IMAGE-LEVEL MICRO-GESTURE CLASSIFICATION VIA VISUAL-TEXT CONTRASTIVE LEARNING

LUT Computer Vision and Pattern Recognition

School of Engineering Science

BM40A0801 Machine Vision and Digital Image Analysis,

Practical Assignment: Image-level Micro-Gesture Classification

2024

Group Members

Andry Andriamananjara (000593633):

Code Implementation (AlexNet)

Report (Introduction and Literature Review)

Socrates Waka Onyando (001690286):

Code Implementation (Visual-Text Contrastive Learning)

Report (Method Description, Experiments & Results, and Discussion)

CONTENTS

1 INTRODUCTION

1.1 Background

In the domain of affective computing, micro-gestures play a pivotal role in revealing people's true inner emotional states [1]. Unlike ordinary gestures, micro-gestures emanate from subconscious mechanisms of the human cognitive system; highlighting their uncontrolled nature and consequently their adeptness at revealing a more accurate picture of one's emotional state as compared to other modalities such as speech, facial expressions, posture and ordinary gestures [1]. Figure 1 illustrates several examples of micro-gestures.

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

Figure 1. Examples of MGs: (a) Crossing fingers; (b) Touching jaw; (c) Touching neck; (d) Playing or adjusting hair [1].

Given the importance of recognizing such microgestures, building systems that are able to discriminate one micro-gesture class from the other, and more broadly one emotional state from the other, is crucial. However, this is a difficult task considering that different gesture classes can be visually similar. This is evident in Figure 1 where two micro-gesture examples, "touching jaw" and "touching neck", are alike. Moreover, there exists diverse movements within the same micro-gesture class [2] as well as the fact that micro-gestures may be subtle and limited to a smaller area of the body.

1.2 Objectives

The main objective of this project is to develop a classification model capable of identifying and categorising micro-gestures at the image level. The provided dataset of images, which has been extracted from the iMiGUE video dataset, comprises 32 distinct classes of micro gestures. The goal herein is to train a model that can effectively classify each image into one of these classes.

1.3 Work Division

For this project, each member of the group explored various models independently with the aim of finding the most effective one. The final model selected for implementation (visual-text contrastive learning) was chosen based on the performance observed during experimentation. The report was collaboratively written by both group members, with responsibilities divided into two main sections: 1) Introduction and Literature Review and 2) Method Description, Experiments & Results, and Discussion.

2 LITERATURE REVIEW

The issue at hand pertains to the classification of images. Numerous methods have been developed to address such image classification problems. Among them are AlexNet [3], one of the well-known Convolutional Neural Network (CNN) methods, and advanced deep learning techniques such as Contrastive Language-Image Pretraining (CLIP) [4].

AlexNet [3], presented by Alex Krizhevsky, was a pioneering work in the current trend toward convolutional networks and deep learning. It emerged as the top CNN model during the ImageNet Large-Scale Visual Recognition (ILSVRC) competition in 2012. Krizhevsky et al. demonstrated that networks with ReLU can learn faster and more precisely than other activation functions such as $sigmoid$ and $\tanh$ more precisely [3]. ReLU addresses the neuron saturation problem by preventing highly positive and negative values from resulting in derivatives close to zero. Additionnally, AlexNet addressed the overfitting difficulty by preprocessing the dataset, applying data augmentation techniques, and implementing dropout for regularization in the fully connected layers. AlexNet accepts a 3D or colored image as input and applies all given convolutional operations to output the most probable class for the given input. Furthermore, it can also be applied in specific sports videos by predicting the given shot [5].

CLIP [4] is a model capable of understanding both text descriptions and images. The model utilizes a pre-trained neural network that can comprehend the relationship between images and text. It is based on three main approaches: (1) Multi-modal training (the model combines images and text simultaneously); (2) Contrastive training objective (the goal is to recognize the exact pair of text and images among all given pairs); and (3) zero-shot learning (the model can understand visual concepts by using natural language guidance) [4]. CLIP, illustrated in Figure 2, works according to the following flow: a text encoder (transformer) encodes the texts; an image encoder based on ResNet or ViT encodes images; contrastive training then takes place where similar text and image representations are brought closer together, while unrelated ones pushed further apart; and finally zero-shot prediction where CLIP is able to classify an image without being explicitly trained on image-label pairs [4].

