ACTIVITY PREDICTION FROM AUTO-CAPTURED LIFELOG IMAGES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

KADER BELLI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JULY 2019

Approval of the thesis:

ACTIVITY PREDICTION FROM AUTO-CAPTURED LIFELOG IMAGES

submitted by **KADER BELLI** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Assist. Prof. Dr. Emre Akbaş
Supervisor, **Computer Engineering, METU** _____

Prof. Dr. Adnan Yazıcı
Co-supervisor, **Computer Science, Nazarbayev University** _____

Examining Committee Members:

Assoc. Prof. Dr. Sinan Kalkan
Computer Engineering, METU _____

Assist. Prof. Dr. Emre Akbaş
Computer Engineering, METU _____

Prof. Dr. Adnan Yazıcı
Computer Science, Nazarbayev University _____

Assoc. Prof. Dr. Murat Koyuncu
Information Systems Engineering, Atılım University _____

Assist. Prof. Dr. Gökberk Cinbiş
Computer Engineering, METU _____

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Kader Belli

Signature :

ABSTRACT

ACTIVITY PREDICTION FROM AUTO-CAPTURED LIFELOG IMAGES

Belli, Kader
M.S., Department of Computer Engineering
Supervisor: Assist. Prof. Dr. Emre Akbaş
Co-Supervisor: Prof. Dr. Adnan Yazıcı

July 2019, 58 pages

The analysis of lifelogging has generated great interest among data scientists because large-scale, multidimensional and multimodal data are generated as a result of lifelogging activities. In this study, we use the NTCIR Lifelog dataset where daily lives of two users are monitored for a total of 90 days, and archived as a set of minute-based records consisting of details like semantic location, body measurements, listening history, and user activity. In addition, images which are captured automatically by cameras located at users' chests are available for each minute together with text annotations, which promotes the multimodal nature of the dataset. We train and evaluate several classification methods on the text and image data separately, and on their combination as well. Specifically, for text data, we encode the words using a one-hot encoding, and train SVM and MLP models on bag-of-words representations of minutes. For image data, we train two different convolutional neural networks (CNN) in two different ways: training from scratch and fine-tuning an ImageNet [1] pre-trained model. Finally, we propose a multi-loss, combined CNN-MLP model which processes image and text data simultaneously, uses fusion methods to merge the two sub-models, and can handle missing input modalities. We also put effort into a con-

tribution to the NTCIR LifeLog dataset by manually labeling 90,000 images into 16 activity classes.

Keywords: lifelog, multimodal classification, machine learning, deep learning

ÖZ

OTOMATİK YAKALANMIŞ HAYAT GÜNLÜĞÜ GÖRÜNTÜLERİNDEN FAALİYET TAHMİNİ

Belli, Kader
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Dr. Öğr. Üyesi. Emre Akbaş
Ortak Tez Yöneticisi: Prof. Dr. Adnan Yazıcı

Temmuz 2019 , 58 sayfa

Hayat günlüğü analizi, veri bilimcilerinin büyük ölçüde ilgisini çeken bir konu başlığı haline gelmiştir, çünkü hayat günlüğü faaliyetleri sonucunda büyük, çok boyutlu ve çok modlu veriler üretilmektedir. Bu çalışmada, iki kullanıcının günlük yaşamlarının toplamda 90 gün boyunca izlendiği ve konum, vücut ölçümleri, müzik dinleme geçmişi ve kullanıcı faaliyeti gibi ayrıntılardan oluşan dakika bazlı kayıtlar halinde arşivlendiği NTCIR Lifelog veri setini kullanıyoruz. Ayrıca her dakika için, kullanıcıların göğüs hizasına yerleştirilmiş kameralar tarafından otomatik olarak çekilen görüntüler, metin açıklamaları ile birlikte veri setinin çok modlu yapısını destekleyecek şekilde verilmektedir. Bu çalışmada, çeşitli sınıflandırma yöntemlerini metin ve resim verileri ile bunların kombinasyonları üzerinde eğiterek, öğrenme performanslarını değerlendirdik. Metin verileri için, kelimeleri tek boyutlu vektörler halinde düzenledik; SVM ve MLP modellerini bu vektörler üzerinde eğittik. Görüntü verileri için ise, iki farklı evrişimsel sinir ağı (CNN) mimarisini iki farklı şekilde eğittik: sıfırdan eğitme ve ImageNet [1] veri seti üzerinde önceden eğitilmiş mimariye hassas ayar yapma. Son olarak, görüntü ve metin verilerini aynı anda işleyen, iki alt modeli birleştirmek için

füzyon yöntemlerini kullanan ve eksik verileri telafi edebilen birleşik bir CNN-MLP modeli önerdik. Ayrıca, 90.000 görüntüyü 16 faaliyet sınıfı ile etiketleyerek NTCIR LifeLog veri setine katkı sağladık.

Anahtar Kelimeler: hayat günlüğü, çok modlu sınıflandırma, makine öğrenmesi, derin öğrenme

Hayatın kıyılarında delice esip duran ‘Rüzgâr’a...

ACKNOWLEDGMENTS

I would like to express my grateful thanks to my lovely family; my father, my mother and dear sisters Elif and Sema, who never stop pushing me to successfully end my studies. I admit that this thesis could not be finalized without their endless support.

I feel my greatest gratitude to my supervisors Prof. Dr. Adnan Yazıcı and Assist. Prof. Dr. Emre Akbaş, who never leave me alone on this extended journey. I would definitely be lost without their guidance! I also thank Dr. Frank Hopfgartner and his team for sharing the dataset with us, which has been the starting point for our journey.

I am grateful to my friends, who are a small group of warm-hearted people, for motivating me every time I feel exhausted and upset.

And to my brand-new employer, proud to say, Google, for being my biggest motivation at the last stages of this study.

Last but most importantly, I warmly thank "Elveda Rumeli" (Farewell Rumelia), my all-time miracle, and the dream team who created and brought it to us. I always agree that Elveda Rumeli has been the best thing that happened to me in my life! It has been the secret key to all my success and happiness since we met in 2008.
I am grateful for the strong personality and great memories I had, and wonderful people I met thanks to Elveda Rumeli. Specifically, I am thankful for meeting fantastic musician Erdal Güney, with the peaceful voice of whom I keep calm and relaxed.
And finally İrşad Aydın, who has been the most special gift given to me by Elveda Rumeli, and who has been my secret motivation for 10 years, with his guidance, brotherhood, friendship and many more.. Thank you, dear friend, for touching my life from miles away, and making me a better person, making me who I am today.
Vielen Dank für alles!

Kader Belli

July 2019, Ankara

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
ICAISC	International Conference on Artificial Intelligence and Soft Computing
IEEE	Institute of Electrical and Electronics Engineers
MLP	Multilayer Perceptron
MSVM	Multiple-SVM
NTCIR	NII Testbeds and Community for Information Access Research
RBF	Radial Basis Function
ResNet	Residual Neural Network
SVM	Support Vector Machine
VC-Dimension	Vapnik-Chervonenkis Dimension

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

Advances in computer miniaturization and the investigation of wearable devices have led to the rise of the concept of lifelogging, which has been an active research topic in various domains since the 1970s.

Lifelogging enables people to create digital archives of their daily lives, which can include, but not limited to, body measurements, first-person imagery, location tracking and the history of various activities such as listening, reading, watching, browsing, etc. Research has been conducted on the design of user-friendly hardware and software components for lifeloggers. In addition, the concept of lifelogging is of great interest to data scientists because large-scale, multidimensional and multimodal data are generated as a result of lifelogging activities [2, 3].

The main objective of lifelogging research is to monitor and provide insights into users' daily lives by understanding the relationships between user activities and different dimensions of the collected data (See Figure 1.1). The progress starts with the description of lifelog records, which is achieved by classification of archived data into activity classes. Since lifelog data have various dimensions, it is possible to approach the classification problem from different perspectives, which increases the complexity of the problem, significantly.

Lifelog records are bare figures with a set of image attachments before they are described by activity definitions. Activity prediction comes into the picture to interpret these figures and images, and to draw the baseline for lifelogging research by explain-

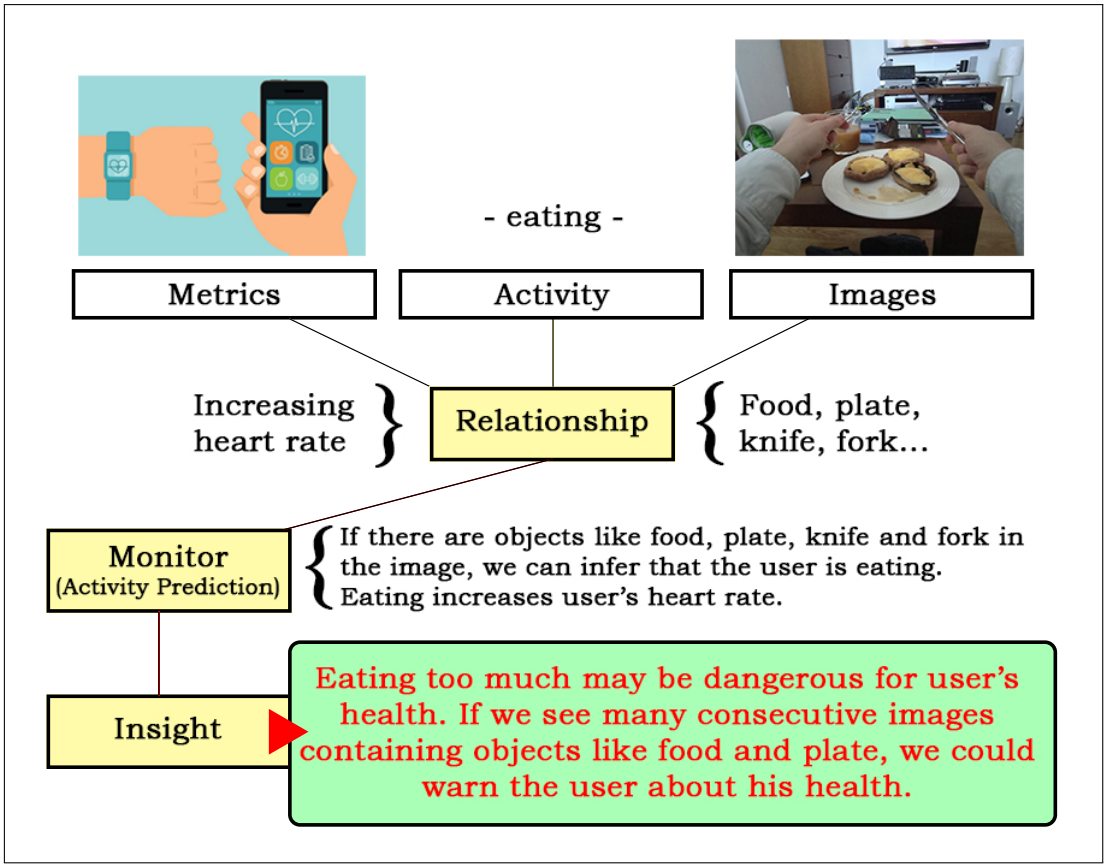

Figure 1.1: Lifelogging Research

ing the relationship between activities and different dimensions of the data.

The extent of this study could be interpreted as a contribution to the lifelogging research baseline. In this study, the main objective is to apply different classification algorithms to different modalities of the lifelog data and make a comparison between their performance in predicting user activities. Next, we propose a multi-loss, combined learning model to process image and text data simultaneously, and merge findings from the two modalities using fusion methods. The proposed model is able to handle missing input modalities thanks to the custom loss function we used for training the model, and it shows better classification performance than the naive activity prediction method in the presence of missing values.

1.2 The NTCIR Lifelog Dataset

The NTCIR Lifelog dataset was published for the first time by Gurrin et al. in [4] in 2016. The first test collection consists of lifelog data from three users and includes images which are captured automatically from a wearable camera, location, and activity definition, which are recorded on a minute-by-minute basis for one month per user. Lifeloggers gathered data in an all-day gathering process, resulting in a wide range of daily activities. The collected data is presented as an XML document, which is also designed in a minute-based structure, and a folder containing image files, which are referenced by their relative file paths in the XML document [4].

The version analyzed in this paper is called the NTCIR-13 Full Phase-2 Lifelog-2 dataset. This version contains 90 days of lifelog data, generated by activities of two users; i.e. 60 days of data from the 1st user and 30 days of data from the 2nd user. In addition to the features available in the first test collection, the Full Phase-2 Lifelog-2 dataset contains minute-based biometric data generated by a smartwatch, which lifeloggers are expected to wear during the day. This version also includes daily health logs (blood pressure, cholesterol, weight, etc.), food and beverage logs, as well as users' listening history for a couple of minutes [5].

Attached to the XML document, visual concept annotations for images are provided as pairs of (image ID, concept) in a CSV file. An example image from the dataset, together with visual annotations and the corresponding lifelog record are available in Figures 1.2 and 1.3, respectively.

1.3 NTCIR Lifelog Tasks

The NTCIR Lifelog dataset, which was first published at the 12th NTCIR conference in 2016, was the first test collection for lifelogging research, and introduced with two initial sub-tasks:

- Lifelog Semantic Access Task (LSAT): Retrieve specific moments from life loggers' archive according to given query sentences.

