Athens Journal of Sciences- Volume 12, Issue 2, June 2025 – Pages 129-150

Batch Transformer Architecture: Case of Synthetic Image Generation for Emotion Expression Facial Recognition

By Stanislav Selitskiy[*]

A novel Transformer variation architecture is proposed in the implicit sparse style. Unlike "traditional" Transformers, instead of attention to sequential or batch entities in their entirety of whole dimensionality, in the proposed Batch Transformers, attention to the "important" dimensions (primary components) is implemented. In such a way, the "important" dimensions or feature selection allows for a significant reduction of the bottleneck size in the encoder-decoder ANN architectures. The proposed architecture is tested on the synthetic image generation for the face recognition task in the case of the makeup and occlusion data set, allowing for increased variability of the limited original data set.

Keywords: *attention, transformer, synthetic image, data augmentation, emotion recognition*

Introduction

Face recognition (FR) is a topic of intensive research on the intersection of pattern recognition and computer vision. Because of its naturally innate nature for humans and noninvasive nature, FR earned its leading place in numerous practical applications in information security, surveillance systems, law enforcement, access control and smart cards. For a decade already, with the pioneering AlexNet architecture, Deep Learning (DL) models for FR have surpassed human accuracy in the controlled environment. However, in real-life settings, especially under conditions that DL models have not seen during training, such as the so-called Out-of-Distribution (OOD) condition, even advanced models suffer from uncertainty and failures.

However, in the case of Facial Expression Recognition (FER), advanced DL models of various architectures (from simple Engineering Features to Convolutional and Transformer ANNs) demonstrate much lower accuracy metrics than for FR and are prone to catastrophic failures, i.e. errors with high confidence.

Therefore, the idea that the whole spectrum of emotion expressions can be reduced to six basic facial feature complexes (Ekman and Friesen 1971) was challenged in the sense that human emotion recognition is context-based, fuzzy, and has many more components than the "basic six". The same facial feature components and their complexes may be interpreted differently depending on the situational context (Cacioppo et al. 2000, Gross and Canteras 2012).

Yet another concern is that the majority of the FER data sets are staged, either with the help of professional actors or plain amateurs, and their expressions are not

[*]PhD Student, University of Bedfordshire, UK.

https://doi.org/10.30958/ajs.12-2-4 doi=10.30958/ajs.12-2-4

natural, missing important micro-expressions generated by emotion states in the real, naturalistic situations.

The study aims to improve the accuracy of the DL models by expanding the training set with synthetically generated images. The scope is limited by the study of the auto-encoder ANN architecture based on a Transformer-type attention mechanism for synthetic image generation. The hypothesis we will investigate is: whether the scaling functionality of the Transformers' dot-product attention mechanism, applied to images of subjects expressing different varieties of emotion expressions and microexpressions, generate synthetic images that would cover missing variants in the training data set, improving the DL model performance on the test data set?

The intermediate objectives are to modify and improve the problematic aspects of the dot-product attention mechanism to make the encoder-decoder model viable on the limited hardware resources.

The contribution is organized in the following way: current Section 1 overviews the state and problems associated with Machine Learning algorithms employed in Face and Facial Expression Recognition and highlights aims, research questions, and the scope of the research. Section 2 reviews the current literature in the area of the proposed scope and research questions, Section 3 - describes the methodology of the proposed solution for the research questions; Section 4 - data sets which were used in computational experiments. Section 5 lists hardware parameters and model configurable parameters. Section 6 shows the results of the experiments in diagram and table form; Section 7 draws conclusions, briefly reiterates study limitations, and outlines future research directions.

Literature Review

Face and Facial Expression Recognition

The most obvious and non-invasive techniques for person recognition are based on biometric measurements of the physical characteristics of the human face or face parts. More subtle biometric-based approaches exploit other or additional physiological and behavioural characteristics, including electroencephalogram, cardiogram, or gesture signals (Mei and Weihong 2018, Schetinin et al. 2018).

Over three decades of research and development in face recognition allowed to achieve high performance in many applications; nevertheless, the recognition methods are still affected by factors such as illumination, facial expression, and poses as described in (Scherhag et al. 2017, Uglov et al. 2008, Xie et al. 2017). These are effects of the general ML problem of OOD, when training data do not fully represent test data.

Efficient attempts to find solutions to these problems have been made with ML and their DL branch methods and techniques, especially with those which extract hierarchical and semantic structures hidden in images (Frintrop 2010, Mei and Weihong 2018).

The early FR and FER ML models used the Bag of (Engineered) Features (BOF) methodology, which generates a compact representation of the large-scale

image features in which relations between the features are discarded. The representation may be seen either as a vector in the linear feature space or as a feature frequency histogram. The basis for the feature space, or the list of buckets for a histogram, also called terms vocabulary or visual word vocabulary, is constructed from the features extracted from the training set. Features extracted from the test set are associated with the nearest vocabulary terms, and a novel image is represented as a linear combination of the vocabulary terms and the frequency of their occurrence (O'Hara and Draper 2011).

BOF's popularity was built on its simplicity and compactness of the spatially order-less collections of the quantized local image descriptors, which makes BOF implementations algorithmically and computationally lightweight (O'Hara and Draper 2011). BOF has been adopted for computer vision and pattern recognition and, in particular, for face recognition studies (Chen et al. 2017, Passalis and Tefas 2017, Wu et al. 2011), and FER (Kalsum et al. 2018).

