**COMP9444 Project Summary**

<Automated Yoga Pose Classification>

Hesong Zhang(z5532689)

Renfei Yu (z5546743)

Wenxuan Wang (z5464319)

Renhua Luo (z5408176)

Yi Zhao (z5530832)

1. **Introduction**

Yoga is an ancient fitness activity that enhances both the body and mind through various poses. However, identifying each yoga pose accurately is important to ensure right practice, effective monitoring, and timely correction.

Here, we propose a solution: use deep learning technology to automate the yoga pose recognition. This can increase efficiency and ensure the accuracy and real-time nature of the classification.

Our project use a dataset called Yoga-82, which contains a large collection of image covering 82 different yoga poses. Our goal is to develop a neural network model that can identify complex yoga poses in real-time, even when poses are observed from different angles.

In summary, through this project, We aim to promote the technical advancement of yoga practice, so that people who love yoga can practice yoga more safely and scientifically, and also increase the application of computer vision in the health and fitness field.

1. **Related Work**

Deep learning techniques are becoming increasingly common in human pose estimation and activity recognition tasks, including yoga pose classification. A large number of studies have shown that convolutional neural network (CNN) and recurrent neural network (RNN) have remarkable effects in analyzing human motion sequence data.(LAKSHMI and CHETAN, 2021)

For example, Younis, G Kaneriya et al. used pre-trained CNN architecture (such as ResNet and VGG) for transfer learning to classify yoga poses from static images, which significantly improved the classification accuracy compared with traditional manual feature extraction methods. (Younis Ayesha et al., 2022)

**Limitations of existing studies:**

Despite these advances, existing approaches still have the following limitations:

Generalization ability of datasets: Many models are trained and evaluated on limited datasets, and are prone to overfitting and poor generalization performance when there are new datasets with different light conditions, angles, or practitioner diversity.

Class imbalance problem: There is often class imbalance in yoga pose data sets, which reduces the performance of the model on low-frequency categories.

**Future direction:**

To address the above limitations, future research could focus on the following improvements:

Data Enhancement vs. Synthetic data: Using advanced data enhancement techniques or generating synthetic data through generative adversarial networks (Gans) can address class imbalances and improve the robustness of models over diverse datasets.

Lightweight model: Develop a lightweight and efficient deep learning architecture to help enable real-time yoga pose recognition on edge devices.

Time modeling: Using advanced time modeling techniques, you can improve the performance on dynamic yoga sequences.

Domain adaptation: The use of domain adaptation strategies helps models achieve better generalization performance on previously unseen data sets.

By addressing these challenges, the field can move toward more robust, accurate, and practical yoga pose recognition solutions, creating a solid foundation for fitness, healthcare, and personalized training applications.

1. **Methods**
2. Model Architecture Design
   1. The MobileNetV3 architecture implements a progressive neural network structure transitioning from 500 to 250 neurons, culminating in 82 classification nodes corresponding to distinct yoga poses. The implementation incorporates the Adam optimization algorithm with a learning rate of 1e-4, augmented by batch normalization and dropout regularization (rate = 0.3) to enhance model generalization capabilities.
   2. The proposed Custom CNN+MediaPipe framework presents a novel integration of 3D convolutional neural networks with the MediaPipe infrastructure. The methodology involves transformation of 2D input images into 32³ voxel representations, subsequently processed through the custom YogaConv3d architecture. This architecture comprises three sequential convolutional layers with increasing filter depths (64, 128, 256), implementing dropout regularization and batch normalization protocols after each convolution operation.
   3. The MoveNet implementation focuses on anatomical precision through the identification and tracking of 17 critical body landmarks. The methodology incorporates advanced normalization techniques for scale and rotational invariance, coupled with temporal trajectory smoothing and confidence-based threshold mechanisms to ensure robust pose estimation accuracy.
   4. The Xception architecture implementation follows a bifurcated training methodology. The initial phase employs transfer learning, utilizing pre-trained weights with custom classification layers (500→250→82 units) optimized via Adam algorithm (learning rate: 0.001). The subsequent phase implements systematic layer unfreezing with refined optimization parameters (learning rate: 0.0001) utilizing SGD with momentum.
   5. The DenseNet201 implementation leverages pre-trained ImageNet parameters while maintaining frozen weights in the initial 100 layers. The architecture incorporates a custom classification mechanism utilizing Global Average Pooling, followed by a dense layer (1024 units) with dropout regularization. The training protocol extends over 100 epochs with dynamic learning rate adjustment.
3. Training Infrastructure

The experimental implementation utilizes NVIDIA GPU infrastructure with TensorFlow 2.x framework integration. The training progression underwent continuous monitoring through TensorBoard for performance optimization and validation.

