FaceRWKV - Exploring RWKV for Facial Expression Recognition

Lukas Vierling, Christian A. Pesch, Oscar R. Cortina
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{lvierling, capesch, orcortina}@connect.ust.hk

Abstract

Facial expression recognition is crucial for emotion analysis and human-computer interaction. This project introduces a novel architectural design utilizing the RWKV (Receptance Weighted Key Value) for vision tasks. Inspired by the Vision Transformer model, we investigate the effective integration of RWKV after sequencing the input data. Our hybrid design combines RWKV with a pre-trained ResNet feature map, with promising results in comparison to current state-of-the-art transformer-based models in terms of accuracy and complexity. We conduct an ablation study to confirm the significance of our design choices and the necessity of incorporating RWKV. Our findings provide valuable insights into enhancing accuracy and performance in facial expression recognition, benefiting emotion analysis and human-computer interaction. The project's repository can be found at https://github.com/lukasVierling/COMP4771.

1. Introduction

Facial Expression Recognition (FER) is a fundamental task in computer vision that involves accurately classifying facial expressions based on visual cues. It has significant applications in areas such as emotion analysis, human-computer interaction, and affective computing. The ability to recognize and understand human emotions is essential for creating more empathetic and responsive AI systems. Transformers have revolutionized Natural Language Processing (NLP) tasks and achieved remarkable performance. However, their memory and computational complexity scales quadratically with sequence length, which can pose challenges when applied to vision tasks. The two-dimensionality of images often results in large sequence lengths, making transformers less efficient in this context. On the other hand, RNNs offer linear scaling in memory and computational requirements but struggle to match the performance of transformers. In this project, we explore the potential of the Receptance Weighted Key Value (RWKV) architecture for FER. RWKV is an existing architecture that provides an alternative to transformers, offering potential improvements in memory and time efficiency. Our objective is to investigate how incorporating RWKV into our own FER architecture can enhance performance. We hypothesize that utilizing RWKV for sequencing the input data can effectively capture relevant visual cues and improve the accuracy of facial expression classification. Inspired by the success of the Vision Transfromer (ViT) model, we explore the integration of RWKV after sequencing the input data using various methods. To evaluate the effectiveness of our proposed architecture, we compare its performance with current State-Of-The-Art (SOTA) in FER. Additionally, we conduct an ablation study to validate the necessity of our design choices and demonstrate the potential benefits of incorporating RWKV. While we do not present results on memory and runtime in this report, we hypothesis that RWKV can decrease memory usage and runtime complexity of modern vision architectures based on its theoretical background. We aim to contribute to the ongoing research in this field by exploring the potential of RWKV as a viable alternative for vision tasks like FER. By leveraging the strengths of RWKV in sequencing the input data, we aim to improve the accuracy and effectiveness of FER models. The results of our study reveal that RWKV exhibit superior performance compared to established pre-trained convolutional neural networks, accomplishing this feat with an impressively minimal number of model parameters. It is noteworthy, however, that the current model dimensions of our approach do not enable it to surpass the current SOTA performance of transformer based models. Furthermore we demonstrated that RWKV can be successfully leveraged for computer vision tasks, specifically for FER, contributing to the development of more efficient and accurate emotion analysis and human-computer interaction systems.

2. Related Work

FER with RWKV architectures remains an unexplored area in published literature, signaling an opportunity for novel exploration in this domain.

ViT: Dosovitskiy *et al*. [2] showcased the potential of transformers in vision tasks by directly applying a pure transformer architecture to sequences of image patches. ViT demonstrated exceptional performance on various image classification benchmarks (ImageNet, CIFAR-100, VTAB) with significantly reduced computational requirements compared to conventional Convolutional Neural Networks (CNN).

RWKV Architecture: Addressing the limitations of memory and computational complexity in transformers, Guo *et al*. [3] introduced the RWKV architecture. RWKV combines the efficiency of parallelizable training from Transformers with the inference efficiency of RNNs, offering constant computational and memory complexity during inference. Experimental results showcase RWKV's competitive performance similar to equivalently sized Transformers, presenting a promising avenue for efficient sequence processing. In our project we used the RWKV-v1 codebase provided by the authors in their GitHub repository [1].