However, with the hardware (GPU) development, BOF models became specific niche ones, with DL models taking priority, especially Convolutional Neural Networks. The CNN architecture was introduced in the late 80's (LeCun et al. 1989), but became the "mainstream" architecture for image recognition in general and FR in particular when the AlexNet model (Krizhevsky et al. 2012) won the "ImageNet Large Scale Visual Recognition Challenge" (Alom et al. 2018, Deng et al. 2009) competition. A quickly growing family of CNN models suitable for FR, with deeper architectures, such as VGG (Simonyan and Zisserman 2014, Yu et al. 2016), and wider architectures, such as GoogLeNet and ResNet (Ren et al. 2016, Szegedy et al. 2015), followed the suite. The latter architectures used Directed Acyclic Graph (DAG) architecture to create cascade cells of parallel convolution layers of different parameters to allow their self-organisation (McNeely-White et al. 2020, Szegedy et al. 2016).

More recently, Transformer architecture, based on attention to local features (Luong et al. 2015, Vaswani et al. 2017), initially introduced into Natural Language Processing (NLP), was adopted for image recognition (Dosovitskiy et al. 2020). Also, hybrid CNN and Transformer architectures were proposed (Carreira et al. 2022, Han et al. 2021, Jaegle et al. 2021, Liu et al. 2021).

Despite their higher than CNN's computational demand, Transformers gained their popularity in FR partially to the hardware progress. The progress also brought fully-connected deep ANNs (or Multi-Layer Perceptrons (MLP)) back to FR attention also (Liu et al. 2021, Tolstikhin et al. 2021) Still, CNN is considered a viable architecture for the next decade (Liu et al. 2022).

The early works on FER, also used the engineered feature BOF methods (Berretti et al. 2010, Shan et al. 2009, Whitehill and Omlin 2006). CNNs were later employed to solve the facial expression recognition problem (Kim et al. 2016, Liu et al. 2015, Lopes et al. 2017, Mollahosseini et al. 2016, Ng et al. 2015).

Difficulties in the visual expression-only analysis of emotion expression led researchers to consider multi-modal biometric analysis using such information as pulse rate, ECG and EEG (Alarcao and Fonseca 2017, Raheel et al. 2020). Being a generic feature extraction ANN architecture, the CNN models were used for the classification of such hybrid as well (Hammad et al. 2021).

However, BOF algorithms and DL models are poorly generalised in the OOD conditions, especially for FER tasks. Generative AI, a significant part of which is represented by the Transformer-based models, on the one hand, is not directly related to the FR and FER tasks, but, on the other hand, can be used for the training data set augmentation by generating variety synthetic images, hypothetically filling the gaps of original data sets, and reducing the OOD problem.

Transformer Architecture

Transformer architecture for ANN, especially for the encoder-decoder (or just encoder) models (Bengesi et al. 2024), gained tremendous popularity with the publication (Vaswani et al. (2017) in which previous works on the attention mechanisms (Bahdanau et al. 2014, Gehring et al. 2016, Luong et al. 2015) were compiled and repackaged into a multi-head model which was dubbed as "Transformer", and applied to the Natural Language Processing (NLP), and in particular Machine Translation (MT) task.

The attention mechanisms, particularly the dot-product proximity mechanism of Vaswani's Transformer, help to solve the bottleneck's inadequate dimensionality problem of auto-encoders. The too-narrow bottleneck could lose important parameters, while too-wide would ineffectively use computational resources for processing low-useful parameters.

For NLP tasks, such parameters could be sequences of words or, rather, their embeddings into the contextual token (feature) space. In other application domains, such as image (Dosovitskiy et al. 2020), video (Liu et al. 2022), time series processing (Verma and Berger (2021), or even graphs (Min et al. 2022), Transformer architecture can also be beneficial, finding important for the task-related parameters, such as images, pixels, temporal signals, or relations.

Transformers' Problems

However, despite their popularity, Transformer architectures exhibit a number of problems of various kinds; some of them are effectively solved in a practical sense, and some are open discussion topics, such as their poor generalization under the Out-of-Distribution (OOD) conditions (Yadlowsky et al. 2023), catastrophic loss of dimensionality, i.e. degradation to a rank-1 matrix over multiple layers (Dong et al. 2021), loss of plasticity and forgetting (Pelosin et al. 2022, Shang et al. 2023), to be fair, that latter one is a problem in general for Deep Learning (DL) architectures.

On top of these high-level issues, more technical problems were dealt with in various variations of Transformer architectures. The so-called quadratic complexity of Transformers or dot-product attention comes from the transposed "key" and "query" matrices multiplication, which, given high-dimensionality input, occupies a large (quadratic input dimensionality) amount of memory (Phuong and Hutter 2022), which is the main problem and also strains computation resources on matrix multiplication. For NLP tasks, such large matrices are rather exotic for very large text attention windows of tens of thousand tokens (which became less exotic with the

introduction of the Large Language Models (LLM)), but for Computer Vision (CV) tasks, even modest image sizes in the $200 \times 200 - 300 \times 300$ range, cause a problem.

A patchy, so-called "visual words", architecture of the vision transformers, which offered global linear complexity over insignificant local quadratic complexity of the small patches in (Dosovitskiy et al. 2020), and hierarchical sequence of transformer layer, which would create "visual sentences" (Han et al. 2021). These architectures have been followed by a number of various modifications and borrowings back into NLP, primarily in the patch arrangement direction - sparse, overlapping, variable size, and so on (Bertsch et al. 2023, Yu et al. 2023, Zaheer et al. 2021).

Decoder parts, when reconstructing large outputs in CV tasks or LLM applications, may also face large matrix multiplication constraints. The NLP's response to such a challenge was a so-called Masked Language Model (MLM) (Besag 1975, Salazar et al. 2019) when only a portion in the range/target text is revealed for the model in training at a time. Similarly to the patchy "visual words" borrowing into vision Transformers, patchy output revealing of the range/target image was proposed to deal with the decoder large matrix problem (Carreira et al. 2022, Jaegle et al. 2021).