1. Architectural Selection Rationale

The selection of diverse architectural approaches facilitates a comprehensive evaluation of pose classification methodologies. The MobileNetV3 architecture addresses deployment efficiency requirements, while the CNN+MediaPipe framework enables sophisticated spatial analysis. The MoveNet architecture provides precise anatomical reference points, complemented by the Xception model's enhanced feature extraction capabilities through transfer learning. The DenseNet201 architecture contributes to advanced feature reuse through dense connectivity patterns.

1. **Experimental Setup**

The image dataset is obtained from a paper titled “Yoga-82: A New Dataset for Fine-grained Classification of Human Poses”: https://neurohive.io/en/news/yoga-82-new-dataset-with-complex-yoga-poses/。 The dataset, Yoga-82, was developed by researchers Manisha Verma, Sudhakar Kumawat, Yuta Nakashima, and Shanmuganathan Raman, as part of their work at Osaka University, Japan, and Indian Institute of Technology Gandhinagar, India. The dataset was developed to address the challenges of fine-grained classification of yoga poses, with hierachical structures to capture various body positions and pose variations .

The dataset is downloaded from a Google drive with a link contained in the webpage mentioned above. The dataset is intended for non-commercial research and educational purposes.

Yoga-82 contains over 28,000 images of yoga poses across 82 unique classes, each representing a different yoga pose/body position variation. The dataset includes poses from various camera angle and in different environments, with a mix of clean and complex backgrounds. Image formats include JPG and PNG, and of various resolutions.

Yoga-82 uses a 3-level hierarchical structure where:

1. Level 1 represents 6 main categories, such as standing, sitting, balancing, inverted, reclining, and wheel

2. Level 2 represents 20 subcategories beneath level 1 main categories, representing variations such as leg position or orientations (e.g. forward bend or side bend for the Level 1 standing position etc), and

3. Level 3 is the specific yoga pose and there are 82 poses in total

Note that in this project, we focused only on level 3 yoga poses and treat each pose as a separate class. Therefore, we have 82 classes for the pose prediction problem.

1. **Results**

### Key Experimental Results

The evaluation of the five models on the Yoga-82 dataset yielded the following accuracy scores:

* 1. MoveNet: 88.22%
  2. Xception: 79.63%
  3. DenseNet201: 79.00%
  4. 3D CNN + MediaPipe: 77.29%
  5. CNN + MobileNetV3: 72.67%

These results demonstrate a clear performance hierarchy, with MoveNet outperforming the other models by a significant margin.

**Key evaluation findings**

**Outstanding performance of MoveNet:**

MoveNet achieved the highest accuracy (88.22%) on the Yoga-82 dataset, indicating its excellent performance in capturing specific postural features. This is consistent with the design goal of its lightweight and efficient attitude estimation model.

**Similar performance of Xception and DenseNet201:**

Xception (79.63%) and DenseNet201 (79.00%) performed equally. Both excel at feature extraction, but lack the special design for pose estimation that MoveNet does.

**3D CNN + MediaPipe timing analysis capabilities:**

3D CNN uses timing information combined with MediaPipe and has an accuracy rate of 77.29%. Despite the introduction of the complexity of the timing dimension, its accuracy has not been significantly improved.

**CNN + MobileNetV3's efficiency first strategy:**

MobileNetV3 has the lowest accuracy (72.67%), mainly because its lightweight architecture prioritizes computational efficiency over accuracy.

**Deployment consideration**

The experimental results show that MoveNet, with its high accuracy and lightweight design, is the most promising model for practical applications and is particularly suitable for real-time attitude recognition tasks. However, generalization capabilities need to be validated on diverse data sets and optimized to fit edge device requirements before deployment. Xception and DenseNet201 can be used when resources are available, while 3D CNN + MediaPipe is suitable for scenarios that require timing analysis.

**Comparison with literature and state-of-the-art technology**

MoveNet has an advantage in real-time applications with its efficient architecture and outstanding performance (88.22%); 3D CNN + MediaPipe provides a compromise solution for time series modeling; MobileNetV3 focuses on computational efficiency. Overall, MoveNet competes in its class with a balance of efficiency and accuracy, while other models offer flexibility for specific scenarios.

1. **Discussion**

1. Loss and Accuracy

These charts illustrate the training and validation loss and accuracy over epochs for different models.

