phi(x_i) + b) \geq 1 - \xi_i, \forall i \\ & 0 \leq i \leq N, \xi_i \geq 0. \end{aligned} \tag{2.2}$$

where $\mathbf{w}$ is the normal vector to the hyperplane, ξ_i is the slack variable which measures degree of misclassification of x_i, C is the cost parameter that controls the trade-off between the generalization of SVM and classification error, and b is the bias term.

By use of a kernel function, $k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$, it is possible to compute the separating hyperplane without explicitly carrying out the mapping into feature space where $\langle \cdot, \cdot \rangle$ presents inner product. To ensure that a kernel function actually corresponds to some feature space, it must be symmetric, continuous, and positive semi-definite. This kind of kernels called Mercer's kernel and typical choice for kernels are

- Linear kernel:

$$k(x_i, x_j) = \langle x_i, x_j \rangle \tag{2.3}$$

- Radial basis function (RBF) kernel:

$$k(x_i, x_j) = (\langle x_i, x_j \rangle)^d \tag{2.4}$$

- Sigmoid kernel:

$$k(x_i, x_j) = exp(\frac{-\|x_i - x_j\|^2}{2\sigma^2}) \tag{2.5}$$

- Polynomial kernel:

$$k(x_i, x_j) = tanh(\gamma \langle x_i, x_j \rangle - \theta) \tag{2.6}$$

Each kernel corresponds to some feature space and because no explicit mapping to this feature space occurs, optimal linear separators can be found efficiently in feature spaces.

Multiple kernel learning

Since kernel functions can be seen as similarity measures between a pair of instances, it is logical to assign different features x^k with different kernel functions k_k and combine them at the kernel level. Hence learning such a kernel combinations suitable to the problem has been an active area of research over the past few years. One way of learning kernels is via the Multiple kernel Learning (MKL) [29, 64, 65]

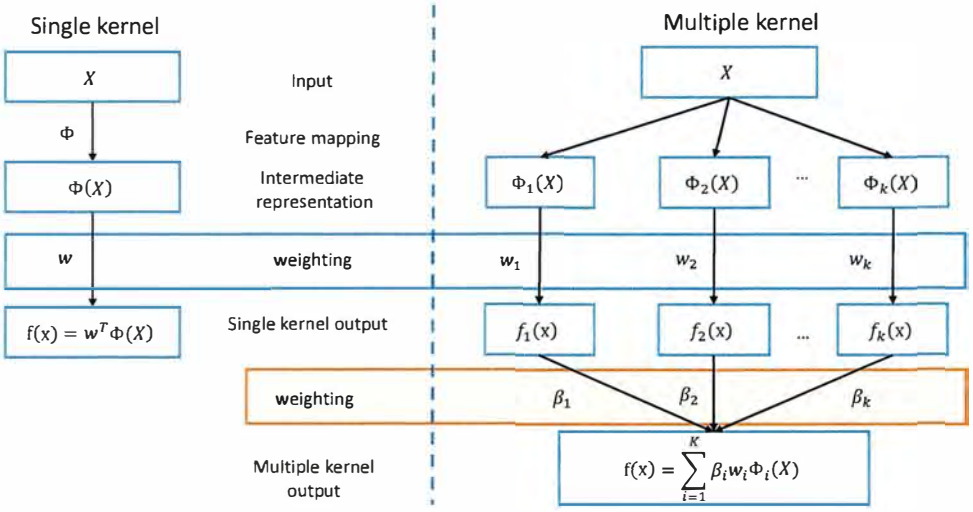


Figure 2.12: Single kernel SVM vs. Multiple kernel SVM. In multiple kernel, it is a weighted concatenation of feature maps induced by base kernels.

framework, in which the kernel k is learnt as a conic combination of the given base kernels $k_1, k_2, ..., k_k$: $k = \sum_{i=1}^{k} \beta_i k_i$, $\beta_i \geq 0, \forall i$. Here β is a coefficient to be learnt in the optimization problem. The difference between single kernel SVM and multiple kernel SVM is shown in Fig. 2.12. Details will be discussed in section 3.3.2.

Related works with kernel fusion

One practical kernel-based approach for feature fusion is to concatenate several features into one single vector and then train a single kernel classifier [66, 31]. For example, in [66], authors adopted SVM to detect semantic concepts in videos using visual, audio, and textual modalities. They used audio, video and text scores, and combined them in a high-dimensional vector before being classified by a single SVM. However, such concatenation requires proper normalization of features extracted from different sources; otherwise, the prediction would be easily dominated by predominant feature. Moreover, this method treats multiple features equally, be-

ing incapable of effectively exploring the complementary information of different modalities.

Another method of kernel-based fusion approach is multiple kernel learning (MKL) [29, 30], which learns optimal composite kernel through combining basis kernels constructed from different features of modalities. For example, Wu et al. [67] combined visual and textual features using l_p-norm MKL for modality classification of medical images. Yeh et al. [68] proposed a slight modified MKL framework that extracts heterogeneous features from data then construct multiple kernels for each feature by selecting different parameters for feature fusion. Since the multiple kernels can come from different sources of feature spaces offering improved classification performance, it has been successfully applied in object detection [69], multimodal affect recognition [70], emotion recognition [71, 72], and Alzheimer's disease classification [73]. Moreover, Poria et al. [70] used the combined feature vectors of textual, visual, audio modalities to train a classifier based on MKL (SPG-GMKL). They used CNN to extract features from the textual data.

2.4 Summary

In this chapter, related works regarding fine-grained recognition, deep neural networks, and multimodal fusion has been reviewed. The reviews in Section 2.1 clearly show that fine-grained bird recognition from an image is the quite challenging task and state-of-the-art algorithms typically adopt part/pose-based CNN, where it includes the annotation process which takes time and effort. Therefore, an investigation of multimodal fusion is crucial to improve the recognition performance.

Moreover, the reviews in Section 2.2 show that how deep neural networks have contributed towards achieving state-of-the-art performance in image recognition, speech recognition, as well as fine-grained recognition. Finally, the reviews in Section 2.3 show that multimodal research has a long history and how recent advances of deep learning renewed the multimodal research areas in terms of popularity and performance. Furthermore, different fusion strategies including early fusion, late fusion, and kernel fusion are reviewed in detail.

Based on these backgrounds, the research interest in this dissertation is to seek possibilities for multimodal fusion on fine-grained bird recognition using deep neural networks. This dissertation is composed of the following two parts. First, the study on the extraction of deep neural features from audio and visual modality and combine these features using multiple kernel learning for fine-grained bird classification is done in Chapter 3. Second, CNN-based multimodal learning models with three fusion strategies (early, middle, late) is proposed to settle the issues of combining multimodality.

Chapter 3

Audio and Visual Deep Feature Fusion using Multiple Kernel Learning

3.1 Introduction

In this chapter, we present a study on classifying bird species by combining deep neural features of both visual and audio data using kernel-based fusion method. Specifically, we extract deep neural features based on the activation values of an inner layer of CNN. We combine these features by multiple kernel learning (MKL) to perform the final classification. In the experiment, we train and evaluate our method on a CUB-200-2011 standard data set combined with our originally collected audio data set with respect to 200 bird species (classes). The experimental results indicate that our CNN+MKL method which utilizes the combination of both categories of data outperforms single-modality methods, some simple kernel combination meth-

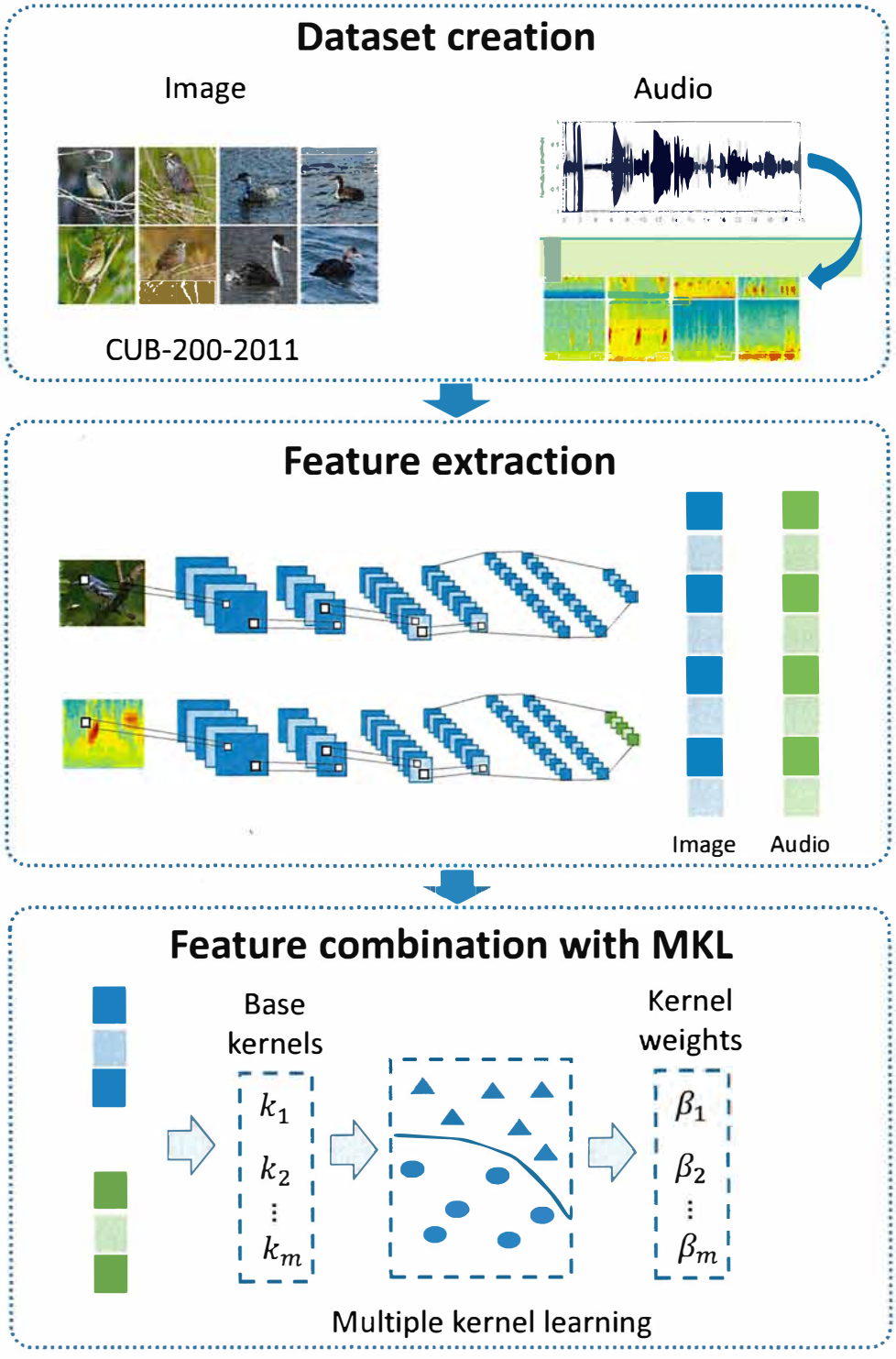


Figure 3.1: Overview of our kernel-based fusion for fine-grained bird classification. After the dataset creation which is presented in Chapter 3.2, the deep neural features from both modalities are extracted using fine-tuned CNN and combined at kernel level using multiple kernel learning.

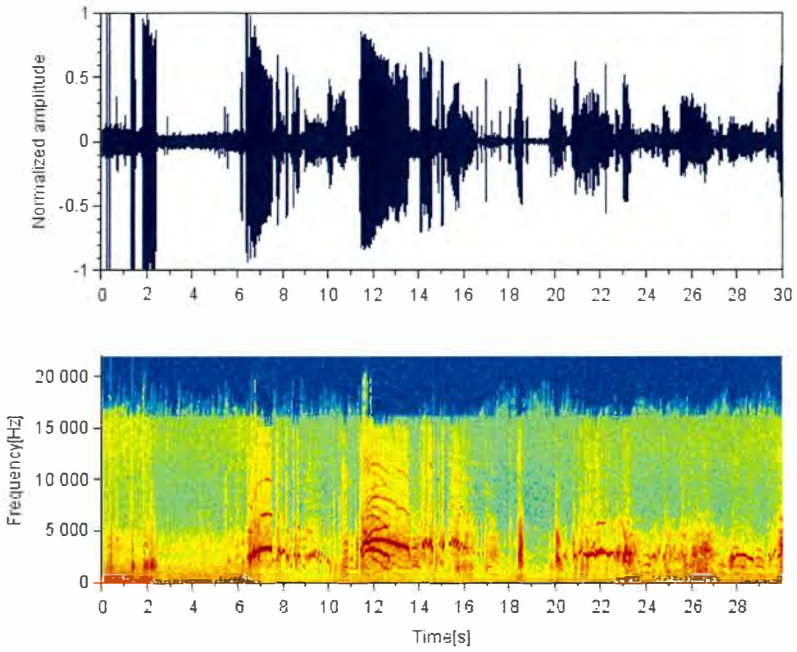
ods, and the conventional early fusion method.