Another approach for dealing with resource constraints of the dot-product attention, is to use another, additive softmax attention mechanism, proposed already in (Bahdanau et al. 2014), but overshadowed by more popular Vaswani's type Transformers. Which, the dot-product-based Transformers, also suffer from instability during warn-up training, which is usually controlled by the following normalization layers. Such an instability was proposed to be dealt better, by including normalization inside Transformers (Xiong et al. 2020). Yet another improvement vector, targeting the persistent memory integration into Transformer architecture, proposes elements of Recurring Neural Networks (RNN) by introducing external memory that would hold the Transformer out data at one moment of time, and then feed it back into the Transformer's input at the following time step (Bulatov et al. 2022, Wu et al. 2022).

Dimensionality Inflation Problem of the Context Features in Auto-Encoder Models

The auto-encoder or encoder-decoder ANN architecture intends to extract important features of the input by mapping it into a limited dimensionality feature space of a bottleneck layer, by reconstructing the original input as a "label". Various mechanisms can be used to avoid this direct and inverse mapping to become identity mapping. For example, probabilistically mixing in a random noise, as in Variational Auto Encoder architecture, which development lead to the popular "diffusion" models.

The problem of such architectures is the dimensionality of the bottleneck, which, if overextended, leads to excessive ANN size on the decoder size. On the other hand, if the bottleneck is too squeezed, important features may be lost. Even for less resource-hungry tasks like image processing, Natural Language Processing (NLP), until recently, was facing that problem.

Traditional NLP tokenizing techniques included the preprocessing stage, on which "stop-words" are removed, remaining words are stemmed and lemmatized (converted to canonical dictionary form), and the Bag of Words (BoW), or Bag of

Features (BoF) in general setting, algorithm is used to map lemmatized words into a linear vector space, spanned on the most frequent and important words dictionary basis (O'Hara and Draper 2011, Struhl 2015). The whole sentence or a larger text is represented as a linear sum of all token vectors (or so-called "embeddings") (Zhang et al. 2010). Such an approach is resource usage effective but does not count in the sentence or larger text structure. For example, such sentences as: "A dog bites a man", "A man bites a dog", and "Dogs bite men" would be represented by the same embedding.

To introduce implicit elements of the linguistic structures, modern NLP models frequently use context tokenizers (Taylor 1953) of the BERT-like family (Devlin et al. 2018) that use the attention mechanism of the Transformer architecture to build such a context. A simple illustration of the BoW and BERT embedding differences would be the former creating "DOG", "BITE", "MAN", and the latter – "nullDOGbite", "dogBITEman", "biteMANnull", "nullMANbite", "manBITEdog", "biteDOGnull". That solves the BoW's structure blindness problem but greatly increases the dimensionality of the embedding space, which is the starting point of LLMs' high computational demands and size.

To keep with the human reader's attention span and produce a coherent flow of text, LLMs have to use long context windows for MLM training of thousands of words. The brute force use of the whole continuous windows is computationally problematic; therefore, another technique of extracting the most valuable and influential context words on the predicted word gave birth to computationally tractable but still huge LLMs - Attention mechanism (Bahdanau et al. 2014, Gehring et al. 2016, Liu et al. 2022) and its Transformer implementation (Vaswani et al. 2017). In such an approach of "attention", learnable matrices are used to compute cosine or Euclidean distances between the word relevance to the projected prediction over the context window sliding, and the most consistent contributor over time is kept and used, in such a way, reducing computational demand.

We experiment with a similar approach applied to the synthetic image generation via encoder-decoder ANN, radically reducing the bottleneck size to the number of the being controlled parameters. For example, in application to emotion generation - to 8 neurons in the final bottleneck layer, associated to 6 "basic" emotion expressions, neutral expression, and a closed-eye expression.

Proposed Solution

For the Generative AI model, we utilize the encoder-decoder architecture that uses a Cosine Distance Batch Primary Components (BPC) variation of the Transformers' dot-product attention mechanism to perform an implicit feature importance assignment on the encoder side.

An optional additional and more traditional variation of the Cosine Primary Element (CPE) Transformer attention mechanism, intended to select relevant data in the training and test batches, was preliminarily experimented with on a simple MNIST data set, producing promising results. However, on a more diverse and

complex data set of facial expressions BookClub, CPE performed worse, additionally consuming limited computational resources and was excluded from the final architecture.

A traditional visual transformer approach is used to address the quadratic complexity of the dot-product family of attention mechanisms. This approach divides the input image into patches, applies dot-product attention separately to each patch, and then combines the results. The following fully connected (FC) layer decreases the output dimensionality before passing information to the next BPC layer.

The meta parameter of the desired emotion expression is supplied together with input image information and then, with residual connections, is appended to each encoder layer. The first Visual Transformer BPC layer mixes meta-parameters into each attention patch, but in the following BPC-FC pair, only FC layer mixes the residual meta-parameters in.

A bottleneck selector of the important features, or rather a cascade of the gradual reduction of the feature dimensionality, was used to guide the decoder's image generation based on the limited number of features of interest.

The decoder side consists of the Learning ReLU (Rectified Linear Unit) activation functions specific to each neuron in the layer. The size of the layer is set according to ANNs composed on the Kolmogorov-Arnold Representation theorem.

We propose modifications of the canonical families of the Transformer attention heads (Vaswani et al. 2017) and Rectified Linear Unit (ReLU) activation functions. Improvements to the former family are aimed at implementing attention not between embeddings in the feature space, but rather attention between features (or dimensions of the features), effectively performing the feature selection in the Lipschitz sense, by projecting them into relevant subspaces.

Cosine and Batch Transformers

Vaswani-type Transformer attention head, or dot-product head, has an arbitrary or optional scaling factor common for all dimensions and embeddings or mini-batch vectors (for example, proportional to the W matrix dimension, etc.), Equation 1.

$$(1)\ k = W_k x + w_{k0},\ q = W_q x + w_{q0},\ v = W_v x + w_{v0}$$
$$B = q^T k/\sqrt{dk}$$
$$y = v\,softmax(B^T)$$

where x is a layer's input, $W_k x$, $W_q x$, $W_v x$ are linear transformations. In this notation, softmax is applied in a "vertical" direction. It should be noted that because vectors k and q, in case of the batch training, become matrices K and Q, hence B are matrices in the general case.