The paper ”Context-Aware Emotion Recognition Networks” [7] introduces a novel approach to emotion recognition called CAER-Net. CAER-Net incorporates context information in addition to facial expressions to provide a more comprehensive representation of emotional responses. The authors propose a two-stream encoding network that separately extracts features from both the face and context regions, and an adaptive fusion network to combine these features effectively.

POSTER++ for facial expression recognition: POSTER++, an enhancement over POSTER, achieves SOTA performance in FER by optimizing cross-fusion, two-stream architecture, and multi-scale feature extraction. This improvement significantly reduces computational costs while maintaining SOTA performance across standard datasets, marking a significant advancement in FER architectures.

The paper ”PAtt-Lite: Lightweight Patch and Attention Network for FER” [11] addresses the challenge of recognizing facial expressions in challenging conditions. The proposed method utilizes a truncated ImageNet-pre-trained MobileNetV1 as the backbone feature extractor and introduces a patch extraction block to capture local facial features. An attention classifier is introduced to enhance the learning of these features. Experimental results show that PAtt-Lite achieves SOTA performance on benchmark datasets, including CK+, RAF-DB, FER2013, and FERPlus, even in challenging subsets.

While a recent GitHub repository [5] tests RWKV on vision tasks, no published work specifically explores RWKV architectures for FER, highlighting a gap in research that our study aims to address.

This project aims to build upon the aforementioned works, particularly the successes of transformer-based architectures like POSTER++, and explore the potentials of RWKV in FER tasks. By conducting experiments with similar architectures as in the ViT paper and speculating on the interchangeability of transformers and RWKV, we seek to address the dependency on transformers of many current SOTA architectures and uncover the possibilities of leveraging RWKV in FER. In line with the approaches of POSTER++ and PAtt-Lite, our study explores the use of popular CNN architectures as feature extractors and incorporates attention-based mechanisms.

3. Data

The experimentation in POSTER++ involved the utilization of three distinct datasets, namely RAF-DB [8], AffectNet [10], and CAER-S [7], for training and benchmarking purposes. Despite our efforts to acquire access to both RAF-DB and AffectNet, regrettably, permission to use these datasets was not granted. Consequently, we solely relied on the publicly accessible CAER-S dataset for conducting our experiments. The CAER-S dataset is a subset of the larger CAER video benchmark. It consists of more than 70,000 frames and images that are sampled from over 13,000 annotated videos. As shown in Figure 1, each image in the dataset is annotated with one of the following seven emotion categories: happiness, sadness, anger, surprise, fear, disgust, and neutral. These classes allow for a comprehensive evaluation and analysis of context-aware emotion recognition techniques. Due to the absence of a dedicated validation dataset in the CAER-S dataset, we created our own validation set by extracting a subset consisting of 10% from the training set. To accommodate the varying sizes of images in the dataset, we resized each picture to a height of 400 and width of 600. These dimensions were chosen as they approximately represent the average size of images in our dataset, minimizing the need for significant reshaping of any individual picture. The dataset exhibits a uniform distribution of classes, obviating the need for ad-

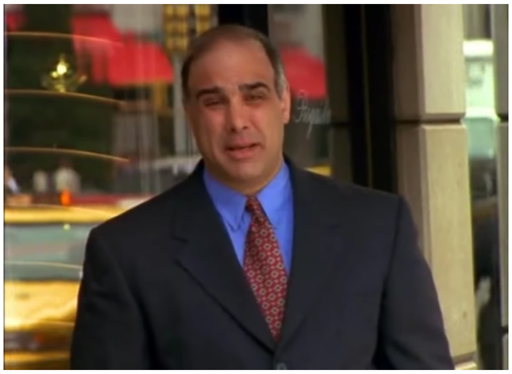

Figure 1. Disgust sample from the CAER-S dataset.

