Emotion Classification from Facial Images as a Meta-Sapiens Task.

Rahul Medicharla
Photogrammetric Computer Vision Lab
The Ohio State University
medicharla.2@buckeyemail.osu.edu

Alper Yilmaz
Photogrammetric Computer Vision Lab
The Ohio State University
yilmaz.15@osu.edu

Abstract

In this paper, we propose discrete emotion classification from facial images as a new downstream task for Sapiens, Meta's state-of-the-art human-vision foundational model. Currently, Sapiens only focuses on the physical aspects of human vision such as pose estimation and body part segmentation. Our model, MotivNet, extends Sapiens' human understanding and attempts to recognize underlying nuances conveyed by human physiognomy. We define three criteria to evaluate MotivNet's viability as a Sapiens task: benchmark performance, model similarity, and data similarity. Throughout this paper, we describe the components of MotivNet, our training approach, and our results compared to current benchmarks. We show that MotivNet achieved results comparable to existing benchmarks and meets the listed criteria, validating MotivNet as a downstream task and pushing forward Sapiens' capabilities as a human-centric model.

1. Introduction

The Vision Transformer [3], referred to as ViT, has significantly advanced computer vision in recent years. ViT's architecture integrates image data with Transformers [18], making it a desired approach to solve tasks such as image classification, anomaly detection, and image segmentation [7]. Built on the ViT architecture, the Masked Autoencoder [6] has become a common approach for generating rich feature spaces from images. The Masked Autoencoder, hereby referenced as MAE, learns to provide non-trivial relationships about an image within the latent dimension, enabling it to be used as a feature space [6].

Recently, Meta has published a new ViT and MAE-based human vision model, Sapiens (2024) [8]. Sapiens is a state-of-the-art (SOTA) foundational model with four primary human-centric downstream tasks: pose estimation, body part segmentation, depth estimation, and surface normal estimation. Built following the MAE pretraining-finetuning paradigm, each of the Sapiens downstream tasks outperformed its counterparts [8], incentivizing us to experiment with the model's capabilities.

In this paper, we propose discrete emotion classification from facial images as an additional Sapiens downstream task. This is necessary because, at present, Sapiens focuses solely on the physical aspects of human vision, without addressing the deeper aspects of humans. Emotion recognition extends Sapiens' human understanding, adding depth to its suite of tasks and enabling a broader range of potential human-centric applications that can be derived.

To evaluate our model's viability as a Sapiens task, we introduce three criteria.

1.The model architecture should not vary significantly from the Sapiens architecture.

2.The data used to fine-tune the model should be similar to the data Sapiens is trained on.

3.The model should achieve results comparable to current benchmarks.

Our model, MotivNet, is able to fulfill all three criteria. MotivNet employs an encoder-decoder architecture with Sapiens as the backbone and ML-Decoder [16] as the classification head. It is fine-tuned upon the AffectNet dataset [14]. We will discuss the components of MAE, Sapiens, ML-Decoder, and AffectNet in detail, demonstrate how MotivNet meets these criteria, and validate it as a viable Sapiens downstream task.

2. Related Information

2.1. MAE

MAE is a flavor of autoencoder [1] aimed at providing rich feature spaces for computer vision tasks. It is built on ViT's principle that an image can be represented as a sequence of non-overlapping patches for transformer-encoder use [3, 6]. It is an encoder-decoder model that projects a masked ViT representation of an image into latent space, and learns to reconstruct the input signal from it [6]. MAE is unique in its patch masking and self-supervision [4, 6], enabling deep feature learning without annotations.

MAE follows ViT in its encoder layer: linear projections

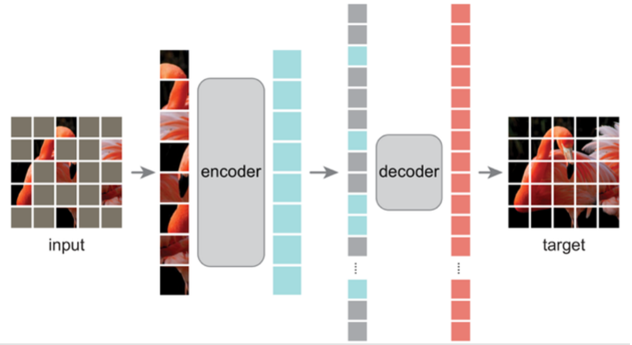

Figure 1. MAE architecture [6]

on patches with positional embeddings [6]. With sufficient pretraining, the encoder of an MAE learns to comprehensively describe an image in the latent dimension, enabling the tensor to function as a feature space. Once pretrained, the MAE can be finetuned on task-specific data for remarkable performance.