3.2 Dataset

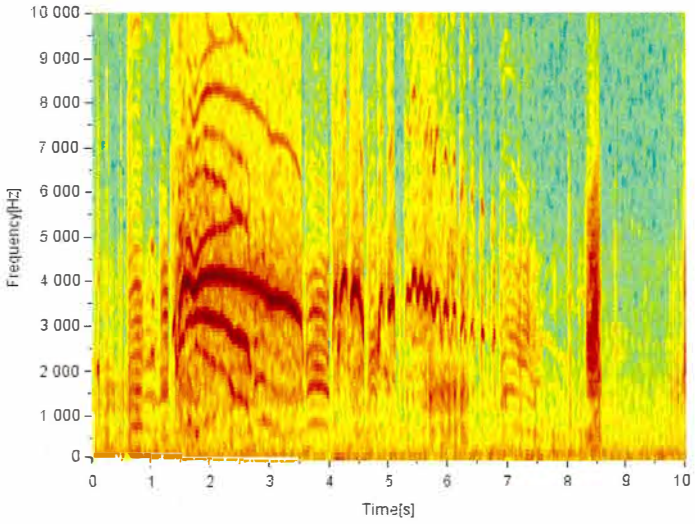
In this study, we use the popular fine-grained CUB-200-2011 [28] bird dataset and our originally collected sound dataset from sharing bird sound database Xeno-Canto[1].

CUB-200-2011 [28] dataset contains 11788 images of 200 species of birds, with each image downscaled to 227 $\times$ 227 pixel. Spectro-temporal features (spectrogram) are extracted from audio recordings that we collected from the Xena-Canto to be used as the audio representation. Based on the 200 species of the CUB-200-2011, we try to harvest at least 10 different audio recordings for each species. As a result, audio recordings over 178 species were collected completely (#recordings $\geq$ 10), audio recordings from 19 species were collected deficiently (0 $<$ #recordings $<$ 10), and 3 species could not be collected. The spectrograms of the audio are obtained by using short-time fourier transform (STFT) over 10 seconds audio frames, windowed with Hanning window (size 512, 50% overlap). The reason is, the sounds of birds are usually contained in a small portion of the frequency range (mostly around 2-8 kHz) as stated in [24], so we only extract features from the range of (0, 10) kHz. In order to focus only sounds produced in the vocal organ of birds (i.e. calls and songs), first we obtained the maximum amplitude of the audio and removed a frame which contains only amplitude less than $1/4$ of the maximum amplitude. Finally, the spectrograms are saved as 227$\times$227 pixel color

[1] http://www.xeno-canto.org



(a)



(b)

Figure 3.2: (a) Radio (top) and spectrogram (bottom) of the black-footed albatross. (b) The spectrogram of 10 seconds duration, which will be fed into the bird classification model.

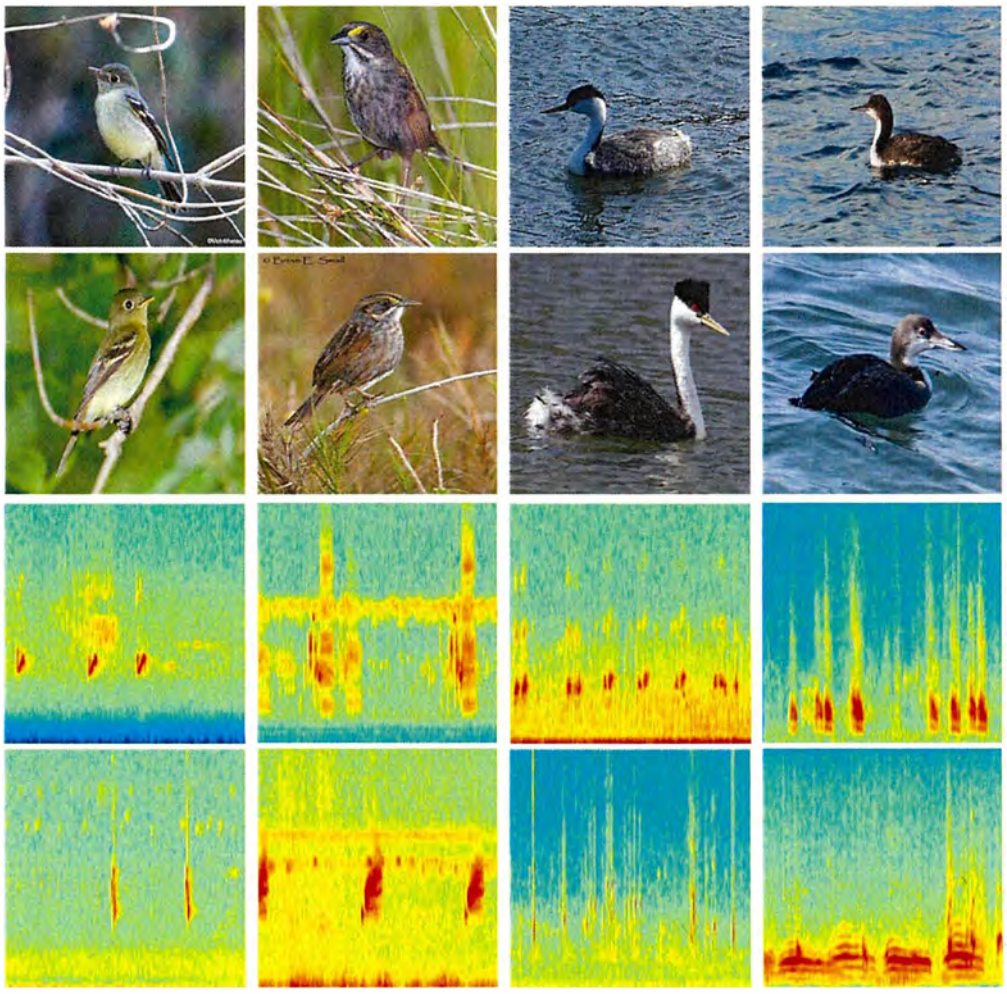
images. The generation of the spectrogram is shown in Fig. 3.2.

Since we could not harvest complete audio recordings for 200 species of CUB-200-2011, we had an imbalanced dataset for audio data. There are a few ways to

address this problem such as under-sampling, over-sampling etc., The idea of under-sampling or data reduction techniques that remove only a majority of class samples is not an ideal solution. Another approach is to apply over-sampling method [74], which have been proposed as a solution to imbalanced datasets. Consequently, simple random minority over-sampling has been performed to balance the dataset through duplicating some random samples of the deficiently collected classes that have $\#\text{recordings} > 5$. As a result, the audio dataset contains 4807 spectrograms of 194 species of birds. Several examples including both images and spectrograms over different bird species are shown in Fig. 3.3. We follow the standard training/test split of CUB-200-2011 dataset suggested in [28]. The sound dataset is split into two halves for training and test set respectively.

Both Audio and Visual CNNs are trained in a supervised manner, thus we create integrated dataset by matching two datasets and corresponding labels using HDF5 [2] file format. The most straightforward and common approach of matching two datasets is to generate all possible combinations of image and audio samples for each class, which creates $N \times M$ pairs of samples when there are N image and M audio samples for a certain class. This approach can be good solution while enlarging dataset, but large duplicate samples can lead to overfitting. Although it was reported in [75] that usually duplicating samples in a dataset has a detrimental effect on the model and accuracy rate. Therefore, random matching has been performed to match two datasets through randomly picking 5 audio samples for each image samples.

[2] https://www.hdfgroup.org/HDF5/



(a) (b) (c) (d)

Figure 3.3: An example of CUB-200-2011 and audio dataset: (a) the yellow bellied flycatcher, (b) the seaside sparrow (c), the western grebe, and (d) the pacific loon.

3.3 Methodology

Extraction of deep neural features of audio and image modalities and combining them for classification using MKL are described in this section.

3.3.1 Feature extraction

We use a fine-tuned CNN as a trainable feature extractor to extract deep neural features for both modalities. CaffeNet [76], a variation of the structure proposed by

Krizhevsky et al. [3] (Section. 2.2), which is conventional CNN for large-scale image classification used to extract CNN-based features from both modality. The CaffeNet consists of 5 convolutional (conv) layers (with ReLU activation, max pooling layers follow the first, second and fifth convolution layer, and local normalization is applied in the first and second convolutional layer) followed by 3 fully connected (FC) layers with ReLU activation and a softmax classification layer as shown in Fig. 3.4.

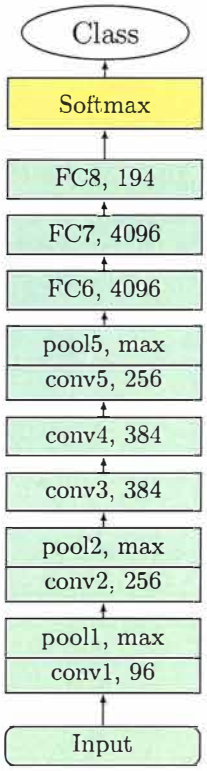


Figure 3.4: The CNN architecture used to train the image and audio modality. The comma separated parameter after the convolutional layer indicates the number of channels.

Images and spectrograms from both datasets were downsized to 227×227 pixels according to the CaffeNet architecture.

We train the image and audio CNN separately by adapting the weights and biases of the first seven layers derived from CaffeNet pre-trained network, discarding

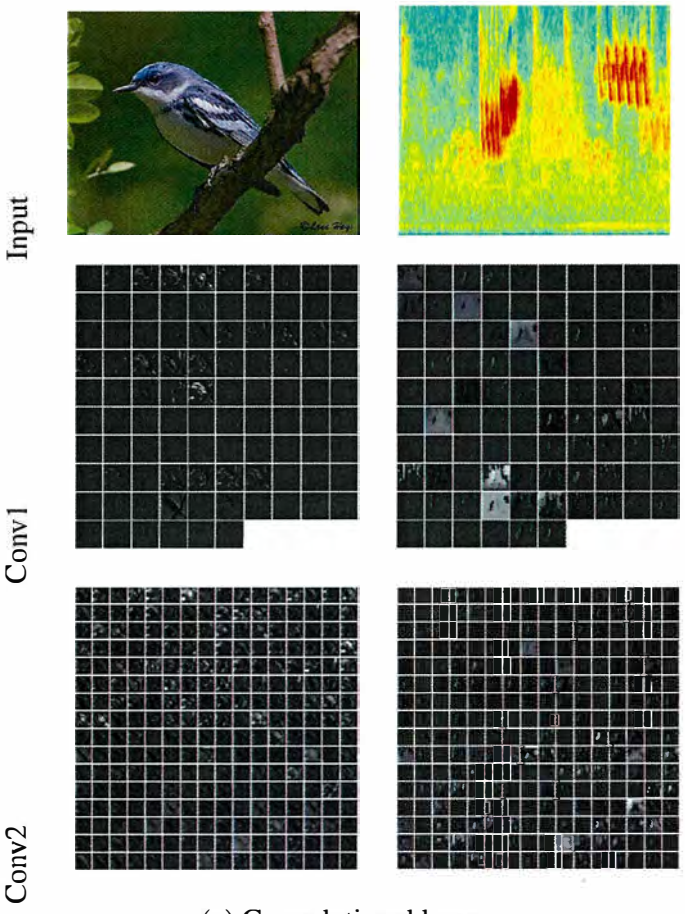
the last fully connected layer. Instead of last fully connected layer of the pre-trained model, we randomly place the initialized new fully connected layer for 200-class bird classification (in our experiment, 194 classes due to the lack of audio dataset).

The CNNs were trained using mini-batch stochastic gradient descent with batch size = 32. We initialized the fine-tuning learning rate as 0.001, which is a tenth of the initial CaffeNet learning rate and dropping it by hand whenever the test error stops improving. After this training, we can extract deep neural features based on the activation values of the inner layers of each CNN. In this paper, we use the activation values (194 features of visual and audio modality respectively as shown in Fig. 3.5b) of the last FC layer of each CNN as the feature vectors to train MKL.