Figure 1.2: An example image and corresponding annotations from the NTCIR Lifelog Dataset

```
<minute id="441">
    <location>
        <name>"Place in Northside, Dublin"</name>
        <latitude>53.350982</latitude>
        <longitude>-6.1623426</longitude>
    </location>
    <activity>walking</activity>
    <bodymetrics>
        <calories>3.1</calories>
        <heart-rate>101</heart-rate>
        <skin-temp>81.5</skin-temp>
        <steps>97</steps>
    </bodymetrics>
    <images>
        <image>
            <image-id>u1_2016-08-16_062145_1</image-id>
            <image-path>u1/2016-08-16/20160816_062145_000.jpg</image-path>
        </image>
    </images>
    <music>
        <song>Runaway Train</song>
        <song-mbid>bb7df441-35b2-4c0d-99b5-6560a5bbbc51</song-mbid>
        <artist>Soul Asylum</artist>
        <artist-mbid>b10db9ad-b4c3-47f3-a7a4-37864b134f65</artist-mbid>
        <album>Black Gold: The Best Of Soul Asylum</album>
        <album-mbid>4f2ff67a-d196-48a6-ba0a-bff6724b94ec</album-mbid>
    </music>
</minute>
```


Figure 1.3: Minute description for the image in Figure 1.2

Example 1.3.1. Find the moments when the user was watching television.

- Lifelog Insight Task (LIT): Generate effective visualizations and insights from life loggers' everyday life.

  Example 1.3.2. Provide insights on how diet affects life loggers' blood sugar level [4].

The extended Full Phase-2 Lifelog-2 dataset was introduced at the 13th NTCIR conference in 2017, together with two additional sub-task, the LSAT and LIT remaining unchanged:

- Lifelog Event Segmentation Task (LEST): Develop approaches to event segmentation from continuous activities.

  Example 1.3.3. Moments when the user was preparing meals at any location are segmented into the "cooking" event.

- Lifelog Annotation Task (LAT): Develop approaches to annotate activity and visual environment of the user at any moment.

Example 1.3.4. At minute id 441, the user was walking, and the context of the image taken at that minute contained ocean, shore, sky, and water. [5]

Recently, in the 14th NTCIR conference which was held in June 2019, the organizers introduce a new task as a substitute for the LEST and LAT tasks:

- Lifelog Activity Detection Task (LADT): Develop approaches to the annotation of multimodal lifelog data in terms of activities of daily living.

In order to reduce the level of subjectivity, the following 16 predefined activity classes are provided in the description of the LADT task:

- Cooking: Preparing meals or making tea or coffee at any location
- Creative Activities: Creative endeavors, e.g. writing, art, music, etc.
- Eating: Eating meals in any location, excluding moments when drinking alone
- F2F Interacting: Face-to-face interaction with people at home or in the workplace, excluding social interactions
- Gaming: Playing computer games
- Houseworking: Working in the home, e.g. cleaning, gardening, etc.
- Physical Activities: Physical activities and sports, e.g. walking, playing sports, cycling, rowing, etc.
- Praying: Praying, worshipping or meditating
- Reading: Reading any form of paper
- Relaxing: Relaxing at home, e.g. watching TV, having a drink, etc.
- Shopping: Shopping in a physical shop, i.e. not online
- Socialising: Socialising outside the home or office
- Time with Children: Taking care of children or playing with children
- Traveling: Traveling by car, bus, boat, airplane, train, etc.

- Using A Computer: Using a desktop computer, laptop or tablet
- Other Activities: Any other activity not represented by the fifteen labels above

The work we present in this thesis was not first designed to provide a complete solution to any of the tasks listed above; however, our approach could be viewed as an initial perspective for LSAT and LAT tasks. In addition, together with the newly introduced LADT task, we could propose our methodology as a solution to this task. In fact, we build our experimentation on activity prediction from lifelog images over the 16 classes provided in the definition of the LADT task.

1.4 Proposed Methods and Models

In this paper, the NTCIR Lifelog dataset, which is described in detail in Section 1.2, is analyzed and processed for lifelogging research. Among many different features of the archived data, minute-based images captured from the first-person perspective are used to classify lifelog records, which are minute descriptions, into activity classes. In order to classify minutes into activity classes, images are processed in three different ways:

1. Using image annotations with text-based classification algorithms
2. Using plain color images with image classification algorithms
3. Using color images together with their annotations, which is performed using a combined classification algorithm

Thus, analysis on the NTCIR Lifelog dataset is performed using machine learning and deep learning methodologies. More specifically, Support Vector Machines (SVM) and Multilayer Perceptrons (MLP) are used to perform text-based classification, whereas deep neural networks, i.e. custom Convolutional Neural Network (CNN) models, are used for image classification.

Next, a multi-loss, combined learning model is proposed to be able to make use of images together with their annotations on a single learning model. This combined

model is trained using a masked loss function so that it could compensate for missing data, which is frequently observed in the NTCIR Lifelog dataset.

The classification performance of the multi-loss combined model is evaluated in comparison with the naive method for predicting activities in the presence of missing input modalities. In the naive method, text-based, image-based and combined learning models are trained separately and the suitable model for activity prediction is determined during test time according to the nature of the individual test record.

The proposed multi-loss combined learning model shows better prediction performance than the naive method both in the absence and presence of the missing data.

1.5 Contributions and Novelties

Our contributions are as follows:

- The original version of the NTCIR Lifelog dataset has 5 activity classes, namely airplane, cycling, running, transportation and walking. With the intuition that these classes are not capable of describing the whole life of lifeloggers, we extend the classification over 16 activity classes by manually classifying a subset of approximately $90\,000$ images taken from the dataset, which requires an effort of approximately 200 person-hours.
- As the dataset contains both textual and visual data, we are able to compare the performance of text-based and image-based classification algorithms on the same set of data.
- We propose a combined learning model, which could learn from text and image data together on a single model.
- We propose a masked loss function, which enables the combined model to continue learning in the presence of missing values.
- We published a part of this thesis as a conference paper titled "Activity Learning from Lifelogging Images" in the International Conference on Artificial Intelligence and Soft Computing (ICAISC) 2019 [6].

1.6 The Outline of the Thesis

This thesis is organized as follows. In the following chapter (Chapter 2), we discuss some of the recent literature on the subject of learning from lifelog datasets, in particular from the NTCIR Lifelog dataset. In Chapter 3, we present the specific methods and models we employ for processing and learning activities from the lifelog data. In Chapter 4, we present our methodology and the results of our experimentation with different classification algorithms, i.e. we compare their performance in terms of classification accuracy. In this chapter, we explain the training and test phases on the dataset after we go through data analysis and preprocessing steps. Finally, in the last chapter (Chapter 5), we present our conclusions.

CHAPTER 2

RELATED STUDIES

2.1 Recent Literature on the NTCIR Lifelog Task

From the day the NTCIR Lifelog dataset was introduced, researchers have focused more on the LSAT, among the tasks listed in Section 1.3.

In 2016, Xia et al. published their research on the integration of location information into images to improve the accuracy of segmentation. They state their finding that the location is an important component in the information retrieval process, and they use artificial neural networks together with a custom ranking function to learn from locations and visual concepts of images. They report that the proposed approach performs well in simple LSAT queries; however, a more complex architecture would be necessary for more complicated scenarios [7].

Safadi et al. offers a framework that uses CNNs to index images, and then pretrained Multiple-SVMs (MSVM) to assign classes to images. In their study, they consider both visual and temporal concepts of the data. They extract visual concepts annotations using well-known network models. Next, they index images according to time, location and activity information. The proposed method appears to give promising results, according to their two-level evaluation criteria [8].

Lin, H. et al. propose a method that uses a deep learning toolkit that allows them to apply several modern deep learning algorithms to the dataset and calculate the correlation between images and classes. They make an effort to find the relevance between images and semantic content by using natural language processing (NLP) tools. They report that image recognition methods with more complex models should

be employed for better performance in learning activities [9].

In 2017, Lin, J. et al. design a framework using CNNs to query images and minutes of life loggers. They make an effort to find a solution to close the gap between images and event-level query topics specified by the organizers. They use CNNs for feature extraction from images; then, build a matching between features and relevant events. Together with feature selection and temporal smoothing methods, they obtain considerably good performance in classification. [10].

Recently, Yamamoto et al. introduce a common approach to solve the three tasks of the NTCIR Lifelog research: LAT, LSAT, and LEST. They analyze both images and locations using visual indexing and location indexing methods. They describe their approach to using the proposed methodology commonly for the three tasks. Query processing, relevance score calculation, and temporal smoothing are some of the methods they employ for the study, in which they demonstrate high performance, and clarify the effectiveness and limitations of their approach [11].

2.2 Recent Literature on Lifelog Research

The related studies listed in Section 2.1 are the main studies which have been published for NTCIR Lifelog tasks. In addition to NTCIR Lifelog, ImageCLEF Lifelog and UbiqLog are some of the well-known datasets which have been published in recent years as a contribution to lifelogging research and have generated considerable interest among researchers.

The ImageCLEF Lifelog dataset is quite similar to the NTCIR Lifelog dataset in nature. It consists of data from three lifeloggers for a period of one month each. The dataset is presented as images (approximately two images per minute) and an XML file specifying semantic location and activity of lifeloggers. Visual concept annotations for images are also made available for the use of researchers, and attached to the dataset as metadata. The dataset is introduced with two information retrieval tasks, namely Lifelog Retrieval Task (LRT) and Lifelog Summarization Task (LST), which are also similar in nature to LSAT and LIT tasks from the NTCIR Lifelog tasks (See Section 1.3) [12].

UbiqLog is announced as the first smartphone based lifelog dataset in 2013, and generated by using a lifelogging framework on mobile devices. The dataset contains user calls, SMS headers, application use, network devices, geographical location, and physical activities as data attributes, which are presented in JSON format. In addition, the data collection framework contains a data model and architecture, which can be used as a baseline for further lifelogging applications [13, 14].

Some of the recent works of literature on lifelog research are listed below.

In 2015, Amlinger published his study in which he compares the performance of different clustering algorithms on small and large size lifelog datasets. In this study, various branches of clustering algorithms are used for activity detection in geospatial datasets, namely partitioning-based, hierarchy-based, and density-based clustering algorithms. The performance of the algorithms are compared using Silhouette coefficient interpretation [15].

Similarly, Del Molino et al. propose a clustering pipeline for the ImageCLEF Lifelog task to summarize the lifelog data. They use image processing techniques to eliminate uninformative images. Next, they assign scores to images representing their relevancy to query events by making use of images itself together with location and activity information given in the metadata. They cluster images into query events according to their relevance score and report that using multiple features of data usually improves clustering performance [16].

Bolaños et al. provide an overview of the leading-edge research published for the task of story-telling from lifelogging data. First, they list improvements and capabilities of current hardware used for collecting lifelog data. Next, they provide a comprehensive categorical catalog of most recent studies on the subject of storytelling from visual lifelogging, which has been a valuable guide for researchers studying in the field [17].

In 2018, Truong et al. propose semantic concepts fusion approach to retrieve meaningful information from lifelogging data. They state that the purpose of their methodology is to efficiently assist users to retrieve events and memories from lifelog data. The query system they develop supports different types of query conditions and uses fusion techniques for information retrieval from lifelog data [18].

Ben Abdallah et al. publish their study in which they propose their multilevel deep learning-based processing for lifelog image retrieval in IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2018. They use deep learning methods at five different phases including data preprocessing, enhancement of metadata image description, semantic segmentation, query - concept matching and image retrieval. This multilevel approach is shown to perform well on some of the well-known learning tasks on the lifelogging research [19].

Recently, Dimicolli et al. publish their study in which they summarize state of the art methods together with their limitations, and future challenges in the subject of activity recognition from visual lifelogs. This study is similar to the one written by Bolaños et al. [17] in nature, but it can be perceived as an updated and extended catalog of current literature on activity recognition from visual lifelogs and is a valuable resource for researchers [20].

2.3 Recent Literature on Multimodal Classification

In their recent study, Baltrusaitis et al. emphasize that our experience of the world is multimodal, i.e. we see objects, hear sounds, feel textures, smell odors, and taste flavors. They define modality as the way in which something happens or is experienced, and a research problem is defined as multimodal when it includes multiple modalities. In this study, recent advances and possible future research topics in the field of multimodal machine learning are presented by a new common taxonomy, which serves as a valuable resource for the field [21].

In the field of multimodal classification, Liu et al. propose a two-component learning model which jointly learn multimodal matching and classification, which they call MMC-Net. The model first learns visual and textual features in the matching component; then, generates discriminative multimodal representations in the classification component. The effort to minimize the loss function of both components on a single model, i.e. multiloss training, results in improved classification performance [22].

Similarly, Zahavy et al. propose their multimodal fusion architecture which uses visual and textual data together for product classification in e-commerce. They train

two separate CNN structures for text and image dimensions and suggests a multimodal decision-level fusion approach, which outperforms state-of-the-art classification methods. Finally, they express their anticipation that multimodal classification will attract considerable interest from researchers in near future [23].

2.4 Summary

A summary of the recent literature on lifelogging research is given in Table 2.1 in comparison with the methodology proposed in this thesis.

Table 2.1: Summary of Recent Literature

Study	NTCIR Lifelog Dataset	Text-Based Classification	Image Classification	Multimodal Learning	Handling Missing Values	Manual Annotation and Classification
Amlinger, 2015 [15]	✗	✗	✗	✗	✗	✗
Ben Abdallah, 2018 [19]	✗	✓	✓	✓	✗	✗
Del Molino, 2017 [16]	✗	✗	✗	✓	✗	✓
Lin H., 2016 [9]	✓	✓	✗	✗	✗	✗
Lin J., 2017 [10]	✓	✗	✓	✓	✗	✗
Liu, 2018 [22]	✗	✓	✓	✓	✗	✗
Safadi, 2016 [8]	✓	✓	✓	✓	✗	✓
Truong, 2018 [18]	✗	✓	✓	✓	✗	✗
Xia, 2016 [7]	✓	✓	✗	✓	✗	✗
Yamamoto, 2017 [10]	✓	✓	✗	✓	✗	✗
Zahavy, 2018 [23]	✗	✓	✓	✓	✓	✗
Proposed Approach	✓	✓	✓	✓	✓	✓

CHAPTER 3

METHODS AND MODELS

3.1 Text-Based Classification

Text classification is defined as the task of classifying documents into predefined classes. More formally, if d_i is a document in the document set D and $\{c_1, c_2, ..., c_n\}$ is the set of classes, text classification is the task of assigning one class c_j to each document d_i in the document set D [24].