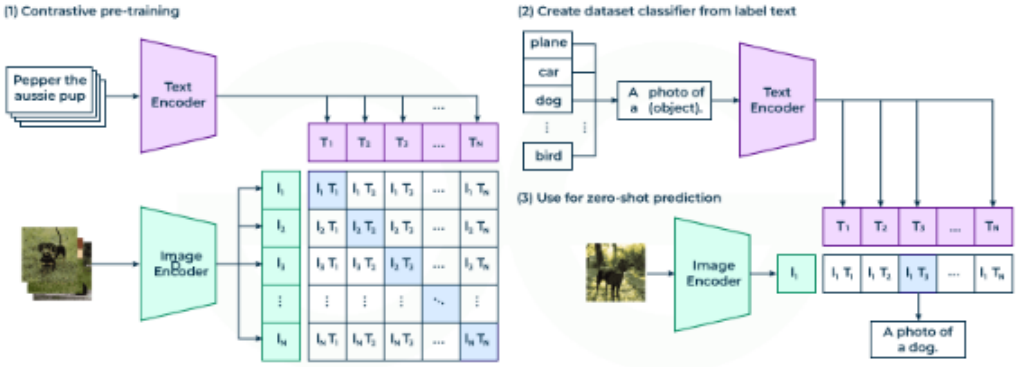

Figure 2. CLIP Architecture [4].

3 METHOD DESCRIPTION

Our proposed method, illustrated in Figure 3, is based on OpenAI's model, CLIP [4], which allows for multimodal learning. Our implementation closely follows the CLIP implementation by Shariatnia [6] with modifications to the loss function and a few parameters. As already mentioned, accurate micro-gesture recognition is limited by the close visual similarity between different micro-gesture classes. Nonetheless, crucial label information that accompanies the visual information may help in the identification of micro-gesture instances. A multi-modal model, such as CLIP, allows for the utilisation of both textual and visual modalities.

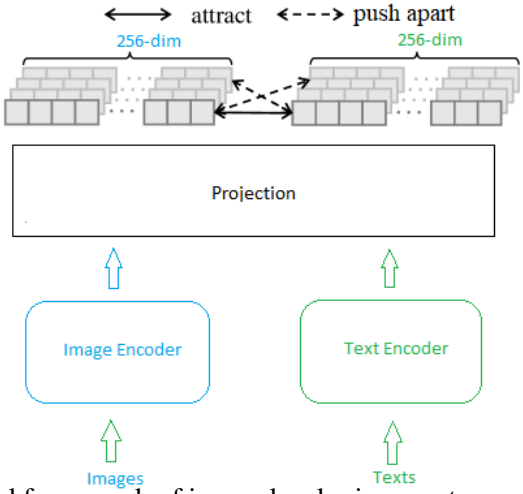

Figure 3. Proposed framework of image-level micro-gesture recognition model

3.1 Image and Text Encoder

In our set-up, a pre-trained ResNet50 model is used to encode images into fixed-size (2048) vectors while a pre-trained DistilBERT model is used to encode the text information (labels) into fixed-size vectors of size 768 [6].

3.2 Projection Head

A Multi-Layer Perceptron is then used to project the text and image embeddings into a lower-dimensional space of size 256 where contrastive learning can take place [6].

3.3 Contrastive Objective

CLIP uses a cross entropy loss function to maximise the similarity between embeddings of related pairs (matching text and image embeddings) while at the same time minimising the similarity between unrelated pairs (mismatched text and image embeddings) [4]. However, our approach utilises a symmetrized variant of the Kullback–Leibler (KL) divergence loss called Jensen–Shannon (JS) divergence, defined by Equation 1 [7,8], given that it yielded better results in our experiments.

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{1}$$

where M is a mixture distribution of P (dot product similarity between image embeddings and text embeddings) and Q (dot product similarity between text embeddings and image embeddings); which in our case is the target square matrix with 1s in case of correct image-text pairing and 0 otherwise.