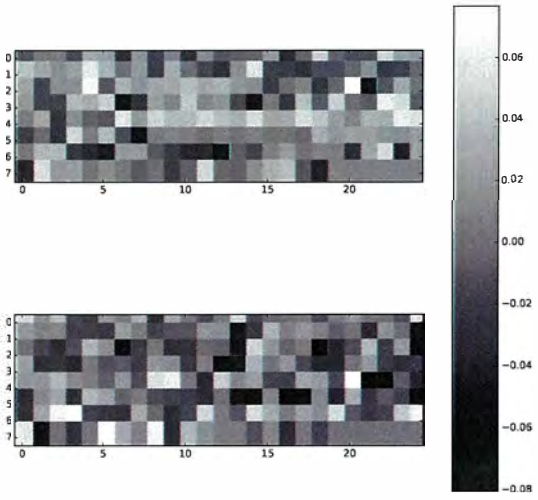
3.3.2 Feature combination with MKL

In order to combine audio and visual features, and train the final classifier using MKL, we create multimodal feature dataset by matching two datasets and corresponding labels. We perform random matching by randomly picking five audio samples for each image sample. Furthermore, we concatenate pair of features to obtain multimodal feature ($194 + 194 = 388$ features) and this combined feature vector along with the labels are used to train a classifier with MKL.

Unlike SVM (Section. 2.3) based on a single kernel function, MKL uses multiple kernels and learns the optimal convex combination of them. Research in MKL [29] has focused on both developing new MKL formulation (a linear combination, non-linear combination and data-independent combination) as well as their optimization. Given a labeled training set $\{(x_i, y_i)\}_{i=1}^{N}$, where x_i is a d-dimensional input vector and label y_i. A linear (convex) combination of base kernels is formu-



(a) Convolutional layer



(b) Fully connected layer

Figure 3.5: Deep neural features of the intermediate layers in CNNs. These are examples of activation values at different layers in CNNs, where (a) shows the features of a different convolutional layer of image (left) and audio (right) network, and (b) shows the last FC layers of both networks which are used to train MKL.

lated as follows:

$$\mathbf{k}(x_i, x_j) = \sum_{k=1}^{K} \beta_k \mathbf{k}_k(x_i, x_j) \tag{3.1}$$

where $\beta_k > 0$, $\sum_{k=1}^{K} \beta_k = 1$, $\mathbf{k}_k$ is a $k^{\mathbf{th}}$ of the K available kernel, and β_k denotes the weight of the $k^{\mathbf{th}}$ kernel.

In the primal form of MKL for classification, x_i is translated via K mappings $\Phi_k(\mathbf{x}) \longmapsto \mathbf{R}^d, k = 1, ..., K$ from the input into K feature spaces $(\Phi_1(x_i), ..., \Phi_K(x_i))$ where d denotes the dimensionality of the feature space. Then MKL solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2} \sum_{k=1}^{K} \frac{1}{\beta_k} \|\mathbf{w_k}\|_2^2 + C \sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & y_i (\sum_{k=1}^{K} \langle \mathbf{w_k}, \Phi_k(x_i) \rangle + b) \geq 1 - \xi_i, \forall i \\ & \sum_{k=1}^{K} \beta_k = 1, \beta_k \geq 0, \xi \geq 0. \end{aligned} \tag{3.2}$$

Here, $\mathbf{w}$ is the normal vector to the separating hyperplane, b is a bias term, C is the trade-off between the generalization of MKL and its training errors which are similar to SVM, and $\langle \cdot, \cdot \rangle$ presents inner product.

We can see that the formulation above imposes an l_1-norm regularization on the kernel weights when $\|\beta\|_1 = 1$, which tends to have sparse optimal solutions (i.e., during the learning most kernels are assigned zero or very small weights). Consequently, a non-sparse version of MKL is proposed by Kloft et al. [64], where an l_2-norm regularization ($\|\beta\|_2 = 1$) is imposed instead of l_1-norm ($\|\beta\|_1 = 1$). Furthermore, Kloft et al. [65] proposed efficient optimization method for arbitrary l_p-norm ($\|\beta\|_p^p \leq 1$) with $p \geq 1$. In this work, l_p-norm MKL [65] algorithm is em-

ployed to learn the optimal convex combination of multiple kernels and optimizing kernel weights. We also try SPG-GMKL [77] implementation.

3.4 Experiments and Results

In this section, we compare our method with single-modality models and other feature combination kernel methods for bird species classification. MKL-based methods explored are l_p-norm MKL and SPG-GMKL. Other compared methods include training combined features using single kernel SVM classifier (CNN+SVM) and an averaging kernel (AverageMKL) which is a simple kernel combination.

In the experiments, we use the implementations of SVM, SPG-GMKL[3], and l_p-norm MKL in Shogun toolbox[4] with one-vs-all approach for the multi-class problem. In all experiments, the cost parameters $C \in \{1, 10, 100, 1000\}$ and other parameters such as kernel parameters are tuned by 5-fold cross-validation on the training set. The performances of all methods are measured by their accuracies on

Table 3.1: The accuracy (%) of SVMs by training with single modality feature and concatenated features with different kernels.

Kernel/Feature	Image	Audio	Multimodal
Linear	35.16	58.10	74.74
RBF	35.27	58.57	**75.74**
Poly	35.14	58.12	74.76
Sigmoid	**36.10**	**59.01**	74.85
Laplacian	**36.10**	**59.01**	74.67
Chi-squared	34.44	58.25	73.67
Additive chi-squared	33.40	57.23	73.65

[3] http://www.cs.cornell.edu/~ashesh/pubs/code/SPG-GMKL/download.html

[4] http://www.shogun-toolbox.org

the test set.

3.4.1 Quantitative results

Since this dissertation aims to integrate image and audio modality, it is necessary to compare the performance between the single modality method and feature combination method.

As a baseline, we first evaluate the performance of a single kernel SVM (i.e. single modality) on each feature individually. We test linear (Eq. 2.3), RBF (Eq. 2.4), Polynomial (Eq. 2.6), Sigmoid (Eq. 2.5), Laplacian, Chi-squared and Additive chi-squared kernels. The results are shown in Table 3.1. The results show that the audio modality performed better than image modality. Furthermore, as shown in Table 3.1, we also trained combined features using a single kernel classifier (CNN+SVM), which is one of the practical kernel-based feature fusion approaches. It can be seen that using only a single kernel classifier for multimodal feature improved the classification accuracy by a large margin. The best performance of each feature and multimodal feature are given in bold font. It shows that RBF kernel achieves the best performance for multimodal feature, and gets third for the single modalities.

For MKL experiments, five RBF with gamma from $\{2^{-13}, 2^{-12}, 2^{-11}, 2^{-10}, 2^{-9}\}$ and four polynomial kernel with powers of $\{1, 2, 3, 4\}$ with similar to [70] are used for evaluation. For l_p-norm MKL methods, we consider norm $p \in \{1, 1.2, 1.5, 2, 4, 8, 16\}$, which are presented as "CNN+MKL" with their norms. The classification performance of different methods are given in Table 3.2. Compared to single-modality methods, all methods that utilize the combination of both features efficiently improve classification accuracy. Moreover as can be seen, "CNN+MKL" methods out-

Table 3.2: The classification performance of different feature combination methods.

Method	Accuracy (%)
CNN+SVM (Early)	75.74
CNN+AverageMKL	74.09
CNN+SPG-GMKL	73.24
CNN+MKL ($p = 1$)	77.61
CNN+MKL ($p = 1.2$)	78.09
CNN+MKL ($p = 1.5$)	78.13
CNN+MKL ($p = 2$)	78.01
CNN+MKL ($p = 4$)	78.09
CNN+MKL ($p = 8$)	**78.15**
CNN+MKL ($p = 16$)	78.11

perform other feature combination methods such as early fusion and AverageMKL. SPG-GMKL Moreover, l_p-norm MKL achieves their best performance at 78.15 with $p = 8$.

3.4.2 Qualitative results

We perform a qualitative study to analyse the effects of our method by comparing single modality and feature combination methods. First, we select some classes where the feature combination methods provide the correct answer while the single modality method produces the wrong classification. Figure 3.6 shows some examples of single modality vs. combined features classification. In the first column, the single modality method predicts the input image and spectrogram as the 'black footed albatross' and 'boat tailed grackle' respectively rather than 'nighthawk'. The single modality method also misclassified in case of the second column. However, the feature combination methods are able to predict right answer proving that com-

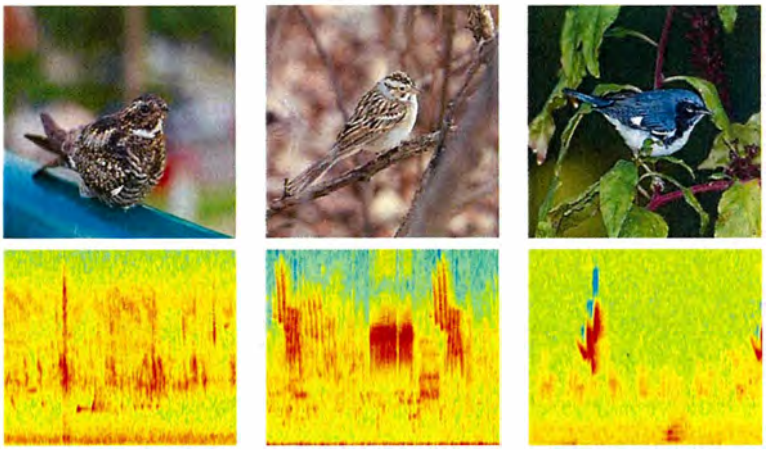


Image	✗	✗	✗
Audio	✗	✗	✓
CNN+SVM	✓	✓	✓
CNN+AverageMKL	✓	✓	✓
CNN+MKL	✓	✓	✓

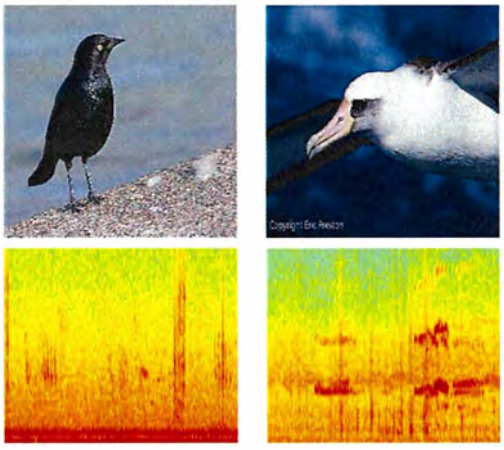


Image	✗	✗
Audio	✗	✗
CNN+SVM	✗	✗
CNN+AverageMKL	✗	✗
CNN+MKL	✓	✓

Figure 3.6: Effects of combining image and audio features. Top two rows show input of sample image and spectrogram of different bird species. The bottom rows show the resulting classification of single modality (Image and Audio) and feature combination methods.

bining features of different modalities can improve the classification performance. As shown in the middle column, we observed when the single modality method provides right answer for spectrogram, the probability of providing the right an-

swer of feature combination methods is higher than classification is correct for an image. Let us mention that the audio modality performed better classification accuracy rather than image modality. We think that audio modality is dominant feature provides informative and discriminative information in our case. Furthermore, the CNN+MKL method is able to provide the right answers while the other methods provide misclassification as shown in the last two columns.

For l_p-norm MKL with different norms, the kernel weights (β) of the base kernels are further analyzed and shown in Fig. 3.7. We can see that l_1-norm MKL produces sparse kernel combinations which focus the weights on one kernel and give zero weights to other kernels. When the norm gets bigger, larger weights are assigned to other kernels, and the distributions of weights are close to uniform.

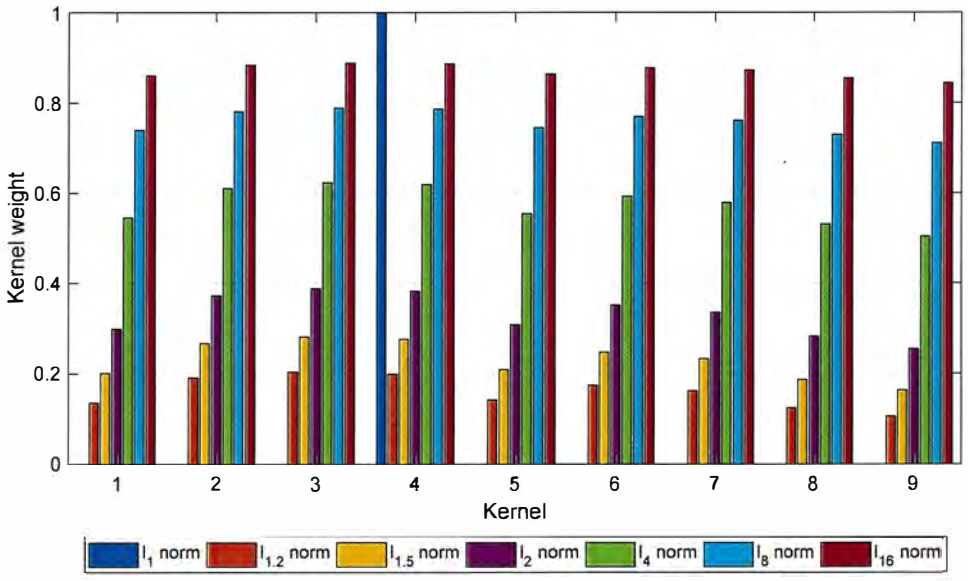


Figure 3.7: Base kernel weights of different l_p-norm MKL.