Such an arbitrary scaling requires using normalisation layers downstream. However, dot-product attention, naturally, as a part of the cosine distance function, can include implicit dimension-wise normalisation Equation 2. The V-transformation is optional in the attention head and, for simplification, is not shown - it could be done outside of the attention head or could be omitted entirely.

(2)

$$k = W_k x + w_{k0},\ q = W_q x + w_{q0}$$
$$q2 = \sum q \otimes q\ ,\ k2 = \sum k \otimes k$$
$$D = \sqrt{q2^T k2}$$
$$B = q^T k \otimes \frac{1}{D}$$
$$y = x\ softmax(B^T)$$

where $q2$ is a sum in the "vertical" vector dimension and is a horizontal vector (not scalar) for the batch training, as well as D is a matrix, also as B, which is calculated as a Hadamard division, and $\otimes$ is element-wise multiplication, or Hadamard product.

The Batch Primary Components (BPC) Transformer is similar to vision Transformers (Dosovitskiy et al. 2020) in terms that it attends not to the nearby embeddings but instead to features or embedding projections to individual dimensions Equation 3; therefore, B is a square matrix of the input vector dimensions for Batch Transformer heads, while for the Vaswani's-type Transformer heads - also square matrix, but mini-batch or sequence dimensions.

(3)

$$k = W_k x + w_{k0},\ q = W_q x + w_{q0}$$
$$q2 = \sum q^T \otimes q^T\ ,\ k2 = \sum k^T \otimes k^T$$
$$D = \sqrt{q2\ k2^T}$$
$$B = q\ k^T \otimes \frac{1}{D}$$
$$y = x^T\ softmax(B)$$

The term "Batch" is used here as an analogy to Batch Normalisation layers, which use another dimension for normalisation than Layer Normalisation. The "Primary Component" (PC) term is used because the Batch attention mechanism scales up or down individual dimensions of the input value, in contrast to the Vaswani-like "Primary Element" (PE) attention mechanism, scaling whole vectors in the input batch, Equation 2.

In the case of the vision transformers, not only the PC approach, Equation 4 is used (to find the most common or relevant pixels in the images, instead of the most relevant images in the batch), but also an input space $\mathcal{X}$ is mapped to a quotient space of patch subspaces $\mathcal{X}_1, \mathcal{X}_i, \mathcal{X}_n$, and then, each subspace is transformed using the Transformer's attention mapping, and range spaces joined into the output space $\mathcal{Y}$ by the direct sum $\bigoplus$:

(4)

$$\pi: X \subset R^m \ \mapsto \bigcup\{i \in n\}\ (X_i \subset R^{m/n})$$
$$y: X_i \subset R^{m/n} \mapsto Y_i \subset R^{m/n}$$

$$z = y_1 \bigoplus \ldots y_i \bigoplus \ldots y_n$$

where n is a number of attention patches.

Also, tensors Key k and Query q may not be the only results of the linear transformation. They can be made non-linear by adding activation functions (Cheng et al. 2023). In this research, both BPC and PE attention head variations are modified to use Hyperbolic Tangent non-linearity, Equation 5.

(5)

$$k = tanh(W_k x + w_{k0})$$
$$q = tanh(W_q x + w_{q0})$$

Plasticity-Learning Neuron-Wise Activation Units

Plasticity-learning neuron-wise activation units, for example, Learning Rectified Linear Unit (LrReLU), instead of using constant scalar slop multiplier a as if in Equation 6, use learnable vector A, Equation 7.

(6)

$$y = a \times x,\ z = relu(y): z_i = max(0, y_i)$$

(7)

$$y = A \otimes x,\ z = relu(y): z_i = max(0, y_i)$$

where $x \in \mathcal{X} \subset R^m$ is a layer's input of dimensionality m, $y \in \mathcal{Y} \subset R^m$ is the layer's output of the same dimensionality, y_i, z_i are parameters along the dimension $i \in m$.

The Equation 8 can be used to roughly emulate member Φ_q from the Equation 10 below, while for the member ϕ_{qp} emulation, neuron-specific hybrid of the Learning ReLU with fully connected layer may be used:

(8)

$$y = (A \otimes x),\ z = W\ relu(y): z_i = max(0, y_i)$$

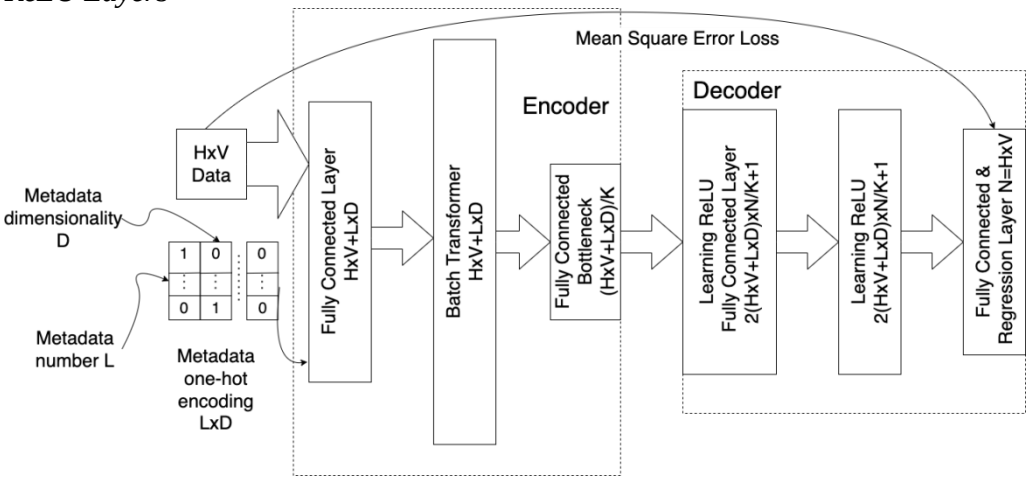
where W is another transformation matrix.