In our problem domain, a document is the sequence of visual concept annotations associated with a lifelog image, and the classes are activities.

We use bag-of-words model vector representation in the preprocessing phase of our study to convert image annotations into input vectors for text-based classifiers. Next, we train SVM and MLP models to learn activities from image annotations.

3.1.1 Vector Representation

In text classification terminology, a document is a sequence of words which is often represented by an array of words. The list of all the words which appear in a set of documents is called vocabulary, or feature set. Hence, a document can be converted to a binary vector, assigning the value 1 if the feature-word appears in the document or 0 in the case of no appearance.

The bag-of-words model is a simplified version of the vector representation, in which each document is represented by a batch of its words, ignoring grammar and word order, but preserving the number of appearances of words in the document.

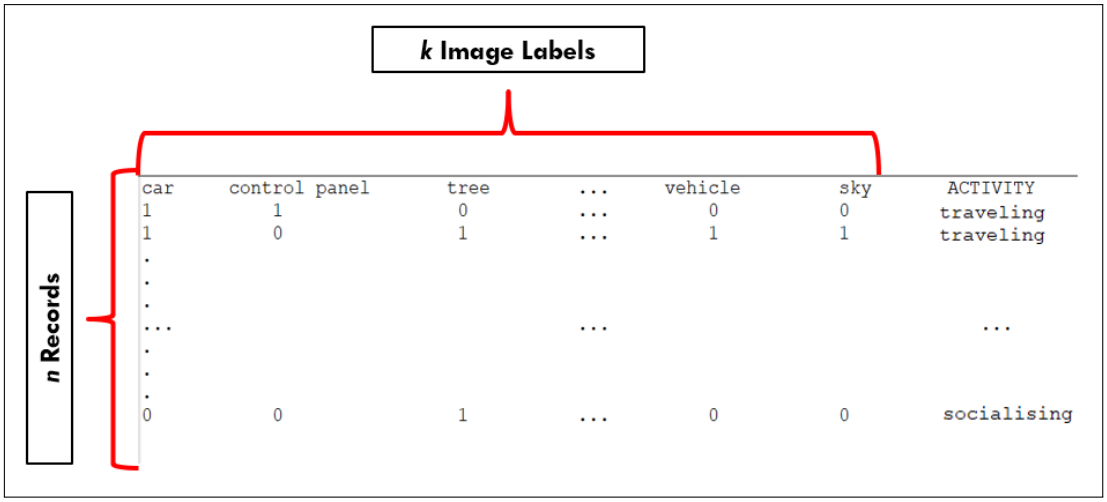

Figure 3.1: Vector Representation of Visual Concepts Annotations

During the preprocessing phase of our study, documents are processed to form a bag-of-words model vector representation since the word sequence, which are the visual concept annotations associated with lifelog images, does not show a specific order of appearance in our data. In our case, however, because every object is stated only once in the list of visual concept annotations, there is no word count either. Hence, the bag-of-words representation results in a set of binary vectors having the size of the vocabulary of image annotations.

A simple sketch of the vector representation is available in Figure 3.1.

3.1.2 Support Vector Machine (SVM)

SVMs were first introduced by Vapnik and Cortes in [25] in 1995. They have been a significant text classifier since they were shown to achieve substantial improvements in text classification by Joachims in 1998 [26].

The idea behind SVM is to find a hypothesis h that will minimize the upper bound on the true error, i.e. the probability of error on a randomly selected sample from the dataset, by efficiently and effectively controlling the Vapnik-Chervonenkis Dimension (VC-Dimension) of the hypothesis space. The VC-Dimension of a function F is defined as the cardinality of the largest dataset that can be shattered by F [26].

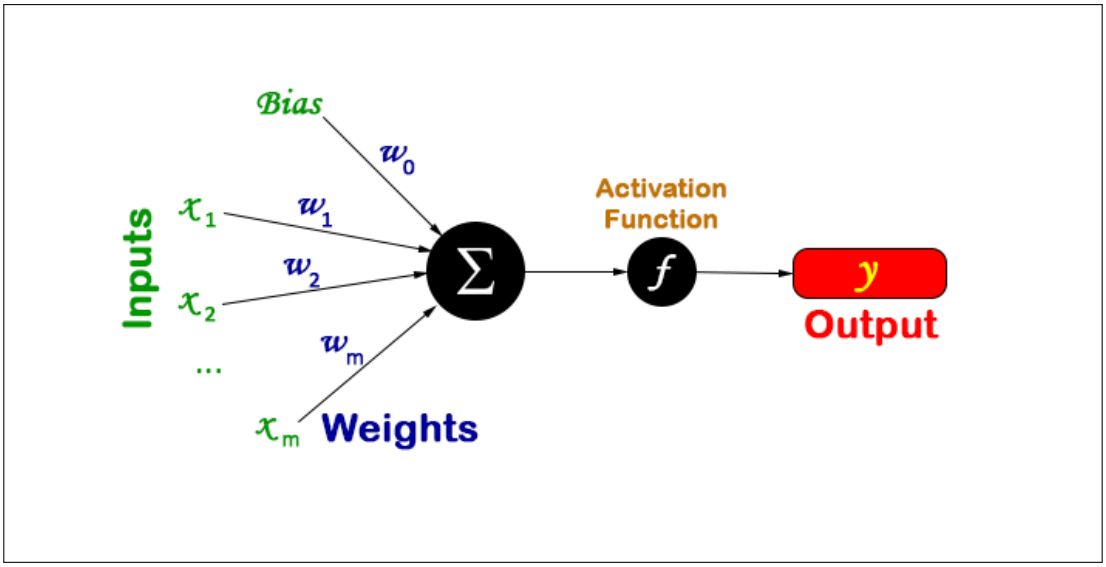

Figure 3.2: Structure of A Simple Perceptron

SVMs can learn linear bounding functions in the simplest form; however, more complex threshold functions can be learned by simply introducing the proper kernel function [26].

Sigmoid function, radial basis function (RBF) and polynomial function are the kernel options which are trained and used in this study together with the linear SVM.

3.1.3 Multilayer Perceptron (MLP)

MLP is a class of fully connected feed-forward artificial neural network, which should be composed of at least three layers: one input layer, one output layer, and one or several hidden layers. In the concept of a fully connected network here, each node is connected to every single node in the following layer [27]. The structure of a simple node, i.e. perceptron, is visualized in Figure 3.2.

In this study, an MLP model with two hidden layers is designed using ReLU (31) as activation function between the hidden layers, and Softmax (32) in the output layer. Additional dropout layers with a loss rate of 0.5 are inserted in order to avoid overfit-

ting; resulting in higher classification accuracy on the test data.

$$f(x) = max(0, x) \tag{31}$$

$$g(x) = \frac{e^x}{\sum_{i=0}^{m} e^{x_i}}, \quad i = 0, 1, 2, ..., m \tag{32}$$

The structure of the proposed MLP model is shown in Figure 3.3. The model performance is measured using categorical cross-entropy (33) as the loss function, and accuracy as the performance metric for classification of minutes into activity classes.

$$CCE(y, \hat{y}) = -\sum_{i=0}^{N}\sum_{j=0}^{M}(y_{ij} * log(\hat{y}_{ij})) \tag{33}$$

where;

$\boldsymbol{y}$: Ground truth label

$\boldsymbol{\hat{y}}$: Predicted label

$\boldsymbol{M}$: Number of categories

$\boldsymbol{N}$: Number of records

3.2 Image Classification

Image classification refers to the task of assigning an input image a class label from a predefined set of categories according to its visual content, i.e. pixel values. Although it is a simple task in nature, it has various practical applications. Moreover, a number of different computer vision tasks (such as object recognition and segmentation) can be reduced to the task of image classification. These are the main reasons why image classification has been one of the core problems in the field of computer vision [28].

In this study, we analyze and use CNN and ResNet, which is a customized version of CNN, models as a solution proposal for the image classification problem.

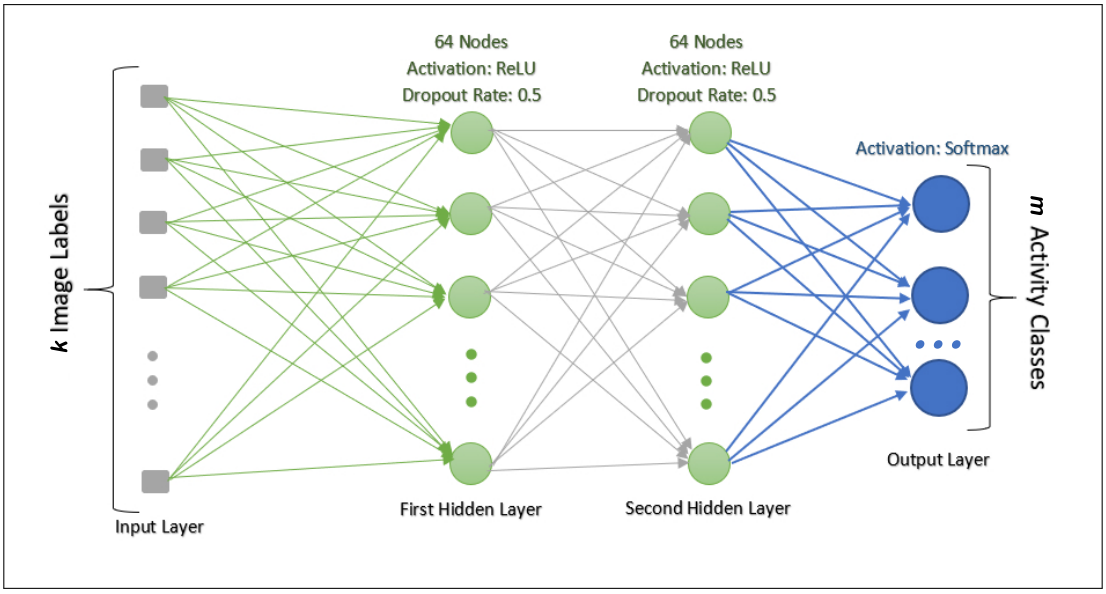

Figure 3.3: The MLP model

3.2.1 Convolutional Neural Network (CNN)

CNN is a customized version of neural network which leverages three important ideas that can help improve a machine learning system: sparse interactions, parameter sharing and equivariant representations.

In traditional neural networks, every output unit interacts with every input unit. Convolutional networks, however, accomplish sparse connectivity by using kernels having smaller size than the input. In this way, it becomes possible to detect small details such as edges and corners within an image which consists of millions of pixels.

Parameter sharing refers to using the same parameter for more than one function in a model. While each element of the weight matrix is used only once in a traditional neural network, each member of the kernel is used at every position of the input in a convolutional neural network, which results in reduced memory requirements of the model.

The special form of parameter sharing in convolutional neural networks causes the layer to gain the equivariance property, which means if the input changes, the output will change in the same way. In other words, if an object in the input image is moved, its representation will move the same amount in the output [29].

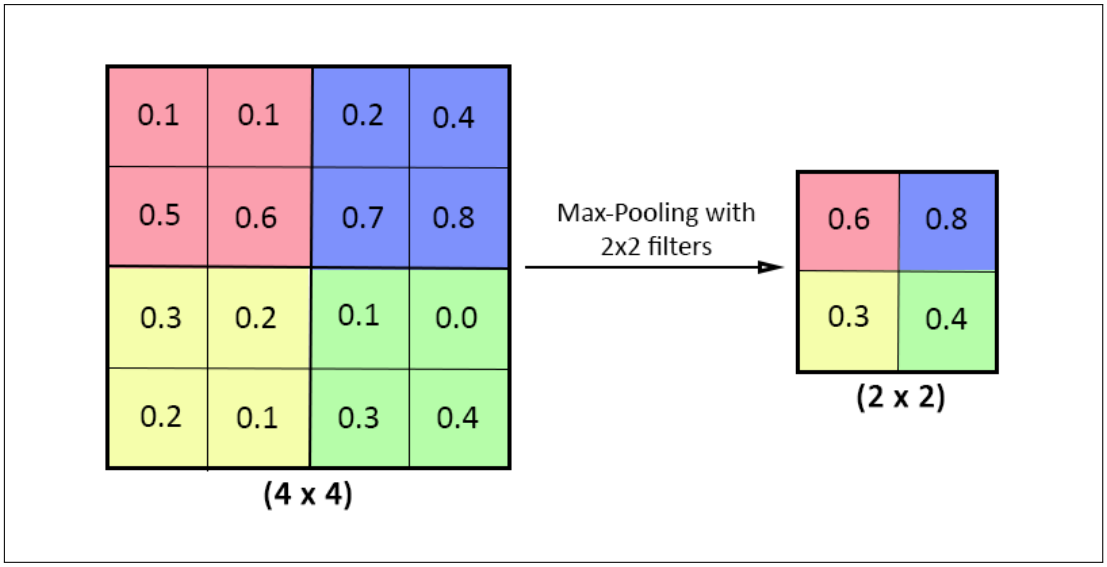

Figure 3.4: A Symbolic Representation of Max-Pooling

CNNs benefit from the assumption of the input having grid-like topology, and construct a model which will process the input more efficiently by making use of convolution (34) and pooling (See Figure 3.4 for a symbolic representation of max-pooling) operations at one or several layers of the network [30].

$$(\boldsymbol{f} * \boldsymbol{g})[\boldsymbol{n}] = \sum_{m=-\infty}^{\infty} f[m]g[n-m] \tag{34}$$

In this study, we use a common CNN structure to learn activities from grayscale and RGB color images. The proposed CNN model consists of 3 convolutional layers, 2 of which are followed by a max-pooling layer. After the last max-pooling layer, image matrices are flattened into one-dimensional vectors. Following the flatten layer, 2 dense layers are added to increase the model depth. ReLU (31) is the activation function which is used in each layer of the model, except for the output layer. In the output layer, Softmax (32) is used as the activation function. In addition, 3 dropout layers are inserted into the model with the purpose of avoiding overfitting.