3.5 Summary

In this chapter, we have presented a study of classifying bird species with audio-visual data using CNN and MKL. Specifically, we used deep CNN to extract features from audio and visual modality and combined these features to perform classification using MKL. In this study, we employed l_p-norm MKL using two kernel functions (RBF and polynomial) in order to enhance the classification performance for bird species. Our experiments indicate that MKL is an effective approach to improve classification performance while fusing different features.

Chapter 4

Multimodal Learning for Fine-grained Classification with Audio-Visual Data

4.1 Introduction

In this chapter, we focus on classifying bird species by exploiting the combination of both visual (images) and audio (sounds) data using CNN, which has been sparsely treated so far. In essence, we propose CNN-based multimodal learning models in three types of fusion strategies (early, middle, late) to settle the issues of combining training data of both modality. The advantage of our proposed method lies on the fact that we can utilize CNN not only to extract features from image and audio data (spectrogram) but also to combine the features across modalities. The overview of proposed method is presented in Fig. 4.1. In the experiment, we train and evaluate the network structure on a comprehensive CUB-200-2011 standard

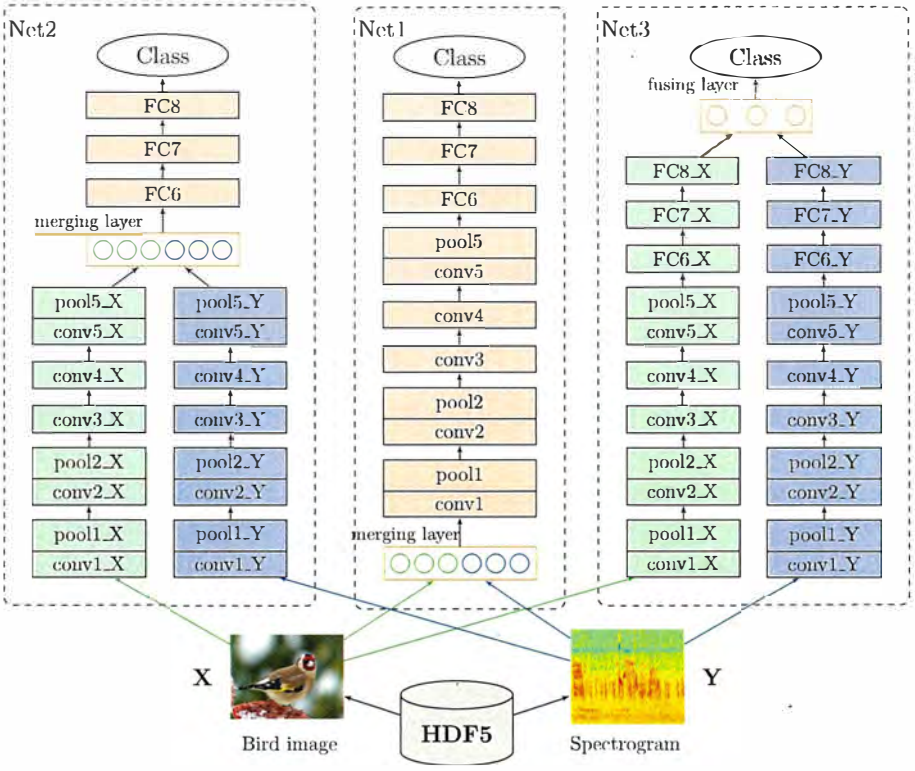


Figure 4.1: Overview of the multimodal learning models with different fusion strategies.

data set combing our originally collected audio data set (Section 3.2) with respect to the data species. We observe that a model which utilizes the combination of both data outperforms models trained with only an either type of data. We also show that transfer learning can significantly increase the classification performance.

4.2 Methodology

Our multimodal architectures extend conventional CNN for large-scale image classification [3]. Our implementation is based on CaffeNet, and can be treated as [76] a variation of the structure proposed by Krizhevsky et al. [3].

4.2.1 Feature Extraction

CNN uses hierarchical features in its processing pipeline. The features from initial layers are primitive while late layers are high-level abstract features made from combinations of lower-level features. The CaffeNet consists of five convolutional layers (with max pooling layers following the first, second and fifth convolution layer) followed by three fully connected (FC) layers and a softmax classifier. Rectified linear unit is applied to every convolutional layer and fully connected layer and local normalization are applied in the first and second convolutional layer. The process through this 8-layer CNN network can be treated as a process from low to mid to high-level features. In Fig. 4.2, it can be seen that low-level features can be extracted in the early layers of CNN and the high-level feature can be extracted in the late layers of CNN. Hence, we hypothesize, that combining the features of different layers in this pipeline can lead to achieve better performance.

4.2.2 Feature Fusion

We propose our method in three strategies to fuse features: early fusion, middle fusion, and late fusion. Early fusion, also known as feature level fusion, is a feature combination scheme that features from multiple modalities concatenated to form a merged feature vector. Middle fusion, also called mid-level combination, combines the high-level features learned by a single network. We use concatenation to combine low and high-level features in order to acquire merged multimodal features and to allow the CNN to learn joint features from merged features. Late fusion, also called decision-level fusion, combines the outputs of single modality

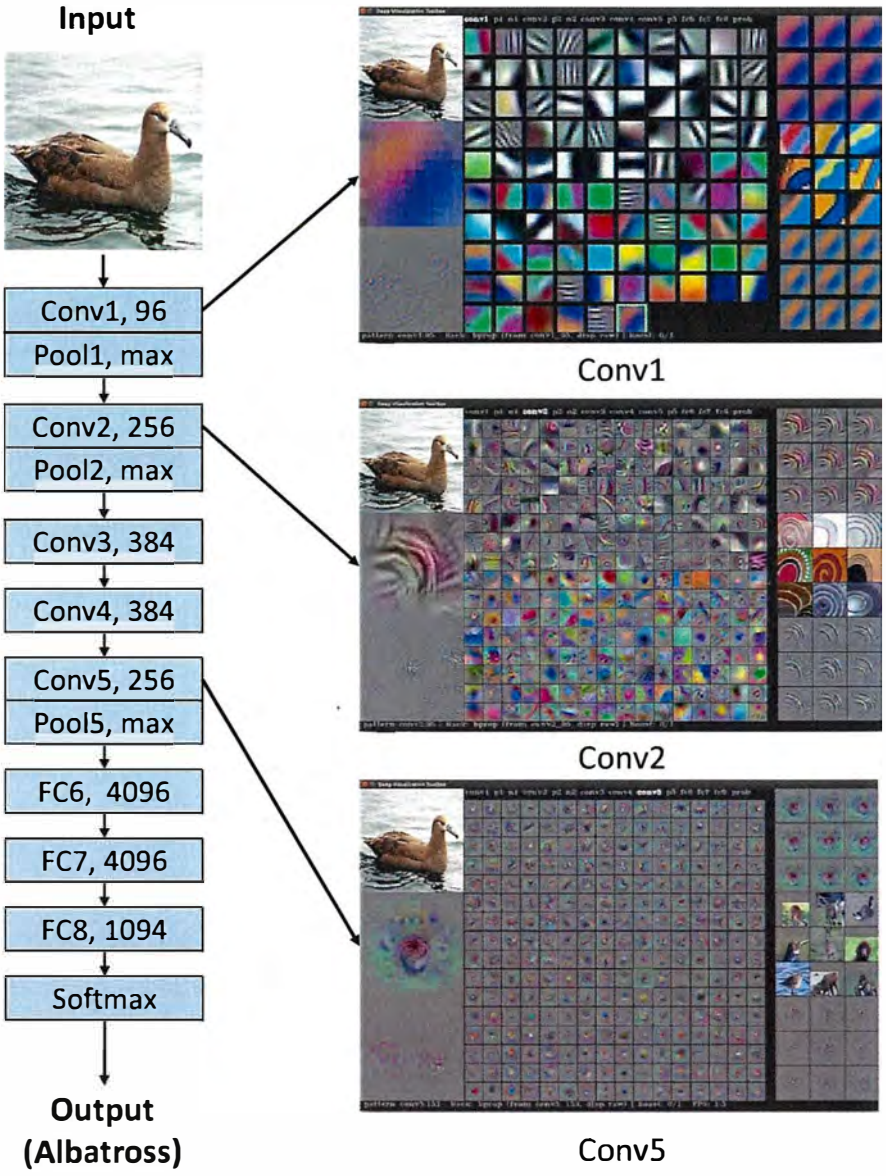


Figure 4.2: Visualization of CNN-based features based on different layers in CNN. The results are produced by using deep visualization toolbox by Yosinski et al. [4].

and determine the final classification. We use a simple summation or a multiplication to perform a fusion of decisions obtained from each network. The summation and the multiplication compute an element-wise sum and element-wise product of each network's output. The feature combination layers can be trained with standard back-propagation and stochastic gradient descent.

4.2.3 Multimodal Fusion Architectures

The proposed multimodal learning models which combine audio and image using CNN by different fusion approaches are described in this section. We exploit the same architecture for both audio and image modalities to focus on evaluating the effectiveness of the feature combination approaches.

Early Fusion Model

One direct approach for combining audio and image is to train a CNN over the concatenated audio and image data as shown in Fig. 4.3. In this strategy, the input vectors related to each modality are concatenated together and then processed together throughout the rest of the CNN pipeline. This model is the most computational efficient comparing to the middle and late fusion models because the number of learnable weight parameters is almost half times less than the late fusion model.

Middle Fusion Model

In the middle fusion strategy, unimodal features are extracted independently from audio and images, then combined into a multimodal representation by concatenating the activations of the last pooling layers of two modalities. The multimodal