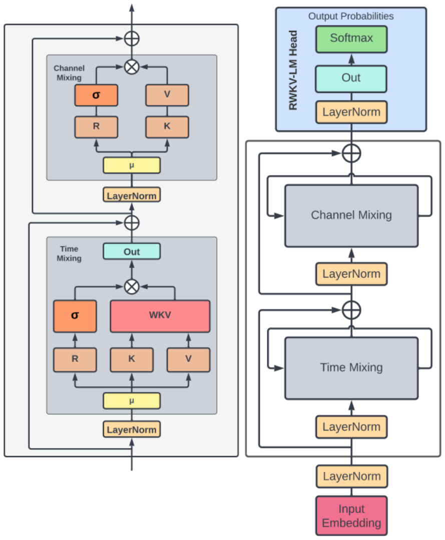

Figure 2. RWKV architecture [3].

ditional preprocessing steps to facilitate its suitability for our application. In our training process, we made a deliberate decision not to employ data augmentation techniques, such as horizontal flips, due to initial experiments indicating slower convergence when using the approaches employed by Poster++ and PAtt-Lite. Considering our limited computational resources, we opted not to incorporate data augmentation into our methodology. Despite the absence of data augmentation, our experiments did not exhibit any indications of overfitting.

4. Method

4.1. RWKV

The RWKV architecture in Figure 2, also known as Receptance Weighted Key Value, is composed of stacked residual blocks. Each block consists of a time-mixing sub-block and a channel-mixing sub-block, both of which have recurrent structures. The time-mixing sub-block incorporates the elements R (receptance), K (key), and V (value). These elements represent the acceptance of past information, the key for attention, and the value for attention, respectively. In the time-mixing sub-block, the current input is linearly interpolated with the input at the previous time step using the receptance vector R and the positional weight decay vector W. The WKV computation, which is analogous to the attention mechanism in Transformers, is performed in the time-mixing sub-block. It involves a weighted summation of the positional interval [1, t] multiplied by the receptance vector $\sigma(R)$. This allows for capturing long-term dependencies in the input sequence. The channel-mixing sub-block also utilizes the elements R, K, and V.

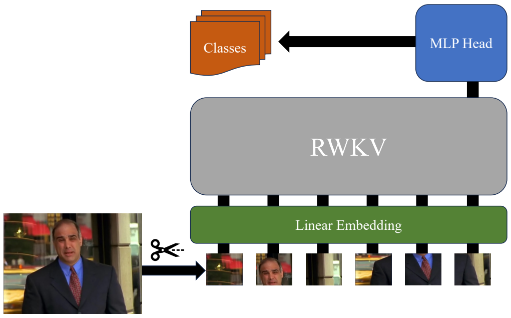

Figure 3. Our baseline architecture inspired by ViT [2].

Similar to the time-mixing sub-block, the current input is linearly interpolated with the input at the previous time step. The resulting interpolated R element is passed through a squared ReLU activation function and element-wise multiplied with the weighted K element to obtain the output. The RWKV architecture supports efficient parallelization using a time-parallel mode, similar to Transformers. Additionally, the RWKV architecture can also be used for RNN-like sequential decoding. In this mode, the output at time step t is used as the input at time step t+1. This leverages the recurrent structure of RWKV and enables efficient processing of longer sequences with a constant memory footprint and speed, regardless of the sequence length. In summary, the RWKV architecture combines time-mixing and channel-mixing sub-blocks with recurrent structures. It utilizes receptance, weight, key, and value elements for multiplicative interactions. It can be efficiently parallelized using a time-parallel mode and supports RNN-like sequential decoding for processing longer sequences.

4.2. Architecture

Our model's architecture shown in Figure 3 draws inspiration from the ViT paper while incorporating a baseline architecture and a hybrid approach.

Baseline Architecture: The baseline architecture begins by breaking down the input image into patches of a specific patch size. Each patch undergoes a linear projection, followed by the addition of optional positional encoding to preserve spatial information. These processed patches form a sequence fed into the RWKV architecture. The final entry's embedding in the sequence is projected through a Multilayer Perceptron (MLP) to derive predictions for the number of classes.