Proposed ANN Architectures

A general-purpose BPC Transformer-based encoder consists of a fully connected layer that mixes input information and metadata that describe it in a few laconic terms, either real numeric or categorical expressed numerically in some form. Then, a BPC Transformer follows, Equations 2-3, and a fully connected layer squeezes the Transformer's output into a narrow bottleneck. The following decoder, firstly, applies the Learning ReLU (or other Learning Unit) to the encoder's bottleneck input to further disable secondary dimensions; then, a fully connected dimensionality

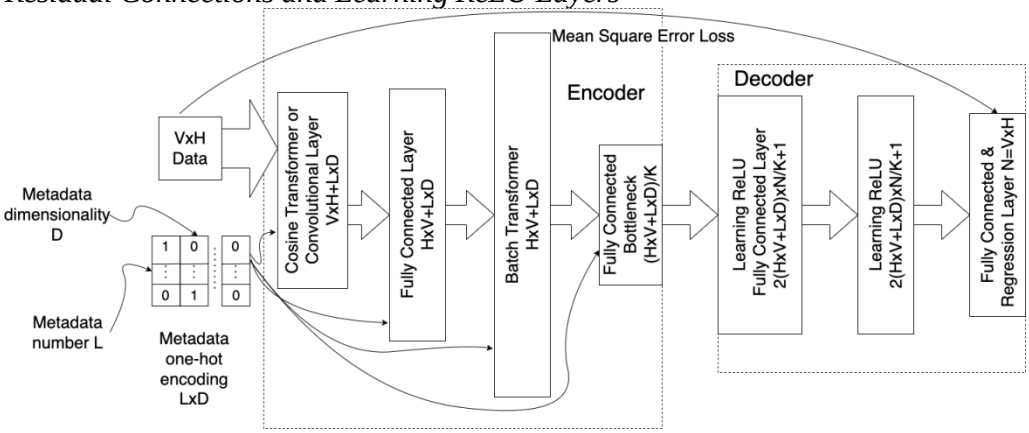
expansion layer follows with the attached second Learning ReLU, Equations 7-8. Finally, depending on the regression or classification task, a fully connected layer and Regression layer follow. This generic architecture is used for image and other parameters prediction tasks, with metadata being a number and dimensionality of predictors (such as emotion), Figure 1, where $H \times V$ is an input data dimensionality, $L \times D$ - number and dimensionality of meta-parameters, N - output dimensionality, and K is a bottleneck compression coefficient.

Figure 1. *Encoder-Decoder Architecture with Batch Transformer and Learning ReLU Layers*



A more specialized architecture that first, applies Cosine Primary Element (CPE) Transformer attention layer to input data, Equations 2, then the similar as described above architecture follows, but with residual or skip connections (previously known as a cascade) (AlFuhaid et al. 1997, He et al. 2015, He et al. 2020) of time metadata, applied to each consecutive layer (Figure 2).

Figure 2. *Encoder-Decoder Architecture with Cosine and Batch Transformers, Residual Connections and Learning ReLU Layers*



The two hidden non-linearity layers decoder was designed with a number of neurons m and $2m + 1$ on the first and second hidden layer, respectively, where m

is a bottleneck dimensionality, Equations 9-10. If we look at the layer transformation of an ANN, as a transformation from a higher dimensional space into the lower dimensional space $f,\ n < m$:

(9)

$$f: \mathcal{X} \subset R^m \mapsto \mathcal{Y} \subset R^n$$

The reason to limit the decoder to two layers was that if one looks at ANN as a Universal Approximation according to the Kolmogorov-Arnold superposition theorem (Kolmogorov 1961), for general emulation of the $f: \mathcal{X} \subset R^m \mapsto \mathcal{Y} \subset R$ process, the 2-layer is a minimal ANN configuration needed (given that activation functions are complex enough):

(10)

$$f(x) = f(x_1, \dots, x_m) = \sum_{q=0}^{2m} \Phi_q \left(\sum_{p=1}^{m} \phi_{qp}(x_p) \right)$$

where Φ_q and ϕ_{qp} are continuous $R \mapsto R$ functions.

The practicality of such ANN, as a Universal Approximator, was disputed in Girosi and Poggio (1989), particularly because of the non-smoothness, hence non-practicality, of the inner ϕ_{qp} functions. However, these objections were rebutted in Kurkova (1991). In Pinkus (1999), ϕ_{qp} activation functions are even called "pathological". Recently, interest to the Kolmogorov-Arnold architecture-based ANNs has returned (Liu et al. 2024, Vaca-Rubio et al. 2024, Genet and Inzirillo 2024).

Considering n as an ANN's output dimensionality, a n-product of Equations 10 was laid into the foundation of the Approximator ANN for Equations 9.

Data Sets

The proposed simple BPC Transformer-based encoder-decoder architecture, was initially tested as a vision transformer encoder head on the handwritten digits MNIST data set. A more convenient than the original Yann LeCun's repository (2013), a MATLAB format version of the data set was downloaded from (Lulu 2023). For non-continuous training, the original partitioning of the data set was used: 60000 training and 10000 testing images.