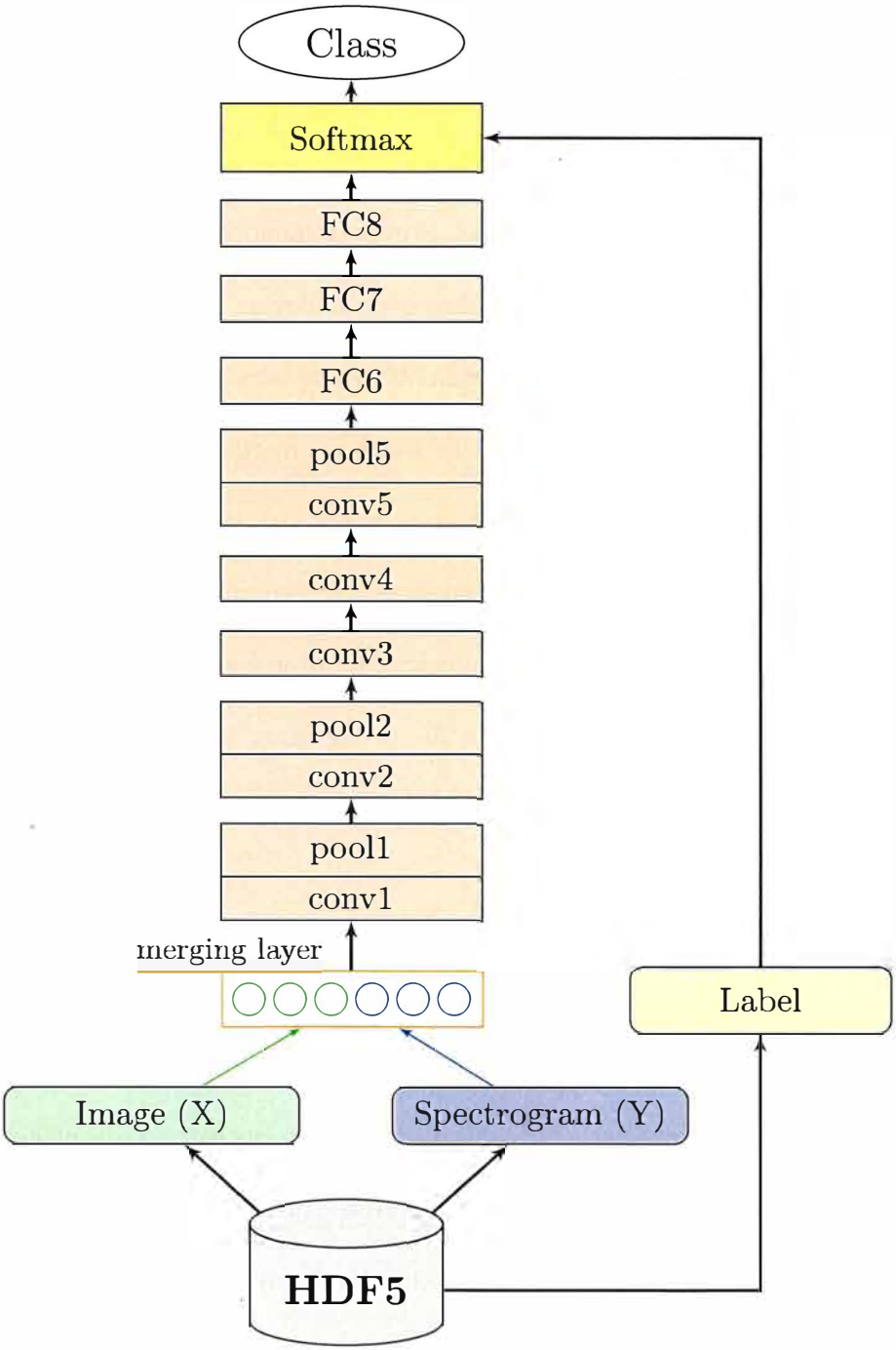


Figure 4.3: The architecture of early fusion model (Net1). 227 × 227 pixel RGB images of two modalities are concatenated at merging layer, which produces 227 × 454 × 3 output volume, and the convolution layers will extract joint features from this merged volume. We use HDF5 format to manage datasets of two modalities, because of it's flexible data storage and unlimited data types.

representation then learned in the following fully connected layers. The middle fusion model is shown in Fig.4.4.

Decision or Late Fusion Model

In contrast to the middle fusion model, extracted unimodal features are separately learned to compute unimodal scores, then these scores are integrated to determine a final score. The late fusion model consists of two-streams processing audio and image data (green and blue) independently, which are fused after last fully connected layers as shown in Fig.4.5. Among the various ways of combining CNNs with late fusion approaches, one straightforward way is to concatenate the output of each network and add an additional fully connected layer on top of this for classification. Instead of using FC layer to combine the two streams, we applied element-wise summation and element-wise multiplication to fuse the outputs of each stream.

4.3 Experiments and Results

The experiments in this section are conducted to evaluate the effectiveness of our proposed architectures (Net1, Net2, Net3). In order to evaluate an advantage of our late fusion approach, we conduct a comparative experiment between Net3 model and two different existing fusion approaches [6, 7]. Besides, we have fine-tuned a pre-trained model for all the models (including comparative models) in order to improve the performance. To improve the repeatability and only focus on the evaluation of the fusion step, we use the well-known CaffeNet model [76] to extract features, train, and fine-tune the CNN with default structure and parameter setting.

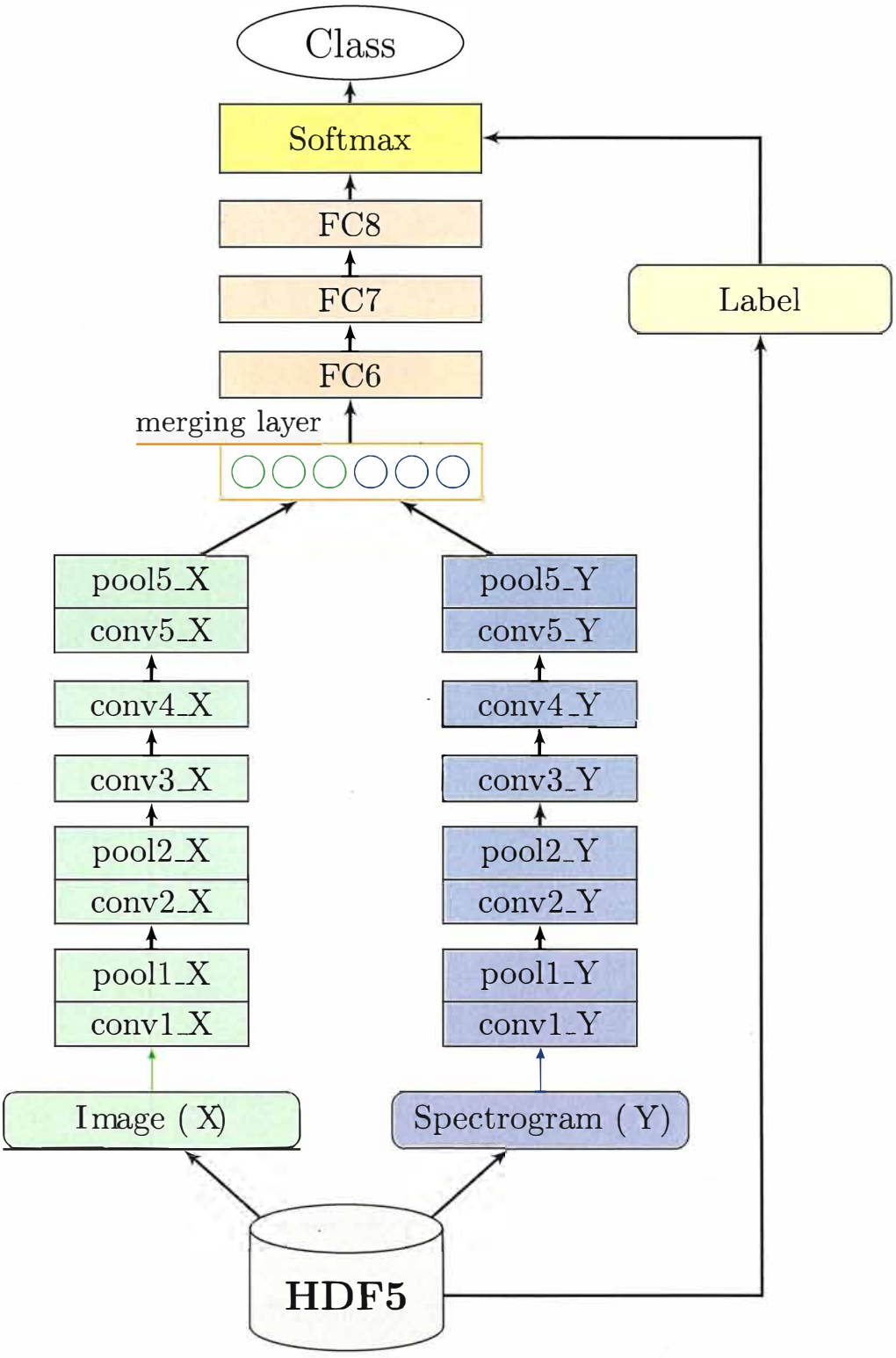


Figure 4.4: The architecture of the middle fusion model (Net2). The activations of the pool5 layers of the two modalities are concatenated at the merging layer and feeding it into the three fully connected layers with softmax at the end.

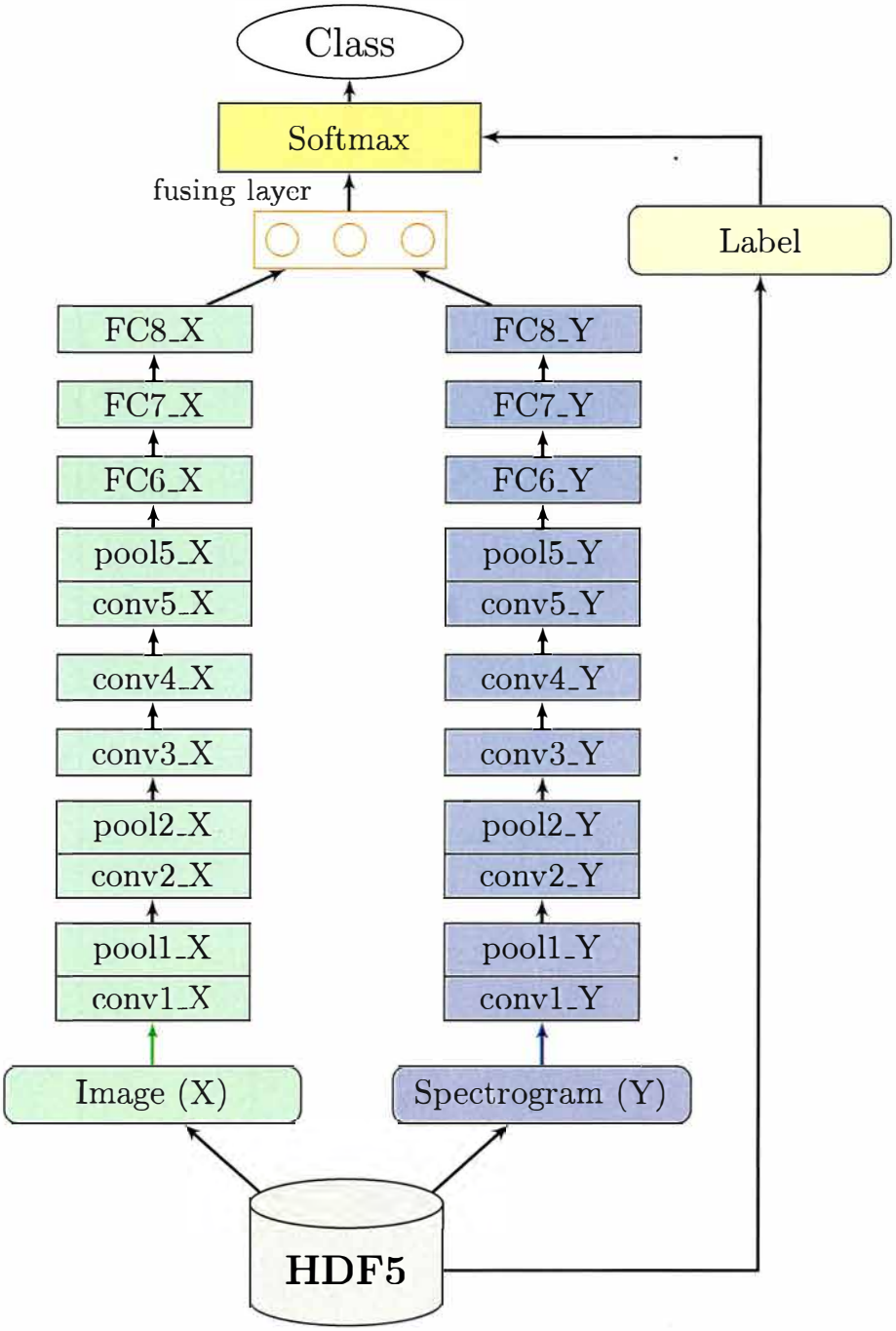


Figure 4.5: The architecture of the late fusion model (Net3). The last fully connected layers of each model hold the unimodal scores for each class, which fused at the fusing layer by summing and multiplying the unimodal scores.

We use stochastic gradient descent to optimize all models. Single modality model and all multimodality models are initialized with learning rates of 0.001 and 0.0001 respectively, and this value is further reduced by hand whenever the test error stops improving. The batch size is set to 32 for single modality and 1 for multimodal learning due to limited resources. We use a momentum of 0.9 and a weight decay of 0.0005.

4.3.1 Quantitative Result: Single Modality v.s. Multimodality Models

The core idea of this work is to address the integration of the image and audio. Therefore, it is necessary to compare the performance between the single modality model (image or sound) and multimodality models (both image and sound: Net1, Net2, Net3). To focus on the evaluation of fusion models, in this experiment, we do not introduce any transfer learning techniques (e.g., fine-tune the pre-trained model to help differentiate between gains from the proposed architectures). Table 4.1 summarizes the results of single modality and proposed models. We can observe that combining two modalities using CNN improves the performance of those only

Table 4.1: Comparative results between individual modality and multimodal CNNs.