In our experimentation, we explored variations by comparing the usage of positional encoding against scenarios without it. Additionally, we considered alternate approaches such as pooling the mean over the output sequence and supplying this aggregated representation as input to the MLP head for classification. Early experiments, not reported in this article, showed inferior performance for the pooling operation. Therefore, we did not further investigate this approach.

Hybrid Architecture: In our hybrid architecture, we integrated a ResNet50 network up to its initial 4 blocks, utilizing these layers to capture low-level features effectively. Operating on the resulting feature map, we extract patches with a patch size of 5 which are then projected on the embedding dimension with a linear layer. During our experiments, we explored the use of pretrained ResNet50 weights in both frozen and unfrozen configurations.

Our approach is influenced by previous studies on FER and transformers. Specifically, we draw inspiration from the ViT paper and experiment with a setup that focuses on the RWKV mechanism to explore its capabilities. However, to enhance the RWKV with the advantages of CNNs, we incorporate common techniques such as CNN feature extractors, similar to the approaches used in POSTER++ and PAtt-Lite. This approach holds significant promise, as the ViT paper demonstrated that this set up can yield good performance for transformers. Given the comparable performance of RWKV models in NLP tasks, as showcased in the RWKV paper, it is reasonable to assume that adopting a similar approach to ViT will enable the RWKV to exhibit its potential in vision tasks. Moreover, CNNs have already demonstrated strong performance in SOTA FER models, making them a logical choice for our hybrid architecture.

5. Experiments

5.1. Scaling Up

Before comprehensive training on the entire dataset, a preliminary phase involved a critical evaluation, specifically a sanity check. This involved testing the model on a limited dataset of 14 images (two per facial expression class) to gauge if our training and architecture work properly. Results indicated proficiency, with the model achieving an accuracy rate of approximately 100% after a mere 8 epochs. This outcome underscores the model's capability to discern subtle nuances within the dataset, showcasing its efficacy in learning and classification. The success in this controlled environment establishes a promising foundation for subsequent training on the complete dataset, reinforcing confidence in the model's potential for our specific application.

In this section, we assessed the scalability of the RWKV architecture across three model versions. Each version investigates the impact of scaling up the architecture on the performance of the model on the test set. The details of the model versions are presented in Tab. 1. For all model versions, we adopt the baseline architecture that incorporates linear projection of the image patches. We decided to conduct the experiments without the utilization of the ResNet50 feature map, to fully rely on the RWKV and test its capabilities. The training process is performed for 25 epochs using the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and $lr = 10^{-3}$. We employ a warm-up strategy for the first epoch, setting the maximum learning rate to 0.1. To ensure stable training, we utilize cosine learning rate annealing and apply gradient clipping with a threshold of 1. The batch size is set to 64. We trained our models on an RTX 4070 Laptop GPU.

Version	n_{layers}	n_{heads}	d_{embed}	Parameters
Small	2	2	128	0.5M
Medium	4	4	256	2.8M
Large	6	6	512	14.8M

Table 1. Small, medium, and large model versions and their hyperparameters, as well as the total number of trainable parameters.

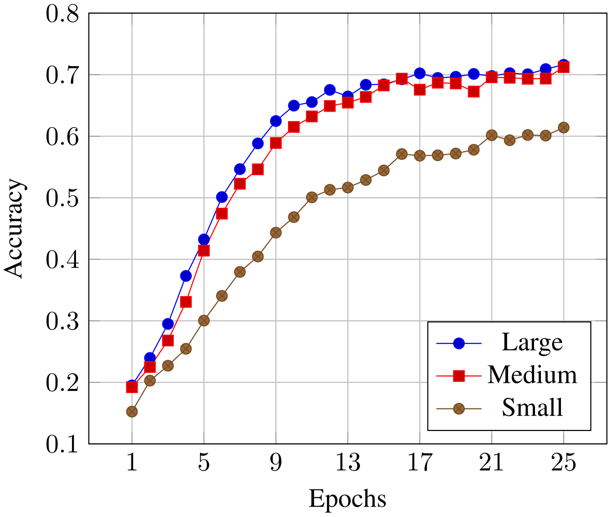

Figure 4. Plot of validation accuracy of the small, medium, and large model trained for 25 epochs.