* Training Accuracy (MoveNet vs. MobileNetV3): The MoveNet model converges faster on the training set,
* Validation Accuracy (MoveNet vs. MobileNetV3): MovNet’s validation accuracy reaches near convergence at earlier epochs, indicating good generalization to new data.
* Training Loss (MoveNet vs. MobileNetV3): MoveNet’s training loss drops quickly and stabilizes, demonstrating high learning efficiency
* MoveNet: Reaches high accuracy quickly with stable training, showing lower validation loss, indicating a well-generalized model.
* MobileNetV3: Moderate accuracy increases over epochs, with a relatively stable gap between training and validation accuracy.
* DenseNet201: Strong improvement in accuracy but shows a higher validation loss than other models, possibly indicating overfitting.

In summary, MoveNet demonstrates better generalization capabilities and lower validation loss, making it preferable choices for selection. The other models also perform well but may require further adjustments to optimize their validation performance.

2. Performance Analysis

* MobileNetV3 vs MoveNet-based Pose Model：MobileNetV3 is compact for mobile use but less precise in posture classification than MoveNet, which excels at real-time keypoint detection.
* CNN+MediaPipe: Offers versatile and accurate posture classification through effective keypoint detection, suitable for general pose estimation tasks
* Xception vs DenseNet201：Xception is efficient for general classification, whereas DenseNet201 captures complex postures better but isn’t suitable for real-time use.
* MoveNet-based Pose Model vs Xception & DenseNet201：MoveNet is ideal for keypoint-based posture tracking, while Xception and DenseNet201 are better suited for traditional classification tasks.
* MobileNetV3 vs CNN+MediaPipe: MobileNetV3 is good for general lightweight tasks, whereas CNN+MediaPipe performs better on precise keypoint detection for pose estimation.

1. **Conclusions**

* MoveNet-based Pose Model is the best choice for posture classification and real-time tracking, suitable for dynamic detection.
* CNN+MediaPipe provides accurate keypoint detection for general pose estimation, adaptable but less specialized for specific postures.
* MobileNetV3, Xception, and DenseNet 201 perform well in general classification tasks but are not ideal for specialized posture analysis, especially in real-time applications.

Strengths of MoveNet:

* Scalability and Efficiency: MoveNet is lightweight, real-time, and ideal for mobile and edge deployment.
* High Accuracy: Delivers precise human key point detection, tailored for tasks needing quick, accurate pose estimation.

Limitations and Weaknesses of MoveNet:

* Pose Complexity: Struggles with complex or multi-person poses, limiting use in crowded scenes.
* Occlusions: Accuracy drops when body parts are occluded or partially visible.
* Customization: Limited flexibility for fine-tuning compared to more configurable models.

**Reference**

LAKSHMI, J.V.N. and CHETAN, K.B. (2021). YOGA POSE CLASSIFICATION USING RESNET OF DEEP LEARNING MODELS. i-manager’s Journal on Computer Science, 9(2), p.29. doi:https://doi.org/10.26634/jcom.9.2.18464.

Younis Ayesha, Li, Q., Nyatega Charles Okanda, Nyatega Charles Okanda, Adamu Mohammed Jajere and Bello, K.H. (2022). Brain Tumor Analysis Using Deep Learning and VGG-16 Ensembling Learning Approaches. Applied Sciences (Web), [online] 12(14), p.7282. Available at: https://jglobal.jst.go.jp/en/detail?JGLOBAL\_ID=202302285936992654 [Accessed 17 Nov. 2024].

Garg, S., Saxena, A. and Gupta, R. (2022). Yoga pose classification: a CNN and MediaPipe inspired deep learning approach for real-world application. Journal of Ambient Intelligence and Humanized Computing. doi:https://doi.org/10.1007/s12652-022-03910-0.

vencerlanz09 (2022). Yoga Pose Classification using CNN + MobileNetV3. [online] Kaggle.com. Available at: https://www.kaggle.com/code/vencerlanz09/yoga-pose-classification-using-cnn-mobilenetv3/notebook [Accessed 17 Nov. 2024].

A. K. Rajendran and S. Chakkaravarthy Sethuraman, “DensePoseCompare: A Comparative Study of DenseNet Models in Yoga Pose Classification,” 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), pp. 471–480, Apr. 2024, doi: https://doi.org/10.1109/icc-robins60238.2024.10534022.

tensorflow (2015). tensorflow/tensorflow/lite/g3doc/tutorials/pose\_classification.ipynb at master · tensorflow/tensorflow. [online] GitHub. Available at: https://github.com/tensorflow/tensorflow/blob/master/tensorflow/lite/g3doc/tutorials/pose\_classification.ipynb [Accessed 17 Nov. 2024].

utkarshsaxenadn (2022). Yoga Pose Detection | Xception | Acc : 75%. [online] Kaggle.com. Available at: https://www.kaggle.com/code/utkarshsaxenadn/yoga-pose-detection-xception-acc-75 [Accessed 17 Nov. 2024].