2.2. Sapiens

Sapiens is a human-vision foundational model built following the MAE pretraining-finetuning paradigm. This paradigm consists of two key stages: large-scale pretraining of a MAE and task-specific fine-tuning with relevant data. Meta first conducted a large pretraining on their MAE with Humans-300M, a proprietary dataset consisting of approximately one billion in-the-wild human images [8]. The dataset is preprocessed to only include unobstructed images and cropped to the human figure [8]. After pretraining, the MAE could reconstruct the original image from minimal input signals with excellent performance.

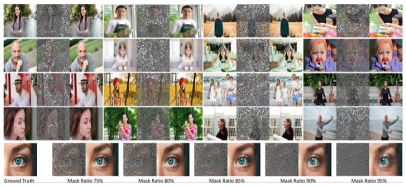

Figure 2. Sapiens MAE reconstruction [8]

For each downstream task, Sapiens followed the same approach: transfer the MAE into another encoder-decoder model and finetune it with a custom decoder and task-specific data [8].

2.3. ML-Decoder

ML-Decoder is a lightweight attention-based classification head with better spatial data utilization than global average pooling, a common accumulation algorithm in dimensionality reduction tasks [10, 16]. ML-Decoder is a variant of the transformer-decoder architecture with the key differences being the removal of self-attention, a group decoding scheme, and non-learnable queries [16, 18]. It provides a scalable architecture that grows linearly with the number of classes, making it suitable for multi-label classification tasks [16].

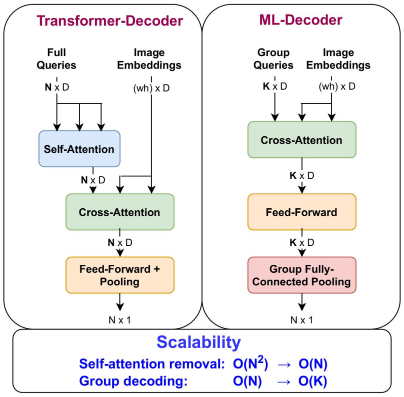

Figure 3. ML-Decoder architecture vs Transformer-decoder [16]

2.4. AffectNet

AffectNet is a facial emotion dataset consisting of more than one million facial images from-the-wild [14]. The images are manually annotated with one of seven recognizable emotions, neutral, happy, sad, surprise, fear, disgust, anger, and contempt [14]. Each facial image is cropped to be 224x224 pixel dimensions and contains only a singular human face [14]. Pulled from across three different search engines with 1250 emotion-related keywords, this dataset provides a comprehensive and diverse set of facial images with corresponding emotion, valence, and arousal values [14].

3. Methods

To build MotivNet we split our development into three stages described below: data selection, model architecture, and training implementation.

3.1. Data Selection

We chose to train MotivNet on AffectNet due to its in-the-wild image attributes [14]. Sapiens was pretrained on images taken from-the-wild [8], enabling strong generalization on such data. Since AffectNet shares this characteristic with Humans-300M [8], MotivNet is aligned with Sapiens' original direction, satisfying criteria number two.

Due to AffectNet's native class imbalance [14], we sampled n samples from each emotion class using SRS without replacement. C is the number of classes and AN is the dataset grouped by class.

$$n = \min_{c \in C} |AN_c| \tag{1}$$

$$n = 3750 \tag{2}$$

We sampled 3750 images per class for a total of 30,000 static facial images for the model to be finetuned upon.

AffectNet came with a presampled validation set of 500 images per class which we use as the test set for model evaluation. The sampled set defined above is randomly split into the training and validation set at runtime with a 9:10 and 1:10 ratio respectively.

3.2. Model

To satisfy criteria number one, we followed Sapiens' approach in defining model architectures for downstream tasks: transfer a pretrained MAE into another encoder-decoder model.

3.2.1. Encoder

We chose to use Sapiens' 308 Facial Keypoint Estimator [8] as the baseline MAE in our model. As this encoder was already finetuned to be proficient on facial images, we felt it was a sufficient starting point to provide rich feature spaces.

3.2.2. Decoder

After initial experimentation, we found that attention-based classification heads had superior performance than simple multi-layer perceptron classification heads. This was due to its dynamic feature selection on the feature space. In practice, we chose ML-Decoder as the classification head, allowing the model to assign variable weights on portions of the latent dimension, effectively filtering the feature space for relevant information based on perceived importance.

3.3. Training Implementation

MotivNet was trained on two NVIDIA RTX A6000 for 30 epochs. We used the AdamW optimizer and the Cosine Annealing with Warm Restarts scheduler [12, 13]. This caused stable weight updates without excessive regularization and randomly assigned learning rates, enhancing generalization and improving convergence on optimal solutions.