By evaluating the performance of three versions of the RWKV architecture on the test set, we can gain valuable insights into the impact of model scaling. Fig. 4 illustrates the results of these experiments. Notably, the accuracy on the validation set indicates that increasing the model size has a positive effect on its performance. However, the graph also suggests that there may be a point of diminishing returns, where further increasing the model size does not lead to improved performance. Although this observation is intriguing, further investigation to confirm this hypothesis is limited by our computational resources. Nevertheless, we can still provide the final testing accuracy in Tab. 2 to provide a comprehensive overview of the performance of each model. In summary, our findings suggest that scaling up the RWKV model can enhance its performance up to a certain point, but there may be diminishing returns beyond that threshold. Further research with more extensive computa-

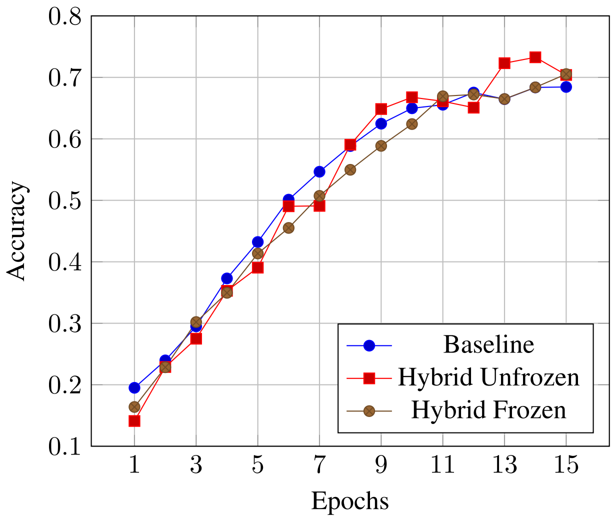

Figure 5. Plot of validation accuracy of the hybrid and baseline models trained for 15 epochs.

tional resources would be beneficial to explore this hypothesis in greater detail.

Version	Accuracy(%)
Small	61.41
Medium	70.41
Large	**70.82**

Table 2. Accuracy on the test set of the small, medium, and large model versions.

5.2. Hybrid Architecture

Version	Accuracy(%)
Baseline	70.82
Hybrid Frozen	70.30
Hybdrid Unfrozen	**73.02**

Table 3. Accuracy on the test set of the baseline and hybrid models.

In this section, we compare the performance of our large model with a hybrid architecture utilizing the ResNet50 feature extractor. We conducted experiments with both frozen and unfrozen layers, we refer to these versions as hybrid frozen and hybrid unfrozen respectively. To optimize computational resources, we observed that the models showed convergence at approximately 15 epochs. Hence, we trained both models for 15 epochs. The rest of the training set up is the same as in the first experiment. The results of the experiments revealed that utilizing the feature map of the frozen ResNet50 did not yield a significant improvement in model performance. However, when the weights were unfrozen, the model was able to fine-tune the initial layers and extract more useful features for FER. This led to an accuracy gain of approximately 2.2% at the best checkpoint. The validation accuracy over the 15 epochs is presented in Fig. 5, and the final test accuracy is reported in Tab. 3. These findings align with the observations made by Dosovitsky *et al.* in their work on ViT [2]. Their research suggests that working on a ResNet50 feature map can enhance model performance for smaller models. Nevertheless, our experiments demonstrate that utilizing the ResNet50 feature map cannot contribute to improving the accuracy of the model, when frozen. In summary, incorporating the ResNet50 feature extractor in our hybrid architecture can yield accuracy improvements, especially when the weights are unfrozen and fine-tuning is allowed.

5.3. Comparison

In this section, we evaluate the performance of our architectures in comparison to the SOTA model, Poster++. Additionally, we consider the performance of the architectures introduced in the original paper from Lee *et al.* [7] that introduced the CAER-S dataset. The comparison reveals that our models, except the small version, achieve comparable performance to the models presented by Lee *et al.*, while maintaining a smaller model size. However, it is important to note that we were unable to surpass the current SOTA model, the transformer based Poster++, which is nearly three times larger than our largest model.