The structure of the proposed CNN model is visualized in Figure 3.5. Similar to our methodology for text-based classification using MLP, we evaluate the model performance using categorical cross-entropy (33) as the loss function, and classification accuracy as the performance metric.

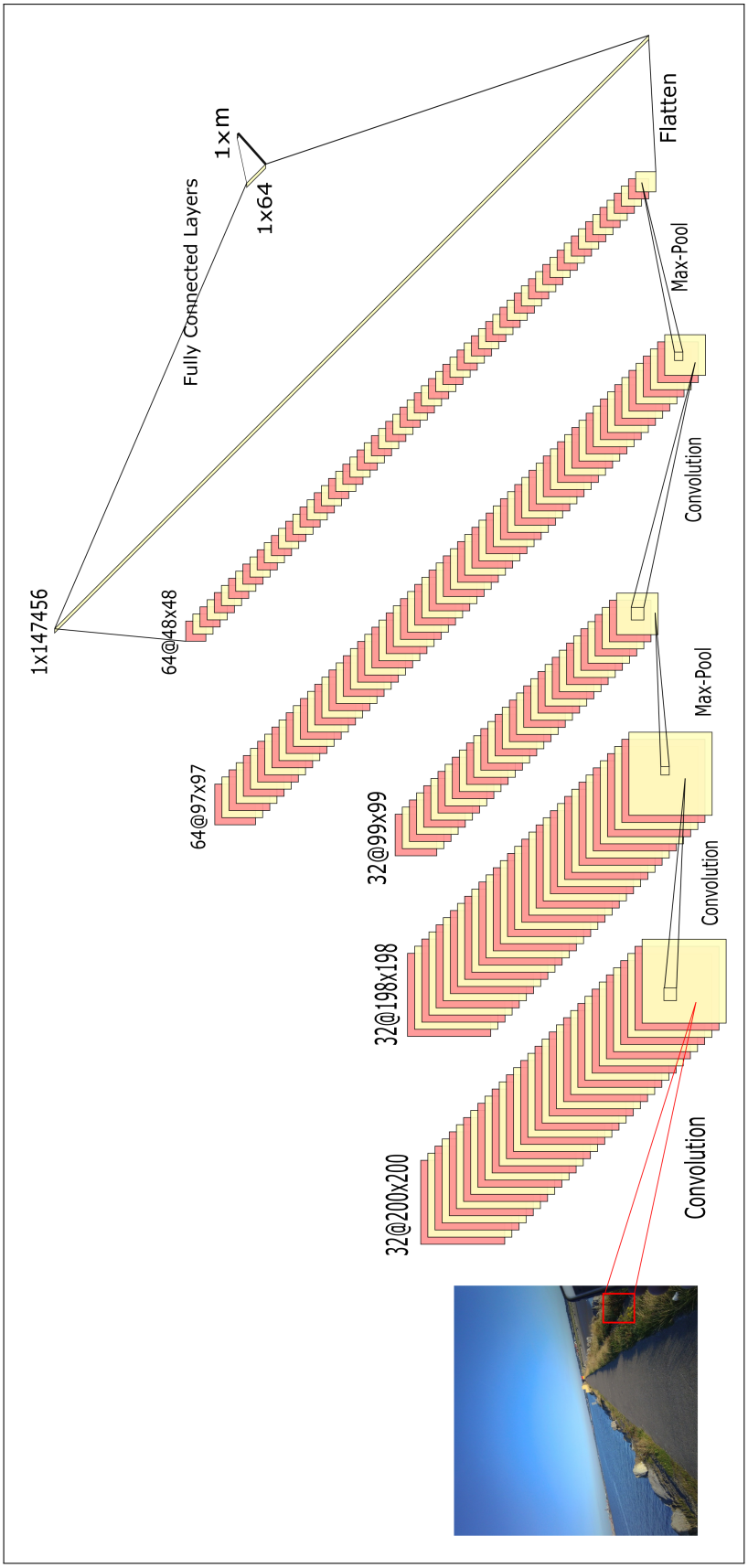

Figure 3.5: The CNN Model

3.2.2 Residual Neural Network (ResNet)

Unlike traditional neural networks, in which each node is connected to nodes in the next layer only, in a residual block of the ResNet model, each layer feeds the next layer together with more nodes in subsequent layers. The residual learning method increases depth and complexity of the learning network, which is shown to considerably improve the learning performance in visual recognition tasks [31].

The residual neural network we use in this study is comprised of a series of identity blocks (See Figure 3.6) and convolution blocks (See Figure 3.7), which are stacked on top of each other to obtain a 50 layer learning model [31]. The model, which is pretrained on the ImageNet dataset [1], is fine-tuned on the NTCIR Lifelog dataset, and classification performance is evaluated using categorical cross-entropy (33) as the loss function, and classification accuracy as the performance metric.

3.3 Learning from Image and Text Data

In the version of the NTCIR Lifelog dataset which is used in this study, approximately 50% of the available records have both text and image dimensions, so we have an intuition that we could create a combined artificial neural network which will expect two inputs, the first one being textual input and the second one as visual input, and learn from the two dimensions together on a single model to increase classification performance. For this purpose, we extract layers before the Softmax layer from the best performing text and image classifiers, which are MLP and ResNet-50 respectively, and concatenate them at the merged layer of the new combined network structure. The resulting learning model is visualized in Figure 3.8. We use categorical cross-entropy (33) loss to train the model and accuracy as the performance metric.

3.4 Handling Missing Values

In the NTCIR Lifelog dataset version which is used in this study, there exist records of 4 categories:

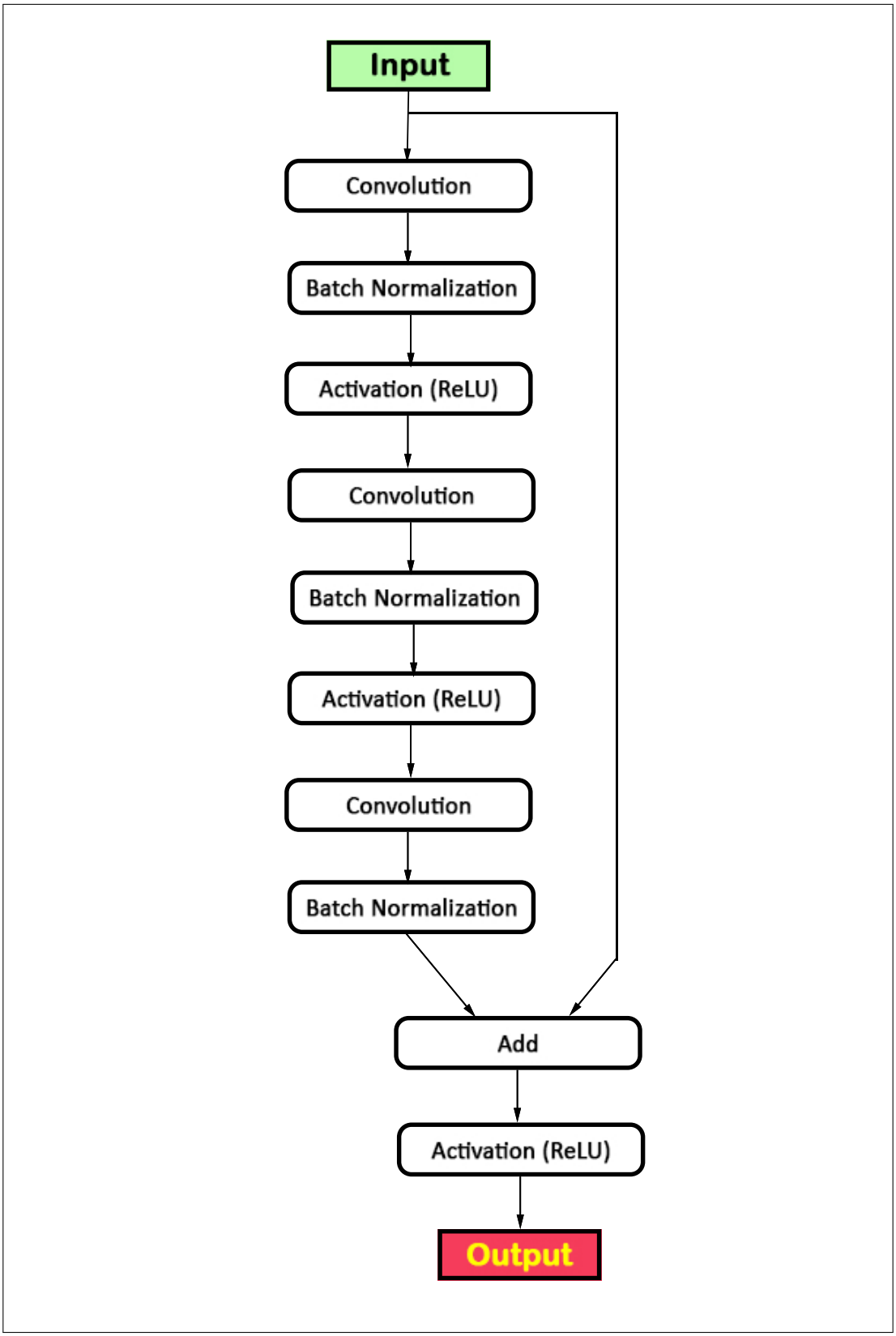

Figure 3.6: Symbolic Representation of An Identity Block

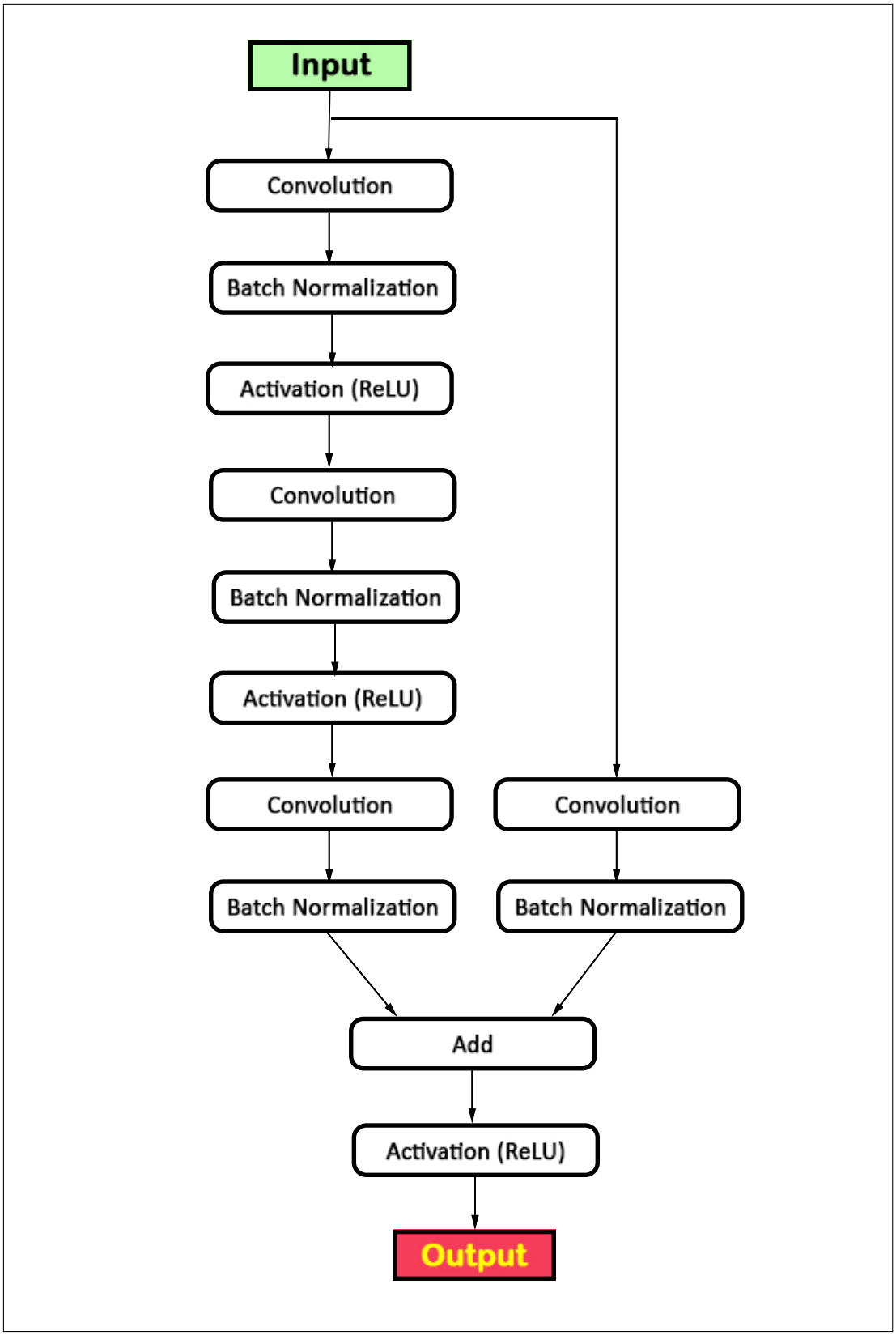

Figure 3.7: Symbolic Representation of A Convolution Block

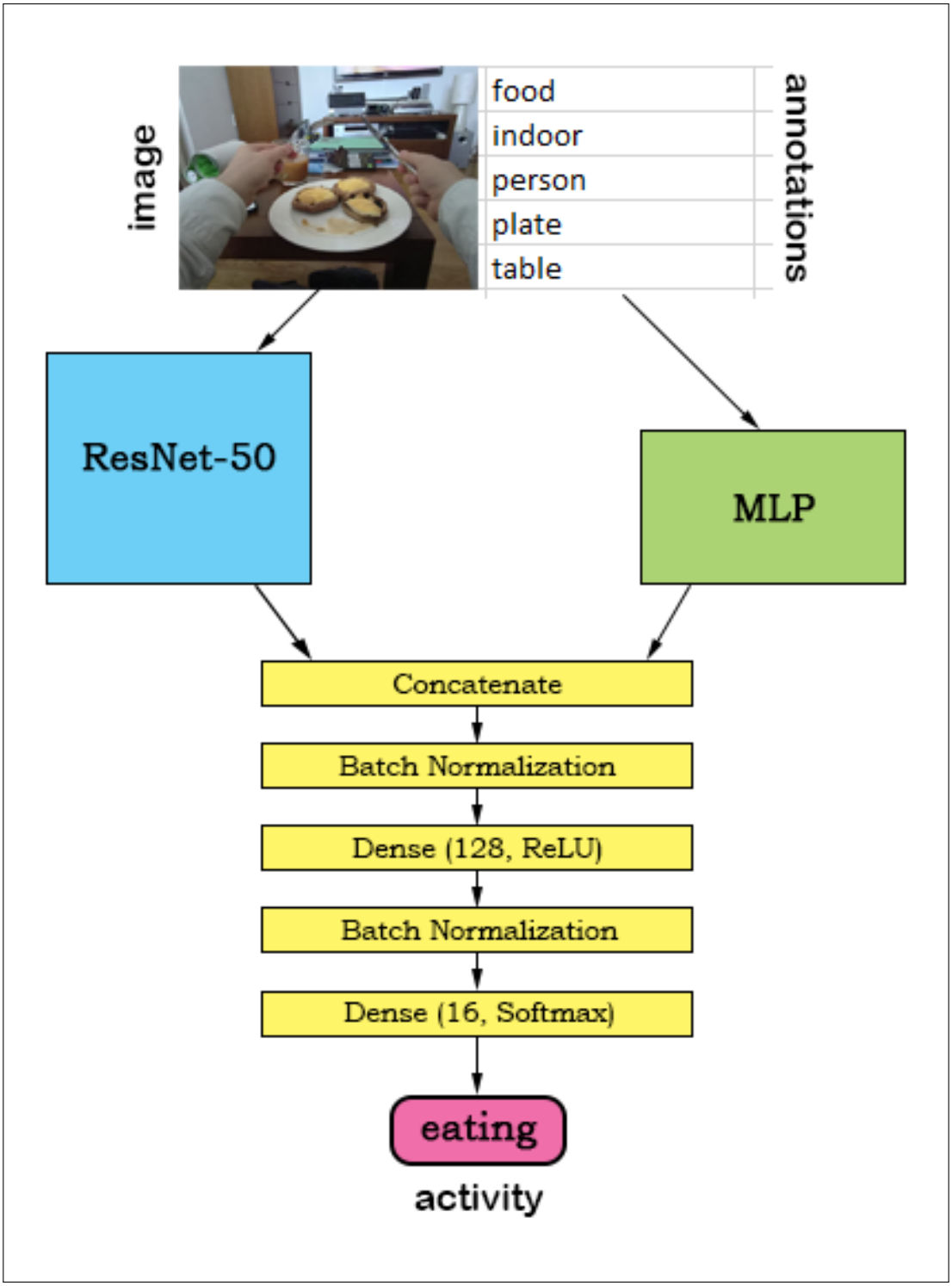

Figure 3.8: Structure of the Combined Learning Model

1. Images with annotations and corresponding minute record from the dataset

2. Images without annotations, with the corresponding minute record from the dataset

3. Annotations, with the corresponding minute record from the dataset, but image file not provided

4. Minute records without images or annotations

The combined learning model we propose in Section 3.3 expects both image and text data to be provided for each record, which falls into the 1st category. Our study is not capable of classifying records from the 4th category because there is no image or text data; however, we could propose a classifier which can learn activity classes from records of the other 2 categories. To this end, we first describe the naive method to predict activities in the presence of missing modalities. Next, we introduce the multi-loss combined model which can learn from text and image data together and handle missing modalities thanks to the proposed custom loss function.

3.4.1 Naive Method for Activity Prediction

In the naive method for activity prediction, we train a text-based classifier (MLP), an image classifier (ResNet-50) and a combined classifier on separate training sessions. Next, we determine the proper prediction model for every single test record on run-time according to the nature of the record (See Figure 3.9). In other words, if current test record has only text data as in the 3rd category records, we predict the activity for the record using text-based classifier. Similarly, if the test record has only image data like a 2nd category record would have, we use ResNet image classifier. We use the combined classifier for prediction only if both text and image data are available for the test record, i.e. records from the 1st category.

Prediction performance of the naive method is calculated as the average classification accuracy of the three classifiers.

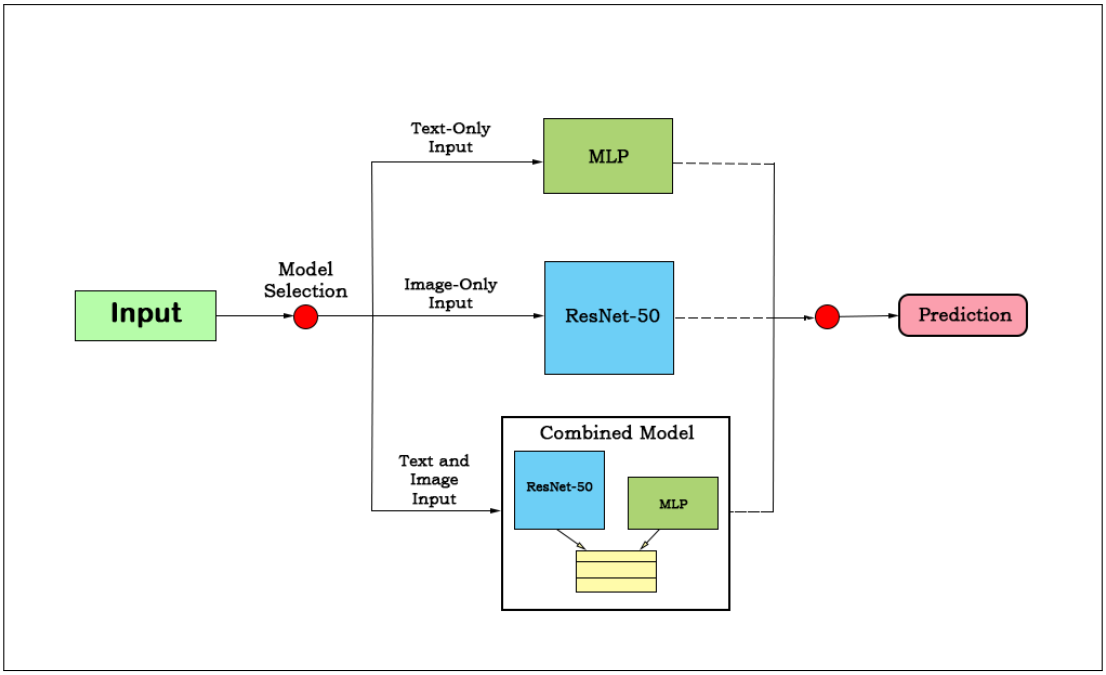

Figure 3.9: Symbolic Representation of the Naive Prediction Method

3.4.2 Multi-Loss Combined Model

The naive method for activity prediction has the drawbacks that we need to deal with training and fine tuning of the three learning models separately, and manually interfere in the testing to determine the right classifier.

In order to overcome these difficulties, we propose the multi-loss combined learning model, which is a multi-input and multi-output model, which takes images and annotation vectors as input, and has two intermediate and one final output (See Figure 3.10)

In this model, in order to compensate the absence of images as in the 3rd category records, a zero-image, i.e. an image of the same shape as other input images having all pixels set to zero, is fed to the model. Similarly, in case of absence of annotations as in the 2nd category, a zero-vector having the size of the number of words in the vocabulary of annotations is fed to the MLP sub-section of the model.

It is necessary to prevent the model from learning from zero-input as it does not contain any valuable information for the record being processed. To this end, while we

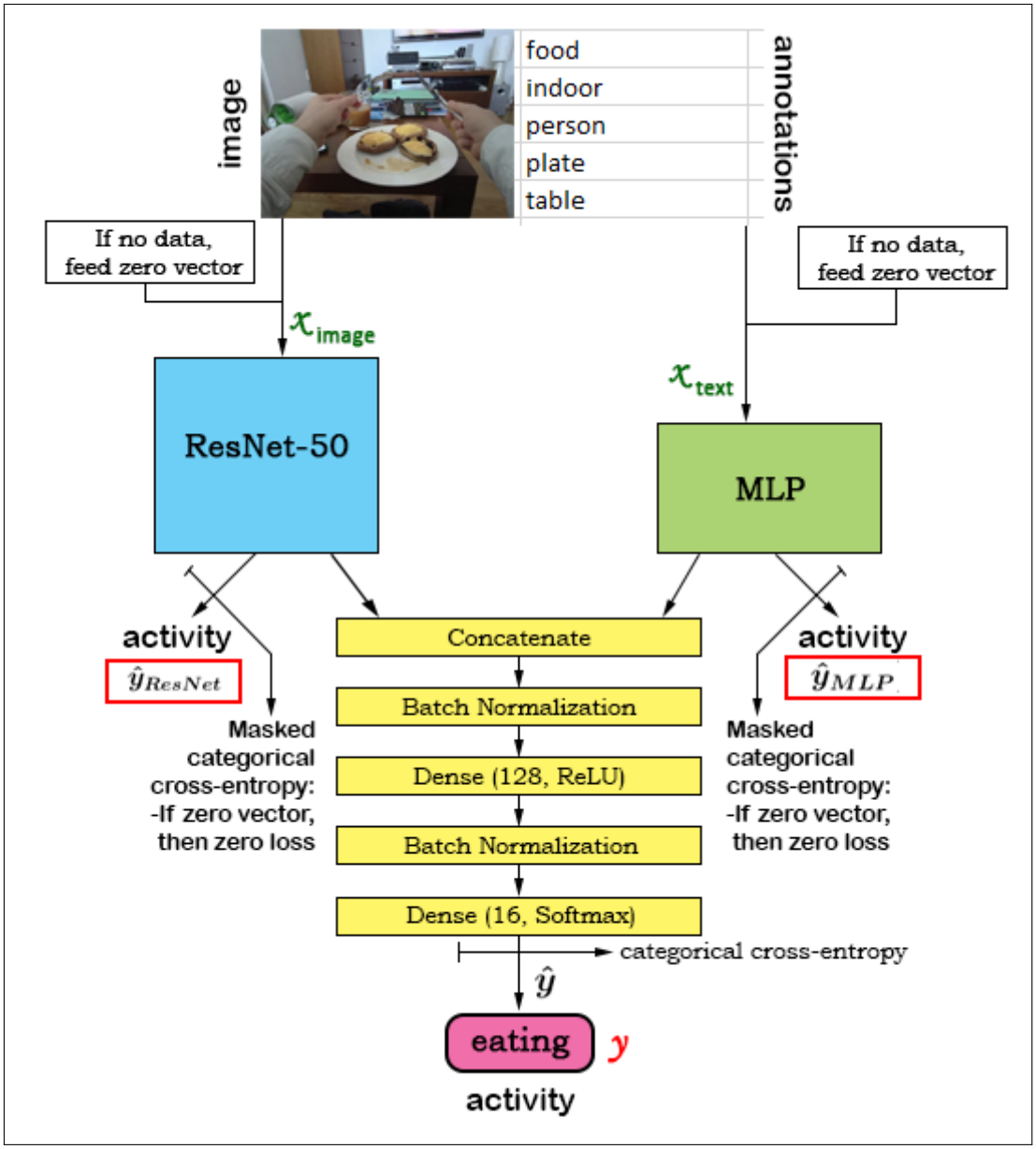

Figure 3.10: Structure of the Multi-Loss Combined Model

use categorical cross-entropy (33) as the loss function for the final output, we introduce a masked version of the categorical cross-entropy, which makes use of the input to create the mask, as loss function for the two intermediate outputs (See Equation (35) for the masked loss function devised for the ResNet-50 image classifier, and Equation (36) for the masked loss function of MLP text-based classifier).