Since the encoder is being finetuned while the decoder is being trained from scratch, we set different learning rates for each. The encoder is being finetuned with an initial learning rate of 1E-8 while the decoder is being trained with an initial learning rate of 1E-5.

MotivNet is trained with a batch size of two and CrossEntropy as the loss function, the standard for multi-label classification tasks. We also used a gradient accumulation technique for every eight batches to effectively make the batch size 16 samples. Before training, each image is interpolated to (768,1024) pixel dimensions and normalized across all channels, as was done with Sapiens' 308 facial keypoint estimator [8].

4. Results

Once trained, we reviewed the loss and accuracy of MotivNet per epoch.

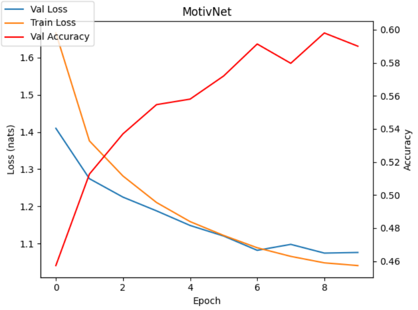

Figure 4. MotivNet Test Metrics

MotivNet stabilized within the first ten finetuning epochs largely due to Meta's pretraining effort. With limited data and finetuning, we achieved an accuracy score on the test set that places us within the top 40 models on the AffectNet FER task, fulfilling the third criteria. Results shown below.

Accuracy	Auroc	F1 score
57.49%	90.39%	57.41%

Table 1. Our results

Loss	Precision	Recall
1.148%	57.99%	57.49%

Table 2. Our results cont.

Model	Rank	Accuracy
Norface[11]	1	68.69%
LFNSB[2]	10	63.12%
Multi-task EfficientNet-B0[17]	18	61.32%
ViT-tiny[9]	30	58.28%
SL + 20% train[15]	36	52.46%
MotivNet(Ours)	?	57.47%

Table 3. Our results compared against AffectNet FER benchmarks

5. Conclusion

Through MotivNET, we show discrete emotion classification from static facial images is a viable Sapiens downstream task. All three of the listed criteria were met: the model architecture was similar to Sapiens, the model had results comparable to current benchmarks, and the dataset used was similar to the one Sapiens was pretrained on. Meta defined Sapiens' goal in their publication as "a unified framework and models to infer these assets in-the-wild to unlock a wide range of human-centric applications for everybody" [8]. We feel that with MotivNet, we have supported this goal and pushed forward Sapiens' ability for human understanding, unlocking an every further range of potential human-centric applications.

In future training, we plan to experiment with other datasets and hyper parameters that satisfy the given criteria to achieve a SOTA model. The creators of AffectNet recently published the AffectNet+ [5] dataset, an improvement from the initial AffectNet dataset. AffectNet+ provides multiple soft labels with different emotion confidence levels rather than a singular hard label [5]. This solves the original class imbalance problem and provides a more real-world application for emotion classification [5]. We plan on experimenting with this dataset.

References

[1] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021. 1

[2] Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*, page 1–15, 2024. 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 1

[4] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022. 1

[5] Ali Pourramezan Fard, Mohammad Mehdi Hosseini, Timothy D. Sweeny, and Mohammad H. Mahoor. Affectnet+: A database for enhancing facial expression recognition with soft-labels, 2024. 4

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 1, 2

[7] Sonain Jamil, Md. Jalil Piran, and Oh-Jin Kwon. A comprehensive survey of transformers for computer vision, 2022. 1

[8] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models, 2024. 1, 2, 3, 4

[9] Jia Li, Jiantao Nie, Dan Guo, Richang Hong, and Meng Wang. Emotion separation and recognition from a facial expression by generating the poker face with vision transformers. *IEEE Transactions on Computational Social Systems*, page 1–15, 2024. 3

[10] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network, 2014. 2

[11] Hanwei Liu, Rudong An, Zhimeng Zhang, Bowen Ma, Wei Zhang, Yan Song, Yujing Hu, Wei Chen, and Yu Ding. Norface: Improving facial expression analysis by identity normalization, 2024. 3

[12] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 3

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 3

[14] 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, 2019. 1, 2

[15] Mahdi Pourmirzaei, Gholam Ali Montazer, and Farzaneh Esmaili. Using self-supervised auxiliary tasks to improve fine-grained facial representation, 2022. 3

[16] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head, 2021. 1, 2

[17] Andrey V. Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, page 119–124. IEEE, 2021. 3

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 1, 2