Architecture	Accuracy(%)	Paramters
FaceRWKV Small	61.41	**0.5M**
FaceRWKV Medium	70.41	2.8M
FaceRWKV Large	70.82	14.8M
FaceRWKV Hybrid Unfrozen	73.02	15.0M
Poster++ [9]	**93.00**	43.7M
CAER-Net-S [7]	73.51	N/A
Fine-tuned ResNet [4]	68.46	58.2M
Fine-tuned VGGNet [12]	64.85	143.7M
Fine-tuned AlexNet [6]	61.73	61.1M

Table 4. Model test accuracy comparison on CAER-S Dataset.

5.4. Ablation study

In this section, we present an ablation study to delve into the intricacies of our architecture. To ensure computational efficiency and minimize our environmental impact, we adopted the small version and trained it for 15 epochs. The final testing accuracy results are summarized in Tab. 5.

Version	Accuracy(%)
Small	54.90
w/ Positional Encoding	54.52
w/o MLP head	54.80
w/o RWKV	14.29

Table 5. Accuracy on the test set of our small model with positional encoding, without the MLP head, and without rwkv after 15 epochs.

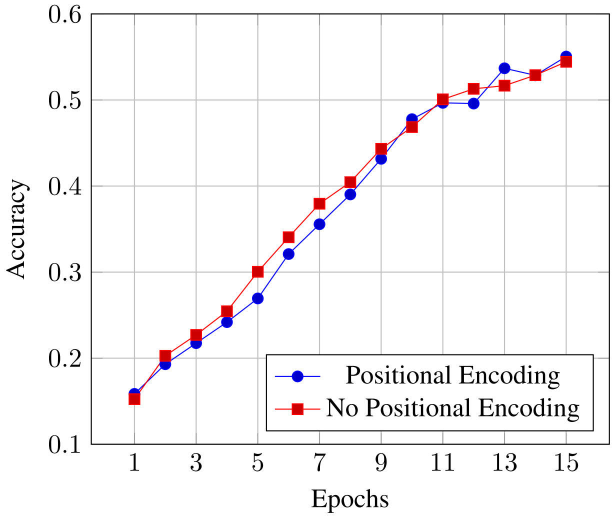

Figure 6. Plot of validation accuracy of the small model with and without positional encoding, trained for 15 epochs.

5.4.1 Positional Encoding

In this section, we explore the impact of incorporating positional encoding into our model. Following the approach described in [2], we introduced positional encoding to the patch embeddings. Notably, we employed a learnable positional encoding scheme. The validation accuracy in Fig. 6 demonstrates that the model with positional encoding exhibits similar performance compared to the model without it. We assume that the usage of positional encoding does not increase the models performance, because the model does not lose positional information due to its RNN characteristics. Yet, the official RWKV GitHub repository recommends the usage of positional encoding for vision tasks.

5.4.2 Linear Projection

To explore the influence of the MLP head in our architecture, we conducted experiments using an alternative design. In this setup, we trained a small model with a modified architecture that replaced the MLP with a linear layer as the classification head. This allowed us to evaluate the impact of the MLP on the model's performance. The results showed that both models performed fairly similarly. The model with the MLP head achieved higher accuracy on the validation set throughout most of the training process, except for the last epoch. However, the testing accuracy was very similar for both models (Tab. 5). It is worth noting that the MLP head potentially provides more expressiveness to the model, as indicated by its consistently good performance. Nevertheless, the results suggest that a linear head would be sufficient for this particular architecture. While the MLP head may offer some advantages in terms of expressiveness, the linear head demonstrated comparable performance, which can be seen in Fig. 7. This finding implies that the additional complexity introduced by the MLP may not be necessary for achieving satisfactory results in this specific FER task.