The masks in Equations (37) and (38) ignore the contribution of the input record to the loss function if all of the values in the record are equal to zero, i.e. it is a zero-image or zero-vector. By this masks, we are able to safely feed zero-records in place of missing values.

$$Loss_1(x, y, \hat{y}_{ResNet}) = -\sum_{i=0}^{N}(c(x_i) * \sum_{j=0}^{M}(y_{ij} * log(\hat{y}_{ResNet,ij}))) \tag{35}$$

$$Loss_2(t, y, \hat{y}_{MLP}) = -\sum_{i=0}^{N}(c(t_i) * \sum_{j=0}^{M}(y_{ij} * log(\hat{y}_{MLP,ij}))) \tag{36}$$

$$c(x_i) = \begin{cases} 1, & \text{if } sum(x_i) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{37}$$

$$c(t_i) = \begin{cases} 1, & \text{if } sum(t_i) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{38}$$

where each record can be represented as a tuple r;

$\boldsymbol{r} = (x, t, y, \hat{y}_{ResNet}, \hat{y}_{MLP}, \hat{y})$

$\boldsymbol{x}$: Image input

$\boldsymbol{t}$: Text input

$\boldsymbol{y}$: Ground-truth label

$\boldsymbol{\hat{y}_{ResNet}}$: Prediction generated by ResNet-50 sub-section of the model

$\boldsymbol{\hat{y}_{MLP}}$: Prediction generated by MLP sub-section of the model

$\boldsymbol{\hat{y}}$: Prediction generated by the complete model

$\boldsymbol{M}$: Number of categories

$\boldsymbol{N}$: Number of records

CHAPTER 4

EXPERIMENTS AND RESULTS

4.1 Environment Setup

Platforms, software, and libraries used for the development of our approach are listed below:

- Programming Languages: Java programming language (JDK 1.8) is preferred in the data preprocessing phase because of file I/O and database facilities, and object-oriented nature. In the learning phase, Python language is used because it provides a variety of library support for image processing and machine learning studies.

- Database Management Systems: The data are stored in a MySQL relational database and a MongoDB NoSQL database with the purpose of better data analysis on the lifelog dataset.

- Computer Vision Libraries: OpenCV is used for importing and reshaping images in the Python environment.

- Neural Network Libraries: Keras with Tensorflow support is used as the platform to design, train and test deep neural networks for classification.

- Data Analysis and Machine Learning Libraries: Scikit-learn library is used to process data and apply machine learning algorithms.

- Development Environment: Java programming is done on IntelliJ IDEA Community Edition. For Python programming, PyCharm IDE is used for small-size

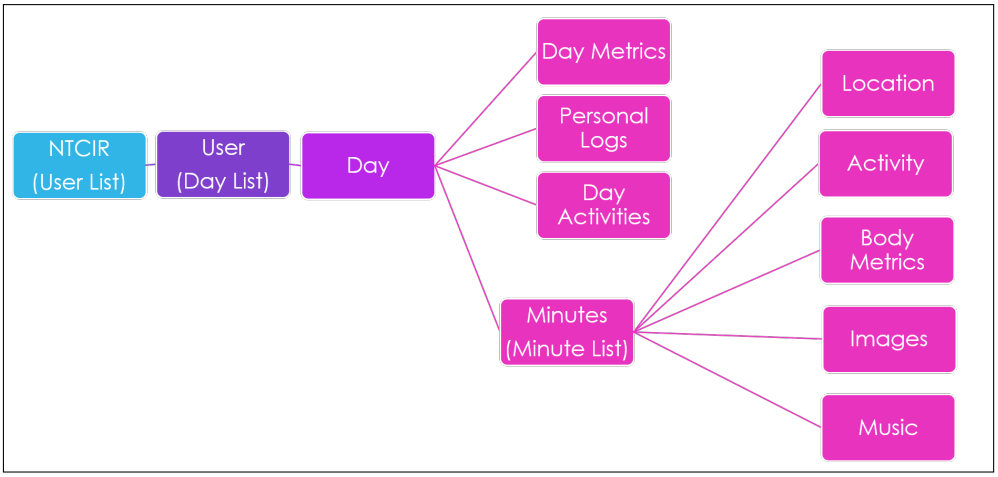

Figure 4.1: The Object Oriented Model

calculations, whereas learning algorithms which require high-performance calculation are run on Google Colab environment with GPU support.

4.2 Data Analysis and Preprocessing

In the NTCIR Lifelog dataset, the main component is the user. Each user is defined by a list of days, each of which is then defined by a list of minutes, in addition to several daily measurements. Each minute is defined by images taken at that minute, together with additional minute-based measurements and details which can be seen in Figure 1.3. From this point of view, the NTCIR Lifelog dataset can be considered as an object-oriented model. Our first effort is therefore to extract an object-oriented model, whose structure is illustrated in Figure 4.1, from the XML document describing the dataset. Next, this model is saved to MySQL and MongoDB databases with the purpose of better data analysis and summarization.

During our efforts to understand the dataset, we realize that the dataset has a lot of missing values, so it is necessary to analyze and preprocess the data to extract significant features. A summary of the results of the numerical analysis of the dataset is given in Table 4.1.

The figures extracted from the numerical analysis and the results of our preliminary research on recent studies have led us to search for a solution for the problem of

Table 4.1: Numerical Analysis of the NTCIR Lifelog Dataset

Feature	**Value**
Number of users	2
Number of days	90
Number of minutes (per day)	1440
Number of minutes (total)	129 600
Minutes with location information	104 118
Minutes with activity definition	11 041
Minutes with body measurements	90 000 (Approx.)
Minutes with images	70 000 (Approx.)
Minutes with music information	763
Number of activity definitions	5
Number of image annotations	361
Frequency of image annotations	Ranges from 2 to 4000
Size of the dataset on disk	26.5 GB (Approx.)
Image dimension	1024 x 768 pixels
Total number of images	110 000 (Approx.)
Number of annotated images	70 000 (Approx.)
Number of minutes having both activity definition and annotated images	9058

Table 4.2: Activities and Frequencies in the NTCIR Lifelog Dataset

Activity	**Frequency**
Airplane	994
Cycling	2
Running	1
Transportation	5743
Walking	2318

classifying minutes into activity classes by using images and image annotations in the NTCIR Lifelog dataset.

In this task, three approaches are possible:

1. Text-based classification, i.e., using image annotations with activity definitions to classify the images

2. Image classification, i.e., using original images with activity definitions for classification

3. Multimodal classification, i.e. using images and image annotations together for classification

In the dataset, 5 activity classes are available with frequencies given in Table 4.2. Among these classes, *cycling* and *running* are eliminated from the dataset in the early stages because they have very low frequencies (2 and 1 records, respectively). Thus, during our initial experimentation, activity classification is performed based on three activity classes; namely *airplane, transportation* and *walking*.

One major problem regarding the available activity classes is that the three classes happen to be inadequate in expressing lifeloggers' daily life, which can be inferred from the figures in Table 4.1, as well. While the total number of minutes is approximately $130\,000$, the number of minutes with activity definition is almost $11\,000$, which results in a ratio of 0.085, i.e. the three (or five) activity classes can represent only 8.5% of the whole dataset.

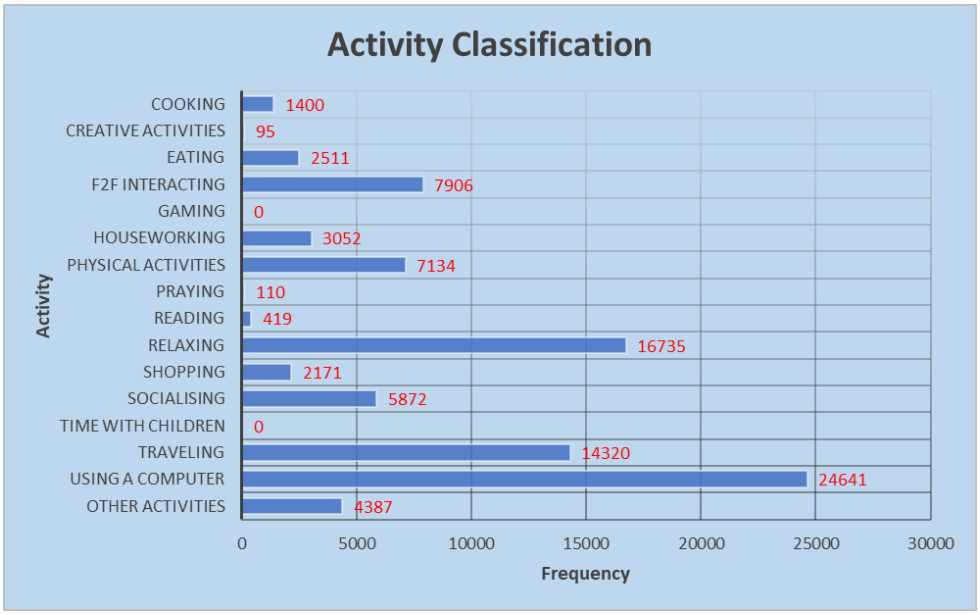

Figure 4.2: Activity Classes and Frequencies

The purpose of the lifelogging research is to provide insight on lifeloggers' daily life, so our research would be more substantive if the classification approach could be generalized to a whole day of lifeloggers. With this initiative, we made an effort to extend the available classification with more activity classes, so that it could better represent daily life. We used the 16 activities which are specified in the definition of the LADT task (See Section 1.3), so that our research progresses in parallel with the NTCIR Lifelog research.

Finally, approximately $90\,000$ images are examined and manually assigned to the 16 activity classes, which requires an effort of approximately 200 person-hours. The classified images are all taken from the images of the 1^{st} user, and span a period of 60 days. The frequency distribution of activities is shown in Figure 4.2, and Table 4.3 gives a numerical summary of the 16-class version of the data.

The experimentation in this thesis, therefore, has two stages of classification. First, we train and evaluate the performance of classification algorithms on the original 3-class data. Next, we apply the same set of learning methods to the manually-classified 16-class data. Hence, we have a chance to observe the effect of increasing input size on the performance of the proposed learning methods.

Table 4.3: Numerical Analysis of the Improved Version of the NTCIR Lifelog Dataset

Feature	**Value**
Size of the dataset on disk	1.5 GB (Approx.)
Image dimension	200 x 200 pixels
Number of users	1
Number of days	60
Number of activity definitions	16
Number of image annotations	570
Frequency of image annotations	Ranges from 1 to 40 000
Number of records with activity definition	90 000 (Approx.)
Number of images with annotations	47 000 (Approx.)
Number of images without annotations	42 000 (Approx.)
Number of records with annotations, but no image file	96
Number of records without images or annotations	40 000 (Approx.)

For the training of the text-based classifier, image annotations are processed and transformed into a vector representation, which is described in detail in Section 3.1.1. Variable values for the representation in Figure 3.1 are realized as in Table 4.4 for the case of 3-class classification, whereas they are upgraded as in Table 4.5 for 16-class classification.

For the training of the image classifier, dataset images are resized to 200×200 pixels due to limited resources, and with the purpose of decreasing training and test duration.

Table 4.4: Values of Variables for 3-Class Classification

Variable	**Value**
n	9055
k	356
activities	airplane transportation walking

Table 4.5: Values of Variables for 16-Class Classification

Variable	**Value**
n	90 753
k	570
activities	cooking creative activities eating f2f interacting gaming houseworking physical activities praying reading relaxing shopping socialising time with children traveling using a computer other activities

In the following sections, the two problem settings, namely 3-class classification and 16-class classification, and our experimentation with different learning algorithms are briefly described. The classification performance of the algorithms is then stated comparatively in the Results section (Section 4.6).

4.3 Training and Test Data

In the context of this study, we have two different problem settings, namely 3-class classification and 16-class classification. The two versions of the dataset, which are used for training and test on these different settings, both have missing modalities at non-negligible ratios.

With this information on mind, we generated two different test sets for each of the two problem settings, which we call *complete* and *incomplete* test sets. In the complete test set, all of the records have both image and text data available. On the other hand, incomplete test set has missing values in different modalities. A record we take from the incomplete test set could have both image and text data, only image data missing annotations, or only text data available. Our experiments run on both of these test sets for each learning model for both of the 3-class and 16-class classification problems.

With the goal of getting the most out of the available data, we train our models using the largest possible subset of the dataset for each learning model. To be more specific, after we separate the test sets from the dataset, we obtain the training data which has missing values in different modalities. In order to train the text-based classifier, we use all of the training records which have image annotations. Similarly, we train image classifiers using all of the training data having images. The combined model is trained using records which have both image and image annotations. Finally, we are able to feed all records which have at least one dimension, i.e. only image, only image annotations or both, to the multi-loss combined model for training.

A numerical summary of the training and test sets is available in Table 4.6.

Table 4.6: Number of Records in Training and Test Sets

Dataset	Content Type	3-Class Classification	16-Class Classification
Training Data	Text-Only	60	58
	Image-Only	5143	29 690
	Image and Text	6329	32 966
	Total	11 532	62 714
Test Data	Complete Data	2717	14 123
	Incomplete Data	4943	26 878

4.4 3 - Class Classification

As explained in Section 4.2, the NTCIR Lifelog dataset has 5 activity classes, 2 of which are eliminated due to very low frequency. In this problem setting, our experimentation is carried out according to the details in Tables 4.4 and 4.6.

Results of the experimentation is presented in Section 4.6.

4.4.1 Text-Based Classification

Text-based classification is performed by using image annotations as input in the form of a binary vector, and the activity class as the output to classification tools. We use the SVM method and an MLP model for the text-based classification task. Details of the proposed methods are stated in Section 3.1. Parameter values for different SVM kernels are determined using cross-validation method.

4.4.2 Image Classification

Image classification is performed using all of the available images which are associated with an activity class in grayscale and RGB color modes. Image files are imported and processed using Python's OpenCV library, and they are given as input to the custom CNN and ResNet-50 models, details of which are given in Section 3.2.