Method		Accuracy (%)
Single modality	Image	16.2
	Audio	46.4
Multimodality	Net1	50.0
	Net2	49.9
	Net3 (summation)	**53.8**

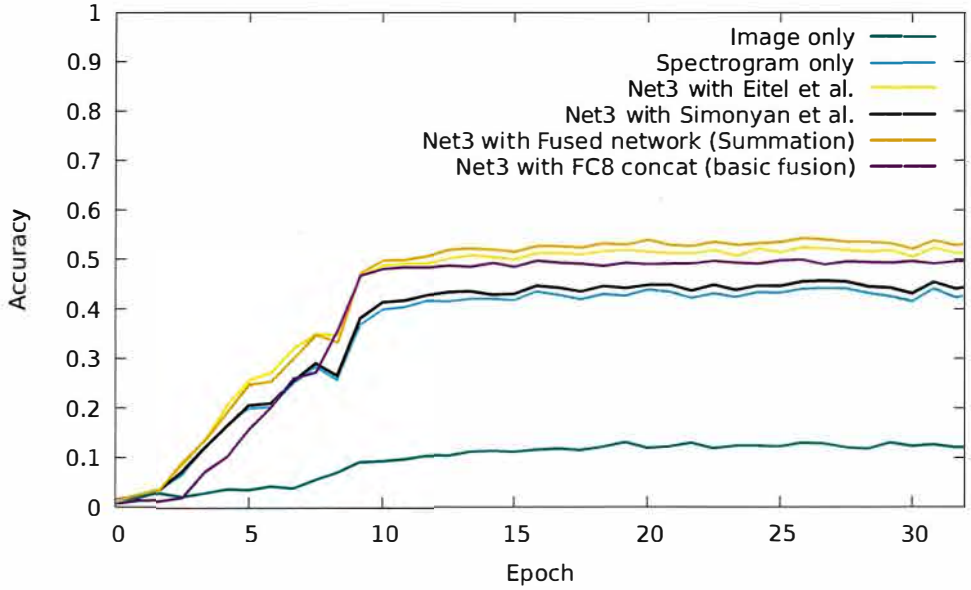


Figure 4.6: Test accuracy vs. Epoch.

using one modality, extracting features separately from image and audio and fusing them at the late stage performs better with significant gain.

Interestingly, the performance of low-level and mid-level fusion models slightly better than the performance of single modality model. One possible reason is that CNN learns features for the predominant modality. In contrast, learning features separately for different modalities results in more independent features, which leads to achieving better performance. Let us mention that the result obtained by multi-modality model is different from simply combining the results of two CNNs trained separately. Indeed, the two modalities' parameters are jointly estimated and thus can be mutually influenced.

Figure 4.6 plots a learning process of single modality (image only and audio only) and Net3 model with different late fusion approaches over learning epochs. It can be observed that the Net3 model with all fusion approaches significantly improve the results. We provide a detailed discussion in the following subsection.

4.3.2 Qualitative Result: Single Modality v.s. Multimodality Models

We perform a qualitative study to analyze the effects of multimodal learning models by comparing single modality and multimodality networks. First, we select some classes where the multimodality models provide the correct answer while the single modality model produces the wrong classification. Then, we study why the multimodality models provide the right answers while the single modality model failed to produce the correct classification.

Figure 4.8 shows some examples of single modality vs. multi modality classification. In the first column, the single modality models predict the input image and spectrogram as the 'barn shallow' and the 'ring-billed gull' respectively rather than the 'red-bellied woodpecker'. However, the multimodality models are able to predict the right answer, because those models provide joint features of different modalities. We observe that when the single modality model provides the right answer for spectrogram, the probability of providing the right answer of multimodality models is higher than when single modality classification is correct for an image. In the second column of Fig. 4.8, single modality model has misclassified the 'clark nutcracker' image as the 'great grey shrike', where other models provide correct answers. Lastly, the Net3 model is able to provide the right answers while the other models provide misclassification on the 'belted kingfisher' (last column).

To better understand the difference between the models, we analyze the feature learned from each network by visualizing the filters of the first convolutional layers shown in Fig. 4.7. We see that each network's filters of different models have

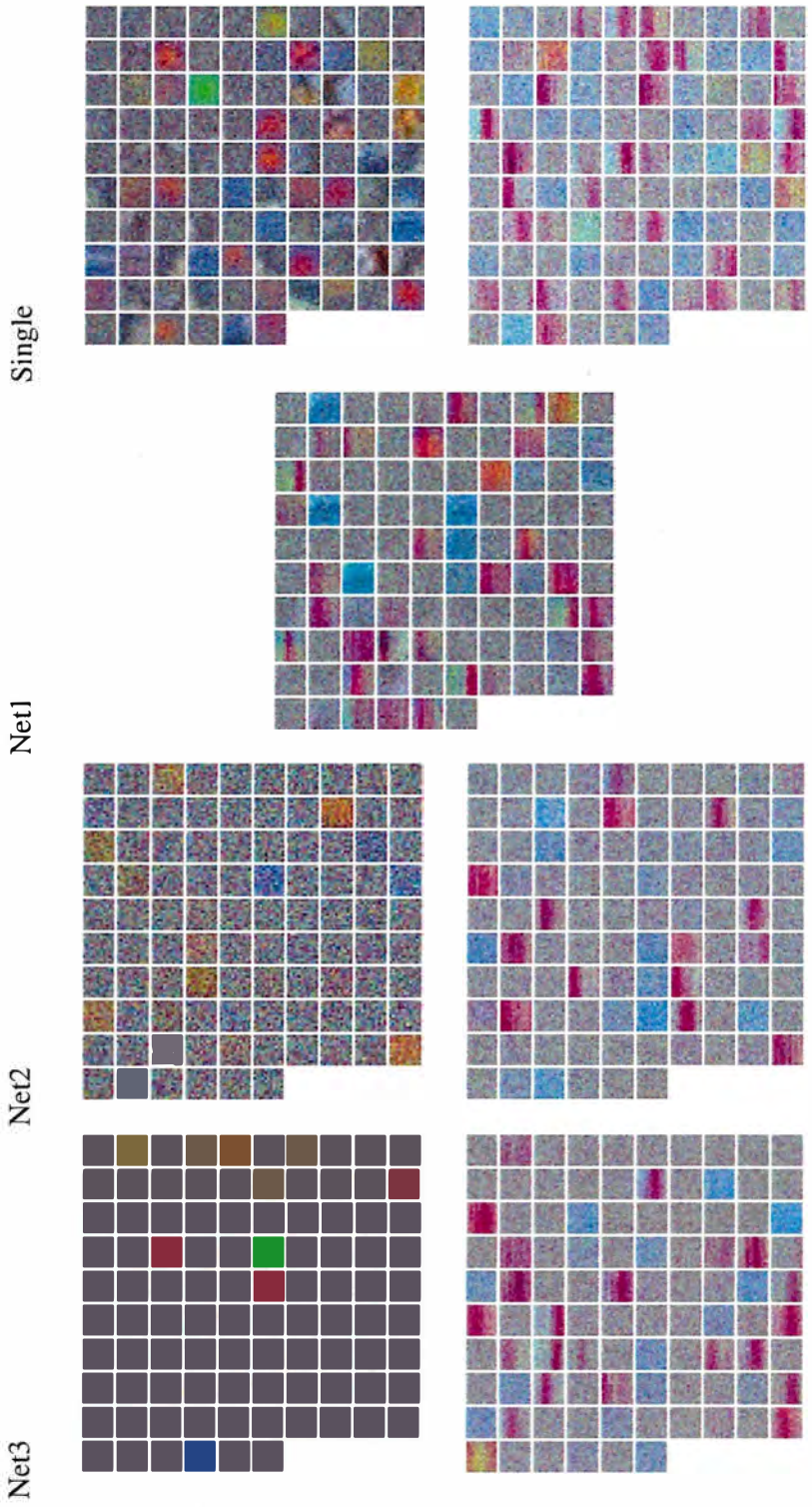


Figure 4.7: Visualization of 96 filters of the first convolutional layer. Left side shows the filters related to the image network, while the right side shows the filters related to the spectrogram network. It can be seen that the filters (left or right side) of different models have a similar pattern. However, the filters of Net1 seems to have mixed filters of both networks.

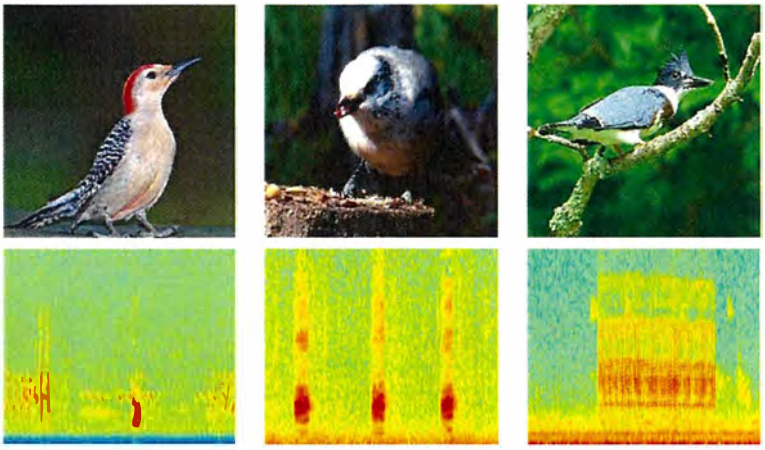
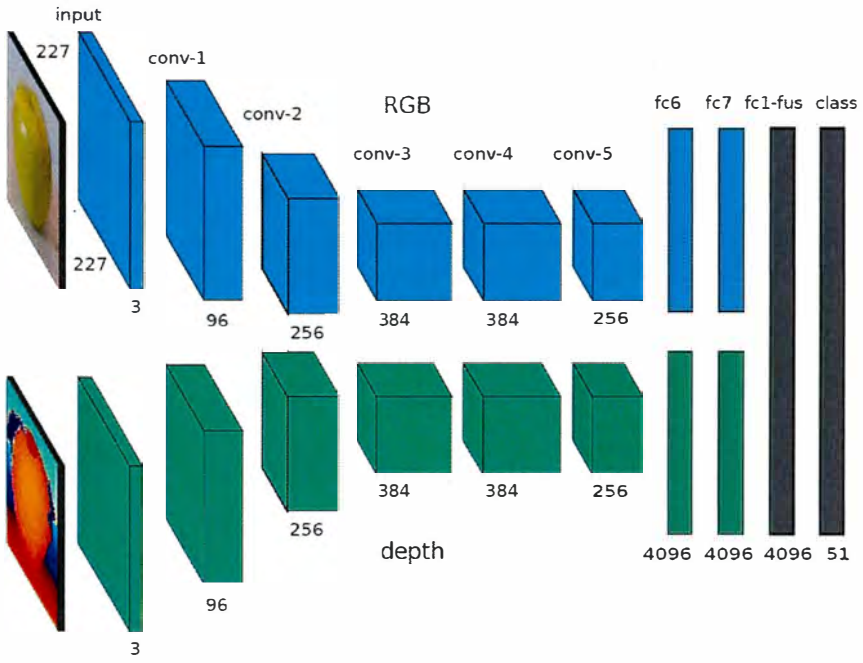


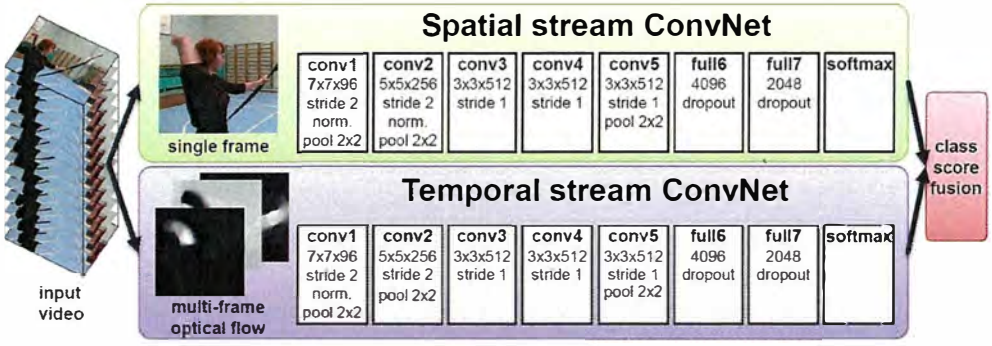
Image	✗	✗	✗
Sound	✗	✓	✗
Net1	✓	✓	✗
Net2	✓	✓	✗
Net3	✓	✓	✓

Figure 4.8: Effects of combining image and spectrogram. Top two rows show sample image and spectrogram of different bird species where are fed into single modality models and multimodality models. The bottom rows show the resulting classification, where multimodal networks provide a correct classification while the classification of single modality model is incorrect.

a similar pattern. Secondly, we see that the single modality model's filters have more meaningful patterns than the multimodality model's filters. As we mentioned before it seems that the learning features separately for different modalities results are more independent features. Finally, it can be seen that the filters of the early fusion model has combined patterns from both networks, but most filters are similar to the spectrogram network. It reveals that the multimodal model (early fusion and mid-level fusion) learns features for the predominant modality.