After the proof of concept verification on MNIST set, a more challenging BookClub artistic makeup data set was used. The BookClub data set, Figure 3, contains images of number $E = |C| = 21$ subjects. Each subject's data may contain a photo-session series of photos with no makeup, various makeup, and images with other obstacles for facial recognition, such as wigs, glasses, jewellery, face masks, or various headdresses. The data set features 37 photo sessions without makeup or occlusions, 40 makeup sessions, and 17 sessions with occlusions. Each photo

session contains circa 168 JPEG images of the 1072×712 resolution of six basic emotional expressions (sadness, happiness, surprise, fear, anger, disgust), a neutral expression, and the closed eyes photoshoots taken with seven head rotations at three exposure times on the off-white background. The subjects' age varies from their twenties to their sixties. The race of the subjects is predominately Caucasian and some Asian. Gender is approximately evenly split between sessions.

Figure 3. *BookClub Data Set Examples*

The photos were taken over two months, and several subjects were posed at multiple sessions over several weeks in various clothing with changed hairstyles, downloadable from https://data.mendeley.com/datasets/yfx9h649wz/3. All subjects gave consent to use their anonymous images in public scientific research.

The makeup design and application also varied in the artists' skills, style and heaviness of the pigments and face area covering percentage. Makeup artists of three levels volunteered their work for the project: mature, semi-professional, and professional. As for the pigment materials, professional artistic and theatrical pigments of Mehron, Inc. production were used in experiments. For occlusions, typical everyday items, likely to be found in households and on the streets, were used.

In addition to the practical usefulness in training and verifying ANN against the makeup and occlusions recognition avoidance, when non-makeup-only photo-sessions compound the training set and makeup and occlusion sessions are used for testing, such a data set is suited very well for benchmarking uncertainty estimation for the real-life conditions of the OOD conditions when test data not being well represented by the training data. The wide variety of lighting conditions, head orientations, emotion expressions, age, gender, and race makes this data set an excellent source of data for the aleatoric uncertainty training.

For synthetic image generation experiments, the data set was partitioned into the subject, makeup, and time-centred photo sessions. All images with makeup and

occlusions, and a few extra sessions without makeup, were selected for the training sub-set (13980 images), while for the test sub-set, only non-makeup and non-occluded sessions (1653 images) were used.

The number of controlled parameters, such as emotion expression, head rotation, subject, makeup or occlusion type, if used together, create bottleneck size that dictates decoder size putting too much computational requirements for the available GPU. Therefore, such controlled meta-parameters were divided into groups. This contribution describes results for the partition with 8 types of emotion expression, for the front head rotation, and no makeup used. That creates an original training set of 838 images and a test set of 227 images.

Experiments

The experiments were run on the Linux (Ubuntu 20.04.3 LTS) operating system with QuadroPro RTX 8000 (with 48GB GDDR5 memory), X299 chipset motherboard, 256 GB DDR4 RAM, and i9-10900X CPU. MATLAB 2023b with Deep Learning Toolbox was used as a programming platform. Models were run with 0.001 learning rate. For MNIST model, the minibatch size was 2048 and epoch number - 250, and For BookClub - 64 for minibatch, and 2000 epochs.

For the proof of concept of the proposed BPC Transformer and LrReLU layers, a simple straightforward encoder-decoder architecture was used, Figure 1. The architecture was applied to generating synthetic images using the MNIST data set to estimate the feasibility of the proposed solution visually. Thus, image size is $N \times M = 28 \times 28$, and metadata L was implemented as a one-hot encoding of digits, by setting one of its 10 variables to 1 and leaving others at 0. Randomly selected test images were assigned all 10 metadata values, and were run through the encoder-decoder, generating a mutation table, in which every digit was recreated in all 10 styles.

The color BookClub images were scaled to $100 \times 100 \times 3$, with 3 color channels. For such size, to fit in the GPU, a hierarchical Patchy Vision Transformer encoder organization had to be used, Figure 2. At the first layer of the hierarchical Vision Transformer, 5×5 patches were used for 3 channels, thus resulting in $m = 3000, n = 75\ 20 \times 20$ BPC Transformers, Equation 3. Then, the output was squeezed by a fully connected layer $K_1 = \frac{1}{3}$ times. The following BPC Transformer and fully connected layers squeezed the data into the final encoder bottleneck L_b. The decoder, similarly to the MNIST model, inflated, with coefficient K_e, bottleneck data back to the original image size.

The source code is publicly available at https://github.com/Selitskiy/BCGen and https://github.com/Selitskiy/ANNLib.

Results

The models' run-time was 8 hours for the head-on single Transformer for the MNIST data set, and 2 hours for the vision Patchy BPC Transformer model for BookClub data set.

The mutation table for the randomly selected test digits from MNIST data set is shown in Figure 4. The top row is original images, the following rows are muted by setting rotating metadata one-hot encoding flags for all 10 possible digit modifications.

Figure 4. *MNIST Digits Mutation Table. Top Row - Original Images, the Next 10 Rows - Mutated Images by Metadata Encoding Rotations*

The model for BookClub data set synthetic image generation is illustrated in Figure 5. The top left image is from the test data set, i.e. not seen by the model during training. Because the subject was not the control metadata, a particular subject expressing the generated emotion may be different consistently, or even different for each emotion (Figure 6).

Figure 5. *BookClub Test Image Emotion Expression Mutation. Top Left - Original Image*