4.4.2.1 Classification Using Grayscale Images

Input images are converted into grayscale images for faster training and with the purpose of understanding the effect of color dimension in image classification. Converted images are then fed to the CNN model.

4.4.2.2 Classification Using RGB Images

Input images are fed to the CNN model without being converted to grayscale images. The model is trained using RGB images as input, and activity classes as output. Classification accuracy of the model is recorded.

The common CNN structure which is used for learning from grayscale and RGB images is introduced and described in detail in Section 3.2.1.

4.4.2.3 Classification using ResNet-50 Architecture

As a more structured classification approach, input images are fed into a sample of ResNet-50 architecture, the structure of which is described in Section 3.2.2.

The training time for the ResNet-50 was longer than the regular CNN model, as a result of the depth and complexity of the network. Correspondingly, ResNet-50 appeared to perform better classification than regular CNN models, which will be shown in the following sections (Section 4.6).

4.4.3 Multimodal Classification

With the purpose of using image and text dimensions of data together to train a single learning model, we use a naive activity prediction method, and two combined neural networks, which are described progressively in Sections 3.3 and 3.4.

4.4.3.1 Combined Learning Model

The combined neural network model expects one image and one text input, and predicts user activity as output. We train the combined model using records which have both image and image annotations available. This is the reason why the size of the training dataset appears to be relatively smaller than the other learning models, which, in turn, affects the classification performance of the model.

4.4.3.2 Naive Prediction Method

In the naive method for predicting activities in the presence of missing values, we use the weights of MLP, ResNet-50 and combined model which were previously trained on separate training sessions. Next, for each record in the test set, we determine the appropriate prediction model according to the content of the record with a simple if-else statement. Thus, the performance of the naive prediction method depends heavily on the performance of the three learning models.

4.4.3.3 Multi-Loss Combined Model

As it can be inferred from the explanations above, both the combined model and the naive method have several disadvantages: The combined model is unable to handle missing input modalities, and the naive method requires manual interference at many different stages.

In order to be able overcome these drawbacks, we propose the multi-loss combined learning model, which expects one image and one text input, but can continue learning when one of the two inputs are not available thanks to the proposed custom-loss function. By this way, we can use a single model to learn from text and image data simultaneously, and still have a large training set.

When we compare the prediction performance of the multi-loss learning model with the performance of the naive model, we observe that results of our experimentation with the proposed multimodal and multi-loss model are promising on the NTCIR Lifelog dataset.

4.5 16 - Class Classification

In this problem setting, annotated lifelog images which are manually assigned to 16 activity classes are used to train artificial neural networks. We use the same set of learning models which are used in 3-class classification, and have the opportunity to observe the effect of increasing data size on performance of the proposed learning models .

4.6 Experimentation Results

Results of our experimentation with 3-class and 16-class data are presented in Table 4.8 and Table 4.9, respectively. Number of trainable parameters for learning models is specified in Table 4.7 as a reference to the size of the models. In addition, we provide confusion matrices of complete and incomplete test sets in the two problem settings in Figures 4.3, 4.4, 4.5 and 4.6.

As we can infer from the experimentation results, non-linear models are able to show better performance than linear models in classifying image annotations. However, image classification algorithms perform better than text-based classification algorithms in our dataset. The color dimension is a factor which can increase classification accuracy. In addition, as the depth and complexity of the network increase, we can observe a significant increase in the performance of classification. Specifically, the classification accuracy of the ResNet-50 architecture on the NTCIR Lifelog dataset is considerably high.

As we can see in the results of both 3-class and 16-class problem settings, a combination of the image and text classification algorithms results in significant improvement in classification accuracy.

Finally, the model which we propose with a masked loss function has the advantage of both allowing learning in the presence of missing values and showing high accuracy in classifying lifelog records into activity classes. The proposed model performs better than the naive prediction method both for complete and incomplete test data.

Table 4.7: Number of Trainable Parameters

Algorithm	**Number of Trainable Parameters**	
	3-Class Classification	**16-Class Classification**
Linear SVM	818 748	10 158 960
SVM with Sigmoid Kernel	1 011 161	11 354 560
SVM with RBF Kernel	1 003 941	10 151 120
SVM with Polynomial Kernel	877 591	10 279 360
MLP	27 523	41 104
CNN (Grayscale Images)	9 465 897	9 466 768
CNN (Color Images)	9 466 473	9 467 344
ResNet-50	23 540 739	23 567 376
Combined Model	23 841 475	23 855 888
Multi-Loss Combined Model	23 843 593	23 885 488

We can infer from the confusion matrices that the classes which have relatively low frequencies are responsible for the decrease in classification performance. Thus, some of the 16 activity classes could be redefined, eliminated or merged to be able to obtain more accurate predictions.

In addition, the body measurements and location data from the NTCIR Lifelog dataset could be included in our final learning model for better prediction.

Table 4.8: Classification Performance on 3-Class Data

Resource	**Number of training records**	**Number of validation records**	**Algorithm**	**Accuracy**	
				Complete Data	**Incomplete Data**
Image Annotations	5111	1278	Linear SVM	0.837	0.566
			SVM with Sigmoid Kernel	0.832	0.563
			SVM with RBF Kernel	0.848	0.572
			SVM with Polynomial Kernel	0.850	0.574
			MLP	0.850	0.574
Grayscale Images	9178	2294	CNN	0.884	0.881
Color Images	9178	2294	CNN	0.892	0.890
			ResNet-50	0.923	0.920
Images and Annotations	5064	1265	Combined Model	0.909	0.891
	-	-	Naive Prediction	0.909	0.914
	9226	2306	Multi-Loss Combined Model	0.932	0.927

Table 4.9: Classification Performance on 16-Class Data

Resource	Number of training records	Number of validation records	Algorithm	Accuracy	
				Complete Data	Incomplete Data
Image Annotations	26 420	6604	Linear SVM	0.657	0.384
			SVM with Sigmoid Kernel	0.635	0.372
			SVM with RBF Kernel	0.662	0.386
			SVM with Polynomial Kernel	0.657	0.384
			MLP	0.663	0.406
Grayscale Images	50 118	12 538	CNN	0.742	0.747
Color Images	50 118	12 538	CNN	0.770	0.774
			ResNet-50	0.806	0.808
Images and Annotations	26 376	6590	Combined Model	0.817	0.805
	-	-	Naive Prediction	0.817	0.814
	50 164	12 550	Multi-Loss Combined Model	0.856	0.857