(a) Two-stream CNN for RGB-D object recognition by Eitel et al. [6].



(b) Two-stream architecture for action recognition method by Simonyan et al. [7]

Figure 4.9: Existing late fusion methods presented in [6, 7]. (a) The input of the network is an RGB and depth image pair. Both streams (blue for RGB image and green for depth image) fused in one fully connected layer (gray) with tensor multiplication. (b) The input video decomposed into spatial and temporal networks, where spatial network inputs video frames (i.e., single image) and temporal network inputs optical flow (i.e., motion across the frames). The softmax scores of two networks are combined by late fusion. The figures taken from [6, 7].

4.3.3 Quantitative Result: Net3 v.s. Existing Late Fusion Methods

To evaluate the effectiveness of our late fusing approach, we conduct comparative experiments on Net3 with a basic fusion method and existing late fusing methods presented in [7, 6] (Figure 4.9). The differences between the late fusion approaches are shown in Fig. 4.10.

- **FC8 concat (basic fusion)**: FC8 layers of each network are concatenated and fed into an additional fully connected layer for final classification. In other words, the FC8 features of each network are fused in a linear combination way (linear weighted fusion).

- **Eitel et al. [6]**: FC7 layers (green and blue) of each network are concatenated and merge into the fusion layer, which performs tensor multiplication of two vectors. The resulting fusion vector is then passed through one additional fully-connected layer for classification. This means this fusion methods is a linear combination of pair-wise interactions between two features. However, this method is not suitable when the features are in different sizes.

- **Simonyan et al. [7]**: Each network focus on learning features from images and spectrograms, respectively, and the final classification is computed as an average of the softmax scores of the two networks. In this fusion method, they do not consider pair-wise interactions between the features. However, this method is suitable when the model consists of different structured network streams.

Table 4.2: Classification performance of fine-tuned Net3 model with different fusion and fine-tuning method.

Method		Accuracy (%)
Fine-tuning weights of pretrained model (summation)		65.0
Two-stage fine-tuning	**Summation (ours)**	**78.9**
	Multiplication (ours)	75.0
	FC8 concat (basic fusion)	75.4
	Simonyan et al. [7]	70.0
	Eitel et al. [6]	72.5

Figure 4.6 plots learning curve of Net3 model with different late fusing methods for each epoch, indicating that averaging the softmax scores gives the lowest performance and our fusing approach performs best.

4.3.4 Fine-Tuning the Pre-Trained Model

Combining multi-modalities at the late stage of CNN has proven to be more effective, thus we apply different transfer learning technique to our last fusion model. Precisely, we conduct additional experiment with fine-tuning CaffeNet pre-trained CNN under Net3 model in this section.

One natural idea for fine-tuning is to train the model by initializing both image and audio CNNs with the weights and biases of the first seven layers derived from CaffeNet pre-trained network (model is trained on the general large scale dataset ImageNet [58]), discarding the last fully connected layer. Instead of last fully connected layer of the pre-trained model, we randomly place the initialized new fully connected layer for 200-class bird classification (in our experiment, 194 classes due to the lack of audio dataset). The experimental result given in Table 4.2.

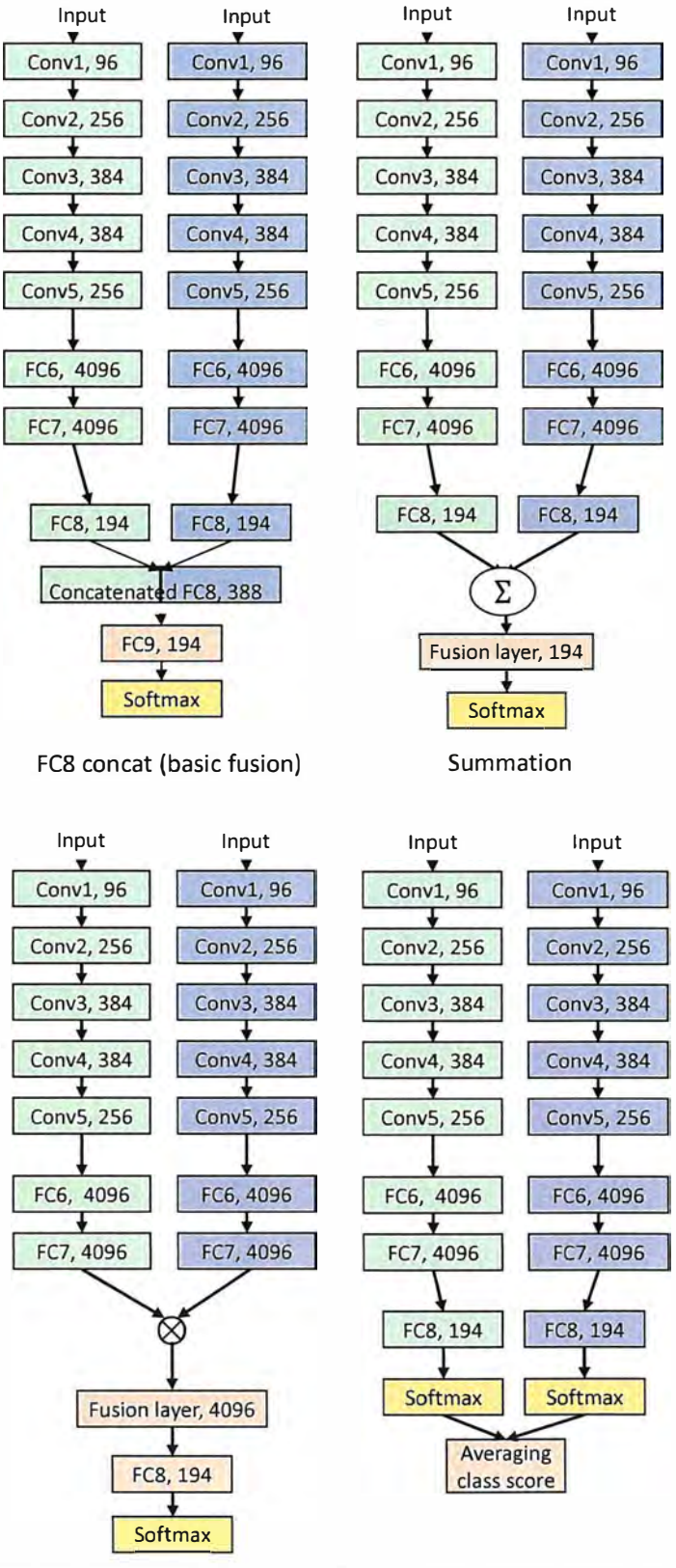


Figure 4.10: Differences between the late fusion approaches.

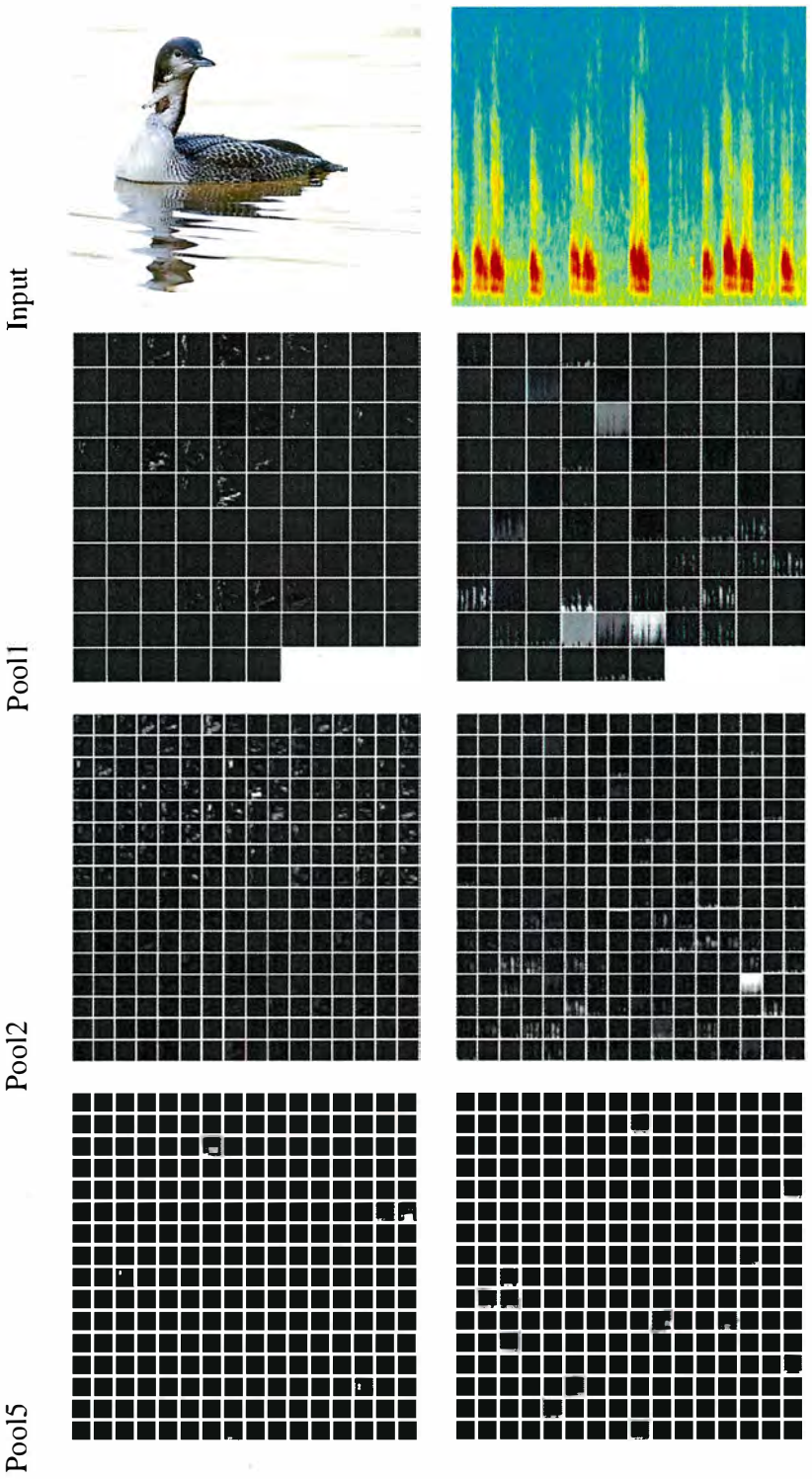


Figure 4.11: Feature visualization of the network layer. These are examples of features of different pooling layer of image (left) and audio (right) network in Net3.

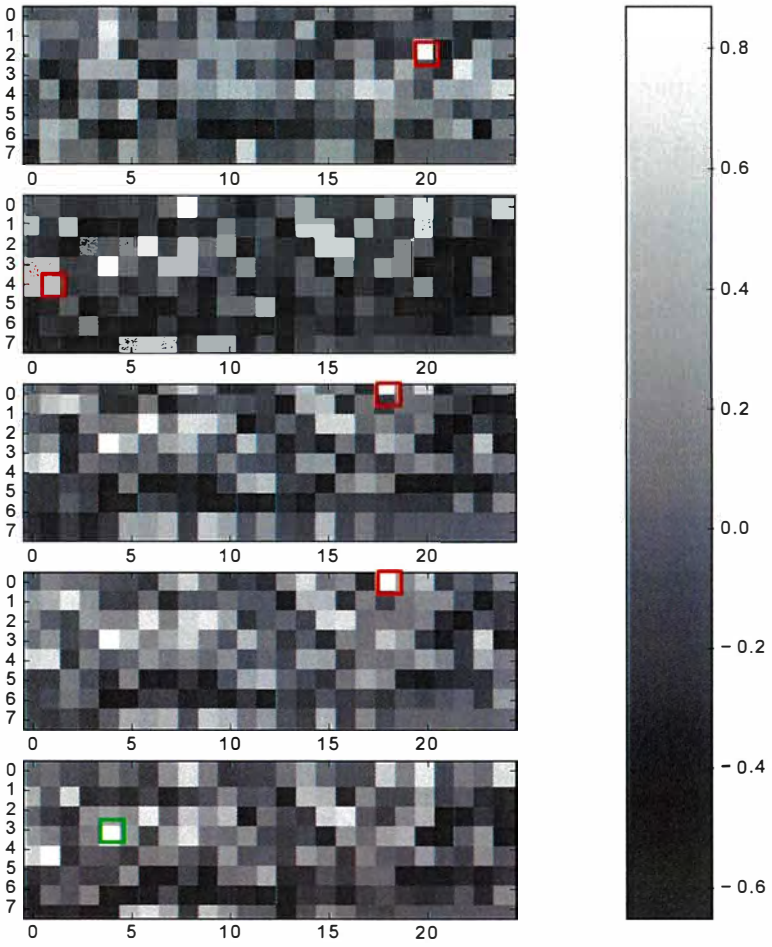


Figure 4.12: Feature visualization of last fully connected layer where top to bottom shows the features of last fully connected layer of image only, spectrogram only, FC8 layer using Eitel et al. [6], FC9 layer using FC8 concat (basic fusion), and fused layer using summation. Here, the red rectangle shows the incorrect answer, the green rectangle shows the correct answer of the classification.