4 EXPERIMENTS & RESULTS

4.1 Experiments

4.1.1 Data Pre-processing

In our experimental setup, we partitioned the dataset into train and validation sets using an 80-20 split ratio. This partitioning was conducted via stratified sampling to guarantee that the distribution of classes remains balanced across both partitions. A preprocessing step was applied to resize each image to dimensions of 224 by 224, which are the input dimensions of our image encoder ResNet50. Subsequently the images were normalised before being inputted into the image encoder. Notably, we utilised RGB modality only in our set-up.

4.1.2 Training

Throughout the training process, both the image and text encoders were trained jointly for a maximum of 250 epochs. At the start, we initialised the learning rates for the image and text encoders at 0.0001 and 0.00001, respectively. However, to enhance convergence, we set up our implementation to adjust these learning rates utilising a learning rate scheduler which was configured to update after every batch. Notably, we experimented with a cross-entropy loss function as implemented by Shariatnia [6] as well as a JS divergence loss function.

4.1.3 Validation

After every training epoch, we evaluated the model's performance on the validation set, considering both the computed validation loss as well as the top-1 and top-5 accuracies. Upon achieving better performance on the validation set, the learnable parameters (weights and biases) of the model's layers were continually saved for potential further inference. This validation set-up enabled us to monitor the model's generalisation to new data and gain insights into its overall performance.

4.2 Results

Table 1 reports the classification accuracies of the two implementations of our approach on the validation set. It is quite obvious that the implementation with the JS divergence loss function gives the superior performance as compared to the cross-entropy loss function implementation adapted from Shariatnia [6].

Table 1. Comparison of MG recognition accuracy with our method using different loss functions

Methods		Modality	Accuracy (%)	
	Loss Function		Top-1	Top-5
Visual-text Contrastive Learning	Cross-entropy	RGB	22.54	27.71
	JS divergence	RGB	**81.13**	**94.33**

5 DISCUSSION

Recent developments in machine learning have seen the rise of multimodal models that can combine different modalities, an example of which is CLIP [4]. These models, which leverage the complementary information that accompanies data (text accompanying images in our case), have demonstrated superiority over their unimodal counterparts [4, 9]. Our choice of a multimodal model was influenced by this assertion in which case we opted for a visual-text contrastive framework to a visual-only model such as AlexNet after a preliminary evaluation of their respective performance on the provided dataset. Regardless, this should not discredit the effectiveness of vision-only models as our comparison was limited to a small selection of models.

In our set-up for the visual-text contrastive learning, we experimented with two loss functions in which case the JS divergence loss function provided superior performance to the cross-entropy loss function. We attribute the sub-optimal performance of the cross-entropy loss function to the construction of the target matrix as implemented by Shariatnia [6] and influenced by the original CLIP implementation by Radford et al. [4]. Shariatnia uses an identity matrix (1s in the main diagonal and 0s elsewhere) of size equal to the batch size [6]. However, this is not suited for our case given that the original CLIP implementation assumed strictly N correct pairs and $N^2 - N$ incorrect pairs in a batch [4]; that is, CLIP assumes each text/label is matched to one and only one corresponding image and vice versa. In our case, there is a possibility of multiple images in a batch matching one micro-gesture class. A KL divergence loss function is therefore the more suitable loss function for comparing the distributions of the target and predicted dot product similarity scores. However, in its plain form, the KL divergence is unbounded and asymmetric; and may therefore result in instabilities during optimization or poor generalisations [8]. The JS divergence is thus the better alternative since it is bounded and symmetric [8].

REFERENCES

[1] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10626–10637, 2021.

[2] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. SMG: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, 2023.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[5] Rabia A. Minhas, Ali Javed, Aun Irtaza, Muhammad Tariq Mahmood, and Young Bok Joo. Shot classification of field sports videos using alexnet convolutional neural network. *Applied Sciences*, 9(3), 2019.

[6] Moein Shariatnia. Simple implementation of OpenAI CLIP model in PyTorch, 2021.

[7] Xia Huang and Kai Fong Ernest Chong. GenKL: An iterative framework for resolving label ambiguity and label non-conformity in web images via a new GENeralized KL divergence. *International Journal of Computer Vision*, 131(11):3035–3059, 2023.

[8] Ponkrshnan Thiagarajan and Susanta Ghosh. Jensen-Shannon divergence based novel loss functions for Bayesian neural networks. *arXiv preprint arXiv:2209.11366*, 2022.

[9] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pages 4348–4380, 2023.