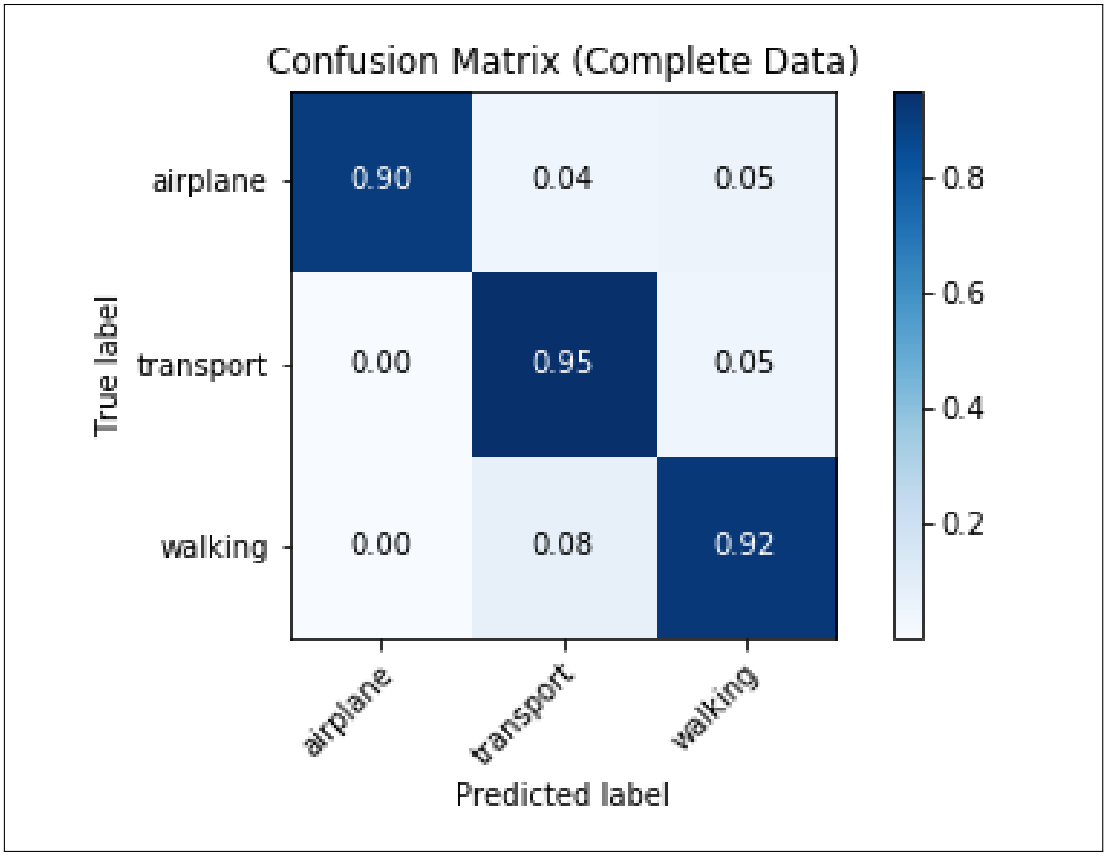

Figure 4.3: Confusion Matrix for 3-Class Complete Test Set

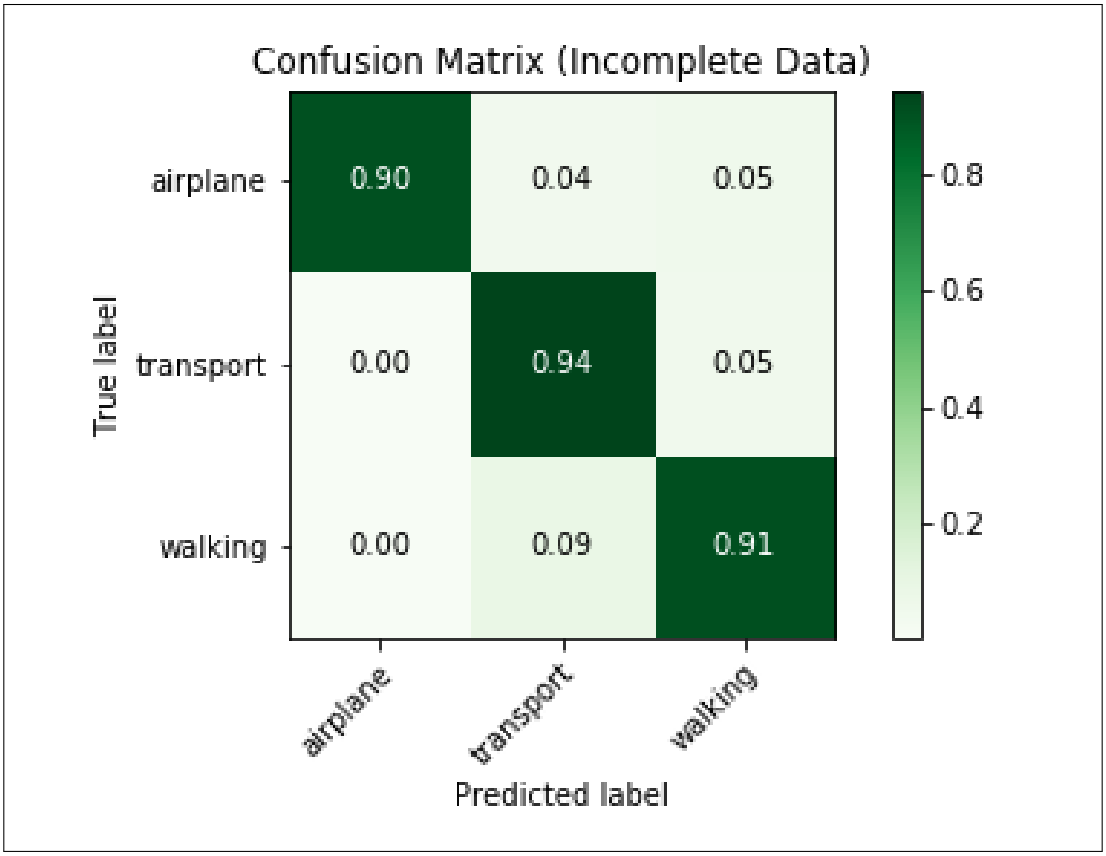

Figure 4.4: Confusion Matrix for 3-Class Incomplete Test Set

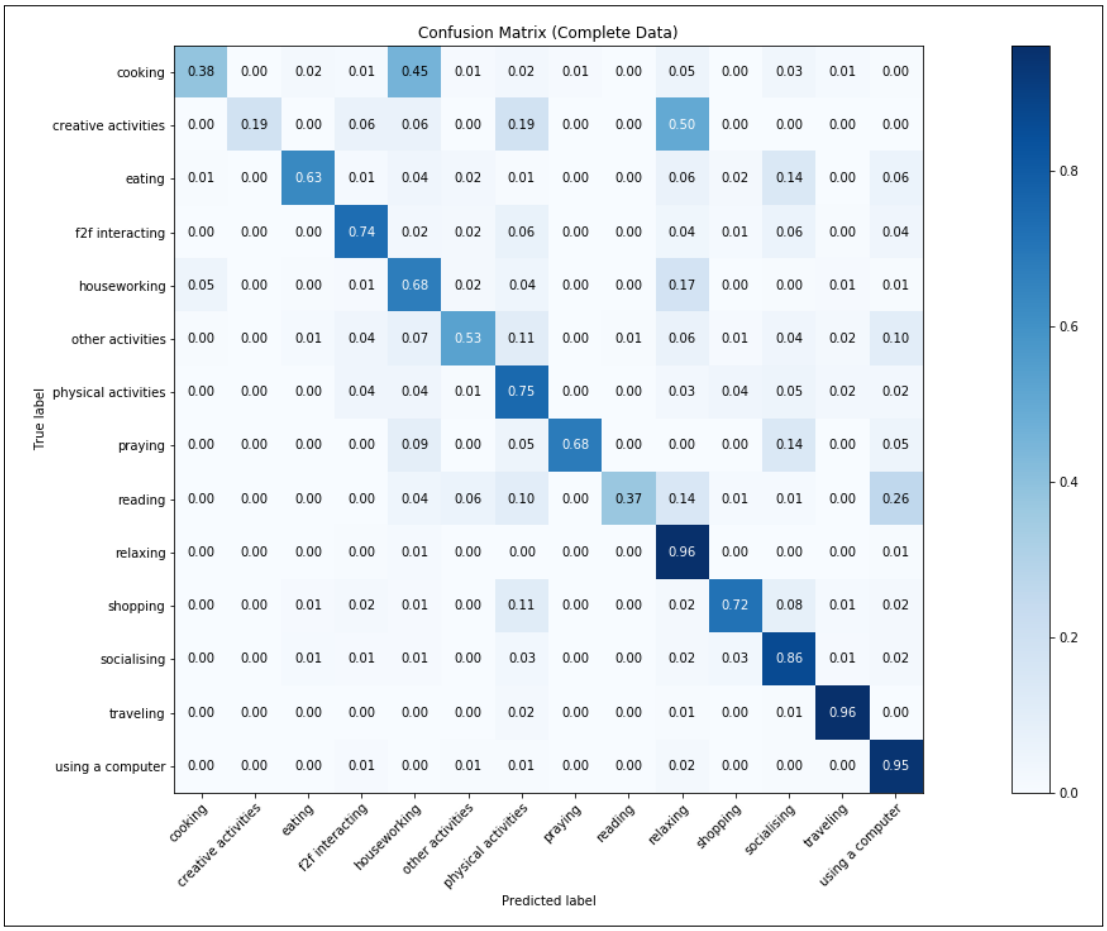

Figure 4.5: Confusion Matrix for 16-Class Complete Test Set

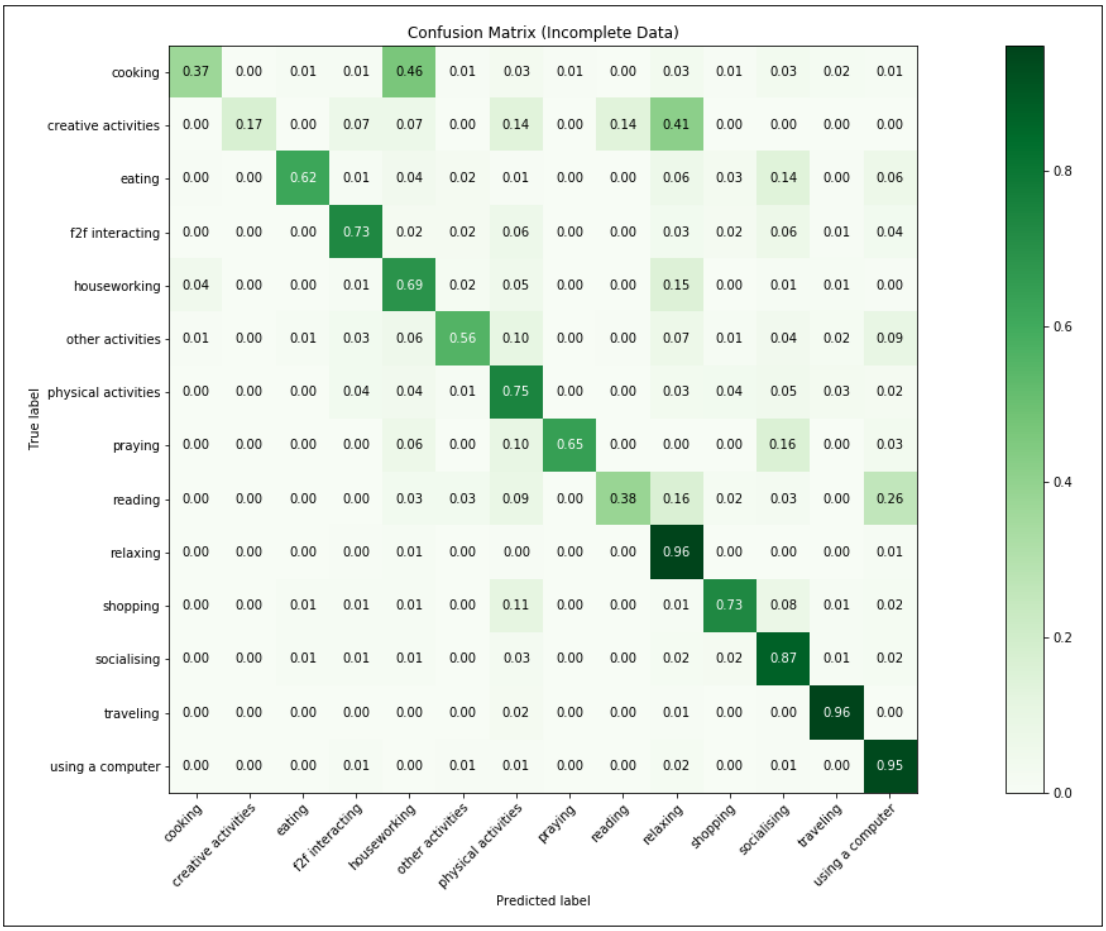

Figure 4.6: Confusion Matrix for 16-Class Incomplete Test Set

CHAPTER 5

CONCLUSION

In this thesis study, the NTCIR Lifelog dataset is described and analyzed from the perspective of relationships between images and activity definitions. The original version of the dataset has 3 activity classes, namely airplane, transportation and walking. With the intuition that these activities are inadequate for describing ligeloggers' whole life, we make an effort to manually classify a subset of the dataset into more generalized 16 activity classes. Next, we carry out a series of experiments on both the original 3-class data and manually classified 16-class data.

In order to classify minute records from lifelog data into activity classes, the images are used with image annotations, with reference to the text-based classification method, and as original images, using the image classification method. Different classification and learning algorithms are trained on input images and annotations. Next, we propose a combined artificial neural network, which takes both images and image annotations as input and learns from the two dimensions together for classification.

In both of the two different versions of the dataset, there are several records missing image or text dimension. With the purpose of handling missing values in the data, we propose a masked loss function to be used at intermediate levels of the combined learning model. We feed zero-vector as input in place of missing dimensions. While we use categorical cross-entropy loss to evaluate the final loss, our masked loss function calculate the loss at intermediate levels by ignoring the effect of the record if it is a zero-vector.

Performance results show that using original images results in better performance than using annotations for the classification problem. In other words, having color

images and image annotations available on hand, better classification of images is obtained by using color images itself on the NTCIR Lifelog dataset.

The combined model, which learns from images and annotations together on a single model, usually performs better than both of the sub-components which learn separately from a single dimension. However, the presence of missing input modalities negatively affect the performance of the combined model by limiting the size of the training data.

Finally, we propose a masked loss function which makes it possible to learn from multimodal data in the presence of missing values in some of the dimensions. The proposed multi-loss combined model is capable of learning from image and text data simultaneously even when there are missing values within the data. The prediction performance of the proposed model is better the naive activity prediction model which can be used in presence of missing values. It also shows better performance than well-known learning models which learn from single dimension.

We suggest that the definitions of the 16 activity classes could be revised, and the body measurements and location data could be included in the learning model. These numerical observations definitely hide lots of valuable information inside, which should be investigated and revealed in some future studies.

REFERENCES

[1] 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*, 2009.

[2] S. Mann, "Wearable Computing: A First Step Toward Personal Imaging," *Computer*, 1997.

[3] P. Belimpasakis, K. Roimela, and Y. You, "Experience Explorer: A Life-Logging Platform Based on Mobile Context Collection," in *NGMAST 2009 - 3rd International Conference on Next Generation Mobile Applications, Services and Technologies*, 2009.

[4] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal, "NTCIR Lifelog: The First Test Collection for Lifelog Research," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, (New York, NY, USA), pp. 705–708, ACM, 2016.

[5] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, R. Gupta, R. Albatal, and D.-T. Dang-Nguyen, "Overview of NTCIR-13 Lifelog-2 Task," in *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*, 2017.

[6] K. Belli, E. Akbaş, and A. Yazici, "Activity learning from lifelogging images," in *Artificial Intelligence and Soft Computing* (L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J. M. Zurada, eds.), (Cham), pp. 327–337, Springer International Publishing, 2019.

[7] L. Xia, Y. Ma, and W. Fan, "VTIR at the NTCIR-12 2016 Lifelog Semantic Access Task," *The 12th NTCIR Conference, Evaluation of Information Access Technologies*, 2016.

[8] B. Safadi, P. Mulhem, G. Quénot, and J.-P. Chevallet, "LIG-MRIM at NTCIR-12 Lifelog Semantic Access Task," *The 12th NTCIR Conference, Evaluation of Information Access Technologies*, 2016.

[9] H.-L. Lin, T.-C. Chiang, L.-P. Chen, and P.-C. Yang, "Image Searching by Events with Deep Learning for NTCIR-12 Lifelog," *The 12th NTCIR Conference, Evaluation of Information Access Technologies*, 2016.

[10] J. Lin, A. G. del Molino, Q. Xu, F. Fang, V. Subbaraju, J. H. Lim, L. Li, and V. Chandrasekhar, "VCI2R at the NTCIR-13 Lifelog-2 Lifelog Semantic Access Task," in *Proceedings of NTCIR13*, 2017.

[11] S. Yamamoto, T. Nishimura, Y. Akagi, Y. Takimoto, T. Inoue, and H. Toda, "PBG at the NTCIR-13 Lifelog-2 LAT, LSAT, and LEST Tasks," in *Proceedings of NTCIR13*, 2017.

[12] D. T. Dang-Nguyen, L. Piras, M. Riegler, G. Boato, L. Zhou, and C. Gurrin, "Overview of ImageCLEF Lifelog 2017: Lifelog Retrieval and Summarization," in *CEUR Workshop Proceedings*, 2017.

[13] R. Rawassizadeh, M. Tomitsch, K. Wac, and A. M. Tjoa, "UbiqLog: A Generic Mobile Phone-Based Life-Log Framework," *Personal and Ubiquitous Computing*, 2013.

[14] R. Rawassizadeh, E. Momeni, C. Dobbins, P. Mirza-Babaei, and R. Rahnamoun, "Lesson Learned from Collecting Quantified Self Information via Mobile and Wearable Devices," *Journal of Sensor and Actuator Networks*, 2015.

[15] A. Amlinger, *An Evaluation of Clustering and Classification Algorithms in Life-Logging Devices*. PhD thesis, 2015.

[16] A. G. Del Molino, B. Mandal, J. Lin, J. H. Lim, V. Subbaraju, and V. Chandrasekhar, "VC-I2R@ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization," in *CEUR Workshop Proceedings*, 2017.

[17] M. Bolaños, M. Dimiccoli, and P. Radeva, "Toward Storytelling From Visual Lifelogging: An Overview," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 77–90, 2017.

[18] T.-D. Truong, T. Dinh-Duy, V.-T. Nguyen, and M.-T. Tran, "Lifelogging Retrieval Based on Semantic Concepts Fusion," in *Proceedings of the 2018 ACM Workshop on the Lifelog Search Challenge*, (New York, NY, USA), pp. 24–29, ACM, 2018.

[19] F. Ben Abdallah, G. Feki, A. Ben Ammar, and C. Ben Amar, "Multilevel Deep Learning-Based Processing for Lifelog Image Retrieval Enhancement," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1348–1354, Oct 2018.

[20] M. Dimiccoli, A. Cartas, and P. Radeva, "Activity Recognition from Visual Lifelogs: State of the Art and Future Challenges," in *Multimodal Behavior Analysis in the Wild*, 2018.

[21] T. Baltrusaitis, C. Ahuja, and L. P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," 2019.

[22] Y. Liu, L. Liu, Y. Guo, and M. S. Lew, "Learning Visual and Textual Representations for Multimodal Matching and Classification," *Pattern Recognition*, vol. 84, pp. 51–67, dec 2018.

[23] T. Zahavy, A. Krishnan, A. Magnani, and S. Mannor, "Is a Picture Worth a Thousand Words? A Deep Multi-Modal Architecture for Product Classification in E-Commerce," *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] E. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," *WSEAS Transactions on Computers*, vol. 4, pp. 966–974, 08 2005.

[25] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, 1995.

[26] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Lecture Notes in Computer Science*, 1998.

[27] J. Nazzal, I. M. El-Emary, and S. A. Najim, "Multilayer Perceptron Neural Network (MLPs) for Analyzing the Properties of Jordan Oil Shale," *World Applied Sciences Journal*, vol. 5, 01 2008.

[28] A. Karpathy, “Image Classification.”

[29] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[30] A. Karpathy, “Convolutional Neural Networks (CNNs / ConvNets).”

[31] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *CoRR*, vol. abs/1512.03385, 2015.