Another method of fine-tuning the model is to train Net3 in two stages. First, training the two stream individually followed by a joint fine-tuning. We train the image and audio CNNs separately, adapting the weights of pre-trained model and learn the weights of the new 194-class output layer. After this training, the networks can be used to perform separate classification with respect to each modality. After then, we train an entire model by setting their learning rate to zero and only training the fusion part of the network to freeze the individual stream networks. As shown in Table 4.2, the transfer learning technique proves it can be improve the performance. The fine-tuned model using pretrained model improves the classification performance by 11.2%, and two-stage training resulted best performance by improving the classification performance with large margin (around 25.1%).

As we can seen on Table 4.2 and Fig. 4.6, averaging the decisions of two networks gives lowest performance compared to other late fusion models. This is because averaging strong modality (i.e., audio) decision and weak modality (i.e., image) decision may degrade the strong modality decision that lead to result in final decision. Furthermore, the basic fusion model and the model with Eitel et al. [6] can capture the possible complementaries among modalities but may lead to conflicts between modalities. On the other hand, in our fusion model, every modality is always potentially useful which leads to making classification collectively.

We also found that the fine-tuning pre-trained model to two modalities and training them simultaneously is significant worse than two-stage training. We think the problem relates to the difference between the size of two datasets and batch size of single modality and multimodality models. In this experiment, we used net surgery[1]

[1] https://github.com/BVLC/caffe/blob/master/examples/net_

to fine-tune two different CNN using one pre-trained model and to fine-tune two pre-trained model into our model.

To confirm that the features extracted from the sample image and spectrogram during the fine-tuning were meaningful, we visualized the activations of different layers, especially the fused layers in Net3. The results are shown in Fig.4.11 and 4.12, that allowed us to confirm the learned features were meaningful and qualitatively resembled the sample image and spectrogram. Moreover, the activations of our fused layer takes advantage by incorporating the features from each stream, when each network failed to produce write answer.

4.4 Summary

In this chapter, we introduced three multimodal CNN architectures in different fusion strategies, which can process jointly the image and audio data for bird classification. Experimental results verified that the two-stream multimodal CNN in late fusion strategy outperforms the others.

In addition, we proposed the summed fusion method to combine multiple CNNs, which shows better performance comparing against several existing fusion methods. Moreover, with the help of two-stage fine-tuning, our method can be more effective.

However, there still exist several drawbacks of our method: (1) Choosing the suitable duration based on the vocal features of birds to be recognized, is essential ingredients of improvement, is missed in our current work. (2) Our method is based on the raw image data, thus part detection and extracting features from

`surgery.ipynb`

pose-normalized regions may improve the classification performance.

Chapter 5

Conclusion

5.1 Overall summary of the current study

This dissertation investigated study on multimodal fusion strategies for fine-grained bird classification with audio-image data utilizing deep neural networks. We proposed that the combination of image and sound provide richer and substantial training signal for bird species classification under CNN framework, which is the first attempt to the best of our knowledge.

Therefore, I restate the research aims and objectives and summarize the main findings and evidence in this study. The main objective of this research was to study the efficient and effective multimodal fusion strategy for fine-grained bird classification by integrating audio and visual data through utilizing deep learning architecture. We summarize the main findings considering main research considerations:

- *What to fuse.* The sound dataset has collected and created corresponding to the bird image dataset. In this study, CNN has been used to learn represen-

tations directly from the raw data and extract a set of discriminative features from audio and visual modality.

- *When to fuse or Level of fusion.* This dissertation aimed to identify effective fusion strategies for fine-grained classification with audio and visual data. In Chapter 3, kernel-based fusion methods has been considered to combine CNN features from both modalities using multiple kernel learning. Experimental results indicate that MKL is an effective approach to improve classification performance while fusing different features. We considered three fusion strategies including early fusion, middle fusion, late fusion in Chapter 4. Based on a quantitative and qualitative analysis of bird species classification, it can be concluded that the two-stream multimodal CNN in late fusion strategy outperforms the fusion strategies.

- *How to fuse.* In Chapter 3, l_p-norm MKL has been utilized to fuse CNN features at kernel-level, which performed better accuracy compared to some simple kernel combination methods, and the conventional early fusion method. The summed fusion method has been proposed to combine multiple CNNs which shown better performance compared to several existing late fusion methods in Chapter 4. Moreover, with the help of two-stage fine-tuning, the proposed method could be more effective.

5.2 Future work

Multimodal research especially multimodal deep learning is an intense multi-disciplinary field of increasing importance and with extraordinary potential. According to the

[8], authors identified core challenges that are faced by multimodal deep learning, namely:

- Representation: Learning how to represent and summarize multimodal data in a way that exploits the complementarity and redundancy.
- Alignment: Identify the direct relations between (sub) elements from two or more different modalities.
- Fusion: To join information from two or more modalities to perform a prediction task.
- Translation: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.
- Co-learning: Transfer knowledge between modalities, including their representations and predictive models.

Consequently, we will focus on representation and translation (cross-model retrieval) problems for multimodal deep learning. We also interested in the feature selection method to obtain key features from multimodal data.

Publications

Journal paper

- Bold, Naranchimeg, Chao Zhang, and Takuya Akashi. "Cross-Domain Deep Feature Combination for Bird Species Classification with Audio-Visual Data." IEICE TRANSACTIONS on Information and Systems 102.10 (2019): 2033-2042.

International conference paper

- Bold Naranchimeg, Chao Zhang, Takuya Akashi, "Bird Species Classification with Audio-Visual Data using CNN and Multiple Kernel Learning", Proceedings of the International Conference on Cyberworlds (CW), pp. 85-88, 2019.

Other journal paper

- Bold Naranchimeg*, Chao Zhang* (*Equal Contribution), Takuya Akashi, "3D Point Cloud Retrieval with Bidirectional Feature Match", IEEE Access 7 (2019): 164194-164202.

Bibliography

[1] S. Branson, G. Van Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," *arXiv preprint arXiv:1406.2952*, 2014.

[2] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, "Deep learning for smart manufacturing: Methods and applications," *Journal of Manufacturing Systems*, vol. 48, pp. 144–156, 2018.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *arXiv preprint arXiv:1506.06579*, 2015.

[5] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[6] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on.* IEEE, 2015, pp. 681–687.

[7] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[8] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[9] E. Tatulli and T. Hueber, "Feature extraction using multimodal convolutional neural networks for visual speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on.* IEEE, 2017, pp. 2971–2975.

[10] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[11] H. Meutzner, N. Ma, R. Nickel, C. Schymura, and D. Kolossa, "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on.* IEEE, 2017, pp. 5320–5324.

[12] A. Torfi, S. M. Iranmanesh, N. M. Nasrabadi, and J. Dawson, "Coupled 3d convolutional neural networks for audio-visual recognition," *arXiv preprint*, 2017.

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[14] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7596–7599.

[15] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1447–1454.

[16] ——, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729.

[17] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, vol. 2, no. 1, 2011.

[18] L. Yang, P. Luo, C. Change Loy, and X. Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3973–3981.

[19] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[20] T. L. Berg, A. C. Berg, and J. Shih, "Automatic attribute discovery and characterization from noisy web data," in *European Conference on Computer Vision*. Springer, 2010, pp. 663–676.

[21] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[22] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 801–816.

[23] S. Kahl, T. Wilhelm-Stein, H. Hussein, H. Klinck, D. Kowerko, M. Ritter, and M. Eibl, "Large-scale bird sound classification using convolutional neural networks," *Working notes of CLEF*, 2017.

[24] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 1744–1748.

[25] N. Takahashi, M. Gygli, and L. Van Gool, "Aenet: Learning deep audio features for video analysis," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 513–524, 2018.

[26] K. J. Piczak, "Recognizing bird species in audio recordings using deep convolutional neural networks." in *CLEF (Working Notes)*, 2016, pp. 534–543.

[27] W. S. Noble *et al.*, "Support vector machine applications in computational biology," *Kernel methods in computational biology*, vol. 71, p. 92, 2004.

[28] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California Institute of Technology, Tech. Rep. CNS-TR-2011-001, 2011.

[29] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of machine learning research*, vol. 12, no. Jul, pp. 2211–2268, 2011.

[30] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, no. Jul, pp. 1531–1565, 2006.

[31] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[32] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4094–4102.

[33] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1445–1454.

[34] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[35] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1153–1162.

[36] E. Gavves, B. Fernando, C. G. Snoek, A. W. Smeulders, and T. Tuytelaars, "Local alignments for fine-grained categorization," *International Journal of Computer Vision*, vol. 111, no. 2, pp. 191–212, 2015.

[37] P. Guo and R. Farrell, "Fine-grained visual categorization using pairs: Pose and appearance integration for recognizing subcategories," *arXiv preprint arXiv:1801.09057*, 2018.

[38] V. Lebedev, A. Babenko, and V. Lempitsky, "Impostor networks for fast fine-grained recognition," *arXiv preprint arXiv:1806.05217*, 2018.

[39] X. He and Y. Peng, "Fine-grained image classification via combining vision and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5994–6002.

[40] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5546–5555.

[41] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4438–4446.

[42] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh, "Deep learning vs. traditional computer vision," in *Science and Information Conference*. Springer, 2019, pp. 128–144.

[43] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[44] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[45] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[46] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8595–8598.

[47] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel, "Handwritten digit recognition with a back-propagation

network," in *Advances in neural information processing systems*, 1990, pp. 396–404.

[48] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[49] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[50] Y. Petetin, C. Laroche, and A. Mayoue, "Deep neural networks for audio scene recognition." in *EUSIPCO*, 2015, pp. 125–129.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[53] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[55] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on*. IEEE, 2013, pp. 8614–8618.

[56] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[57] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 26, 2015.

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[59] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[60] H. Bourlard and S. Dupont, "A mew asr approach based on independent processing and recombination of partial frequency bands," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 1. IEEE, 1996, pp. 426--429.

[61] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 2007, pp. 375–388.

[62] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2623–2631.

[63] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[64] M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg, "Non-sparse multiple kernel learning," in *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, vol. 4, 2008, p. 5.

[65] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien, "Lp-norm multiple kernel learning," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 953–997, 2011.

[66] W. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 2, p. 987184, 2003.

[67] H. Wu and L. He, "Combining visual and textual features for medical image modality classification with lp- norm multiple kernel learning," *Neurocomputing*, vol. 147, pp. 387–394, 2015.

[68] Y.-R. Yeh, T.-C. Lin, Y.-Y. Chung, and Y.-C. F. Wang, "A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection," *IEEE Transactions on multimedia*, vol. 14, no. 3, pp. 563–574, 2012.

[69] S. S. Bucak, R. Jin, and A. K. Jain, "Multiple kernel learning for visual object recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1354–1369, 2013.

[70] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2539–2544.

[71] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proceedings of the 15th ACM on International conference on multimodal interaction.* ACM, 2013, pp. 517–524.

[72] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Emotion recognition in the wild with feature fusion and multiple kernel learning," in *Proceedings of the 16th International Conference on Multimodal Interaction.* ACM, 2014, pp. 508–513.

[73] F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple kernel learning in the primal for multimodal alzheimer's disease classification," *IEEE journal of biomedical and health informatics*, vol. 18, no. 3, pp. 984–990, 2013.

[74] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*. Springer, 2005, pp. 878–887.

[75] A. Chowdhury and J. Alspector, "Data duplication: an imbalance problem?" in *ICML'2003 Workshop on Learning from Imbalanced Data Sets (II), Washington, DC*, 2003.

[76] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675--678.

[77] A. Jain, S. V. Vishwanathan, and M. Varma, "Spf-gmkl: generalized multiple kernel learning with a million kernels," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 750–758.