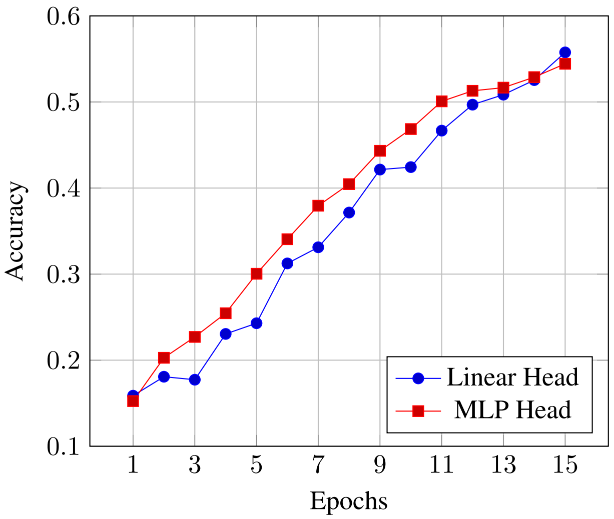

Figure 7. Plot of validation accuracy of the small model with MLP and linear head, trained for 15 epochs.

5.5. Hybrid architecture without RWKV

To thoroughly evaluate the influence of the RWKV mechanism in our architecture, we conducted an experiment in which we replaced the RWKV component with an identity mapping. In this configuration, we relied solely on the ResNet50 feature map and the MLP head for the FER task. This allowed us to isolate the impact of the RWKV mechanism and specifically analyze its contribution to the model's performance. Surprisingly, we observed that the model was unable to learn anything without the RWKV mechanism. The validation accuracy remained consistently low at 14.29%, and the loss did not decrease throughout the training process. This unexpected result suggests that the absence of the RWKV mechanism hindered the model's ability to acquire meaningful knowledge or make progress in learning. One possible explanation for this outcome is that the first linear layer in the MLP, responsible for mapping the high-dimensional feature map ($C \cdot H \cdot W = 76800$) to a lower-dimensional space (128), may have caused a significant loss of information. This potential loss of crucial knowledge could have impeded the model's capacity to effectively classify facial expressions. The findings from this experiment emphasize the importance of the RWKV mechanism in our hybrid architecture and suggest that it plays a critical role in facilitating the learning process and preserving essential information for accurate FER.

6. Conclusion

Throughout this project, we have successfully demonstrated the effective utilization of RWKV in vision tasks. We have observed that the performance of our RWKV model architecture scales with its size, up to a certain point. Additionally, by leveraging the first four blocks of a ResNet50, we have shown that our architecture can extract valuable features that enhance FER performance. Notably, we have also proved that RWKV can handle vision tasks without the need for positional encoding, contrary to the recommendations on the RWKV GitHub repository. Although our benchmark dataset did not allow us to achieve SOTA performance, we have achieved competitive results with fewer parameters, without incorporating any task-specific features like the Poster++ architecture. Throughout this report, we have gained valuable insights into the potential of RWKV for tasks beyond the NLP domain. To further expand on this research, we propose extending the investigation of RWKV in the vision domain based on the following key aspects. Firstly, we recommend exploring newer versions of RWKV that are implemented with increased efficiency, which can significantly reduce training time. Secondly, we suggest conducting experiments with larger models, comparable in size to ViT models. However, such experiments would require substantial computational resources that were not available to us during this project.

References

[1] PENG Bo. Rwkv-lm. https://github.com/BlinkDL/RWKV-LM, 2023. 2

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2021. 2, 3, 5, 6

[3] Bo Peng et al. Rwkv: Reinventing rnns for the transformer era. *Computation and Language*, 2023. 2, 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[5] howard hou. Visualrwkv. https://github.com/howard-hou/VisualRWKV/tree/main/VisualRWKV-v4, 2023. 2

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[7] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoonn Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. 2, 5

[8] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017. 2

[9] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, and Aibin Huang. Poster v2: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149*, 2023. 5

[10] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 2

[11] Jia Le Ngwe, Kian Ming Lim, Chin Poo Lee, and Thian Song Ong. Patt-lite: Lightweight patch and attention mobilenet for challenging facial expression recognition. *arXiv preprint arXiv:2306.09626*, 2023. 2

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5