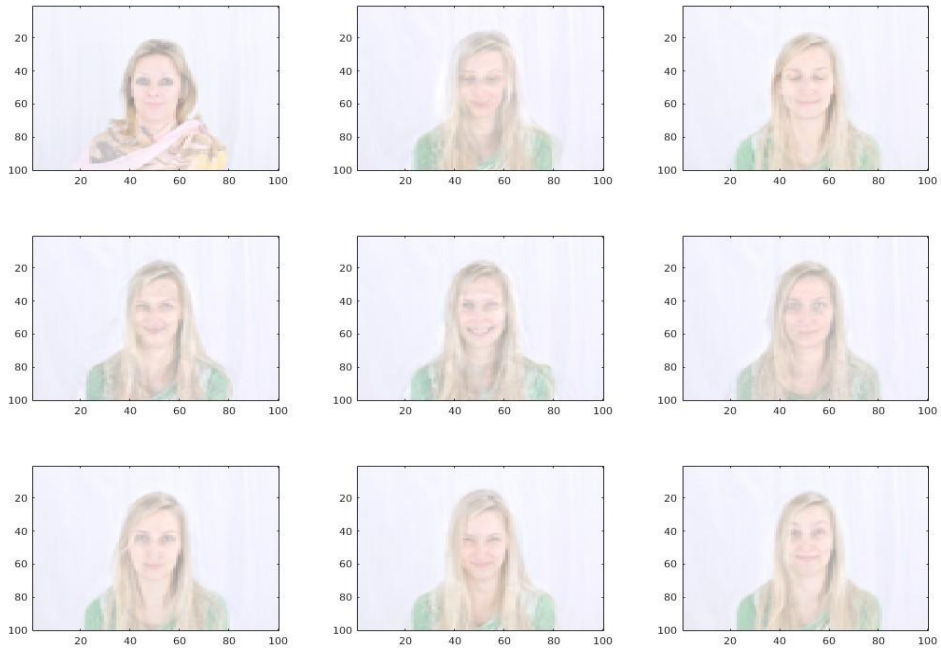
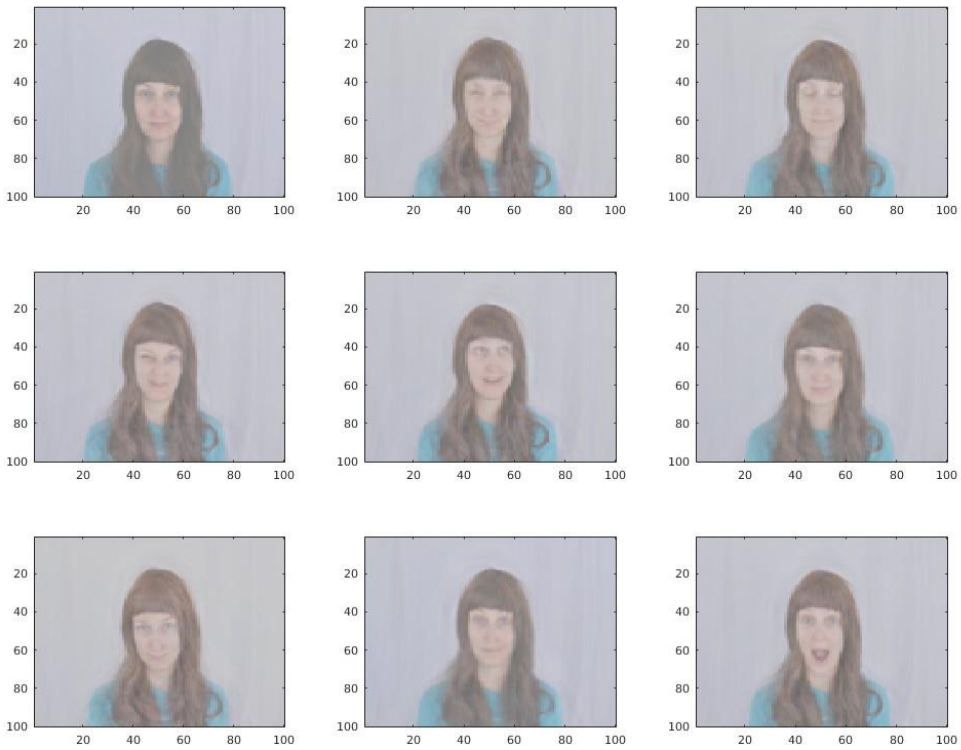


Figure 6. *BookClub Test Image Emotion Expression Mutation. Top Left - Original Image. Case of Multiple Subjects on the Generated Emotions*

For numerical verification, whether augmenting the original training set would improve FER performance of state-of-the-art (SOTA) CNN models, Inception v.2 model was selected, and pure synthetic training set and combine original + synthetic training sets were experimented with.

Of the 838 original images, synthetically generated emotion expression variants for each original image, made the synthetic training set size 6704, and combined original + synthetic training set - 7542. An example of such emotion mutation for one training set image is shown in Figure 7.

Figure 7. *BookClub Training Image Emotion Expression Mutation. Top Left - Original Image*



The resulting accuracy of the Inception v.2 model on FER task using original training data set was significantly improved (from 36% to 44% and to 43%) for completely synthetic, and combined original and synthetic data sets, Table.1. For range accuracy estimates, 5-fold variations of the train and test data sets were implemented. For the confidence level (99%) hypothesis testing, the Wilcoxon one-sided signed rank test was applied to the obtained accuracy distributions for the original, synthetic and hybrid trained results of Inception v.3 model.

The results show that at the 99% confidence level p-value is 0.0312, i.e. the probability of the null-hypothesis (that true distributions of synthetic or hybrid trained model results are not better than the originally trained model results) is slightly more than 3%. This means that the hull-hypothesis can be thrown away, and

the proposed generative encoder-decoder architecture indeed, statistically significantly improves emotion expression recognition of the Inception v.3 model on the experimented with data set.

Table 1. *Accuracy for Model Inception v.2 for Original, Synthetic, and Combined Training Sets*

Metric	Original	Synthetic	Combined
Accuracy	0.3602 ± 0.0208	0.4392 ± 0.0239	0.4329 ± 0.0316

Discussion, Conclusions and Future Work

An obvious limitation of the study is a freshly-backed ad-hock models, not yet optimized, and the results presented here are preliminary proof of concept. Another limitation is a visual estimate of the result without rigorous numeric accuracy or efficiency metrics.

Still, it can be seen that the proposed Batch Primary Component Transformer and Learning ReLU-based encoder-decoder architectures work successfully, performing the task of synthetic image generation by mutating the metadata input parameter. Augmenting organic training data sets with synthetically generated data, improves the accuracy of the SOTA CNN model, such as Inception v.2.

Experiments on other data sets are planned for future work, as well as, optimizing models, allowing them to work with larger image sizes, fully controllable by metadata parameters types of image mutation and generation. Also, experiments with other domains as time series and NLP are planned.

References

Alarcao SM, Fonseca MJ (2017) Emotions recognition using EEG signals: A survey. *IEEE Transactions on Affective Computing* 10(3): 374–393.

AlFuhaid AS, El-Sayed MA, Mahmoud MS (1997) Cascaded artificial neural networks for short-term load forecasting. *IEEE Transactions on Power Systems*, 12(4):1524–1529.

Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, Esesn BCV, Awwal AAS, Asari VK (2018) *The history began from Alexnet: A comprehensive survey on deep learning approaches*. arXiv preprint arXiv:1803.01164.

Bahdanau D, Cho K, Bengio Y (2014) *Neural machine translation by jointly learning to align and translate*. arXiv preprint arXiv:1409.0473.

Bengesi S, El-Sayed H, Sarker MK, Houkpati Y, Irungu J, Oladunni T (2024) *Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers*. IEEE Access.

Berretti S, Bimbo AD, Pala P, Amor BB, Daoudi M (2010) A set of selected sift features for 3d facial expression recognition. In *2010 20th International Conference on Pattern Recognition*, pages 4125–4128. IEEE.

Bertsch A, Alon U, Neubig G, Gormley MR. (2023) *Unlimiformer: Long-range transformers with unlimited length input*. arXiv preprint arXiv:2305.01625.

Besag J (1975) Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)* 24(3): 179–195.

Bulatov A, Kuratov Y, Burtsev M (2022) Recurrent memory transformer. *Advances in Neural Information Processing Systems* 35: 11079–11091.

Cacioppo JT, Berntson GG, Larsen JT, Poehlmann KM, Ito TA, et al. (2000) The psychophysiology of emotion. *Handbook of Emotions* 2(01): 2000.

Carreira J, Koppula S, Zoran D, Recasens A, Ionescu C, Henaff O, Shelhamer E, Arandjelovic R, Botvinick M, Vinyals O, et al. (2022) *Hierarchical perceiver*. arXiv preprint arXiv:2202.10890.

Chen BC, Chen YY, Kuo YH, Ngo TD, Le DD, Satoh S, Hsu WH (2017) Scalable face track retrieval in video archives using bag-of-faces sparse representation. *IEEE Transactions on Circuits and Systems for Video Technology* 27(7): 1595–1603.

Cheng X, Chen Y, Sra S (2023) *Transformers implement functional gradient descent to learn non-linear functions in context.* arXiv preprint arXiv:2312.06528.

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Devlin J, Chang MW, Lee K, Toutanova K (2018) *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Dong Y, Cordonnier JB, Loukas A (2021) Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, 2793–2803. PMLR.

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2020) *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv preprint arXiv:2010.11929.

Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17(2): 124.

Frintrop S, Rome E, Christensen HI (2010) Computational visual attention systems and their cognitive foundations: A survey. *ACM Trans. Appl. Percept.* 7(1): 1–39.

Gehring J, Auli M, Grangier D, Dauphin YN (2016) *A convolutional encoder model for neural machine translation*. arXiv preprint arXiv:1611.02344.

Genet R, Inzirillo H (2024) *Tkan: Temporal Kolmogorov-Arnold networks*. arXiv preprint arXiv:2405.07344.

Girosi F, Poggio T (1989) Representation properties of networks: Kolmogorov's theorem is irrelevant. *Neural Computation* 1(4): 465–469.

Gross CT, Canteras NS (2012) The many paths to fear. *Nature Reviews Neuroscience* 13(9): 651–658.

Hammad M, Pławiak P, Wang K, Acharya UR (2021) Resnet-attention model for human authentication using ecg signals. *Expert Systems* 38(6): e12547.

Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y (2021) Transformer in transformer. *Advances in Neural Information Processing Systems* 34: 15908–15919.

He K, Zhang X, Ren S, Sun J (2015) *Deep residual learning for image recognition*. arXiv preprint arXiv:1512.03385.

He R, Ravula A, Kanagal B, Ainslie J (2020) *Realformer: Transformer likes residual attention*. arXiv preprint arXiv:2012.11747.

Jaegle A, Gimeno F, Brock A, Vinyals O, Zisserman A, Carreira J (2021) Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, 4651–4664. PMLR.

Kalsum T, Anwar SM, Majid M, Khan B, Ali SM (2018) Emotion recognition from facial expressions using hybrid feature descriptors. *IET Image Processing* 12(6): 1004–1012.

Kim BK, Roh J, Dong SY, Lee SY (2016) Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces* 10(2): 173–189.

Kolmogorov AN (1961) *On the representation of continuous functions of several variables by superpositions of continuous functions of a smaller number of variables*. American Mathematical Society.

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25.

Kurkova V (1991) Kolmogorov's Theorem Is Relevant. *Neural Computation* 3(4): 617–622.

LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD (1989) Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1(4): 541–551.

LeCun Y, Cortes C, Burges C (2013) *MNIST handwritten digit database* [Online; accessed 12. Nov. 2023].

Liu M, Li S, Shan S, Chen X (2015) Au-inspired deep networks for facial expression feature learning. *Neurocomputing* 159: 126–136.

Liu R, Li Y, Liang D, Tao L, Hu S, Zheng HT (2021) *Are we ready for a new paradigm shift? a survey on visual deep mlp*. arXiv preprint arXiv:2111.04060.

Liu Y, Sun G, Qiu Y, Zhang L, Chhatkuli A, Gool LV (2021) *Transformer in convolutional neural networks*. arXiv preprint arXiv:2106.03180.

Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S (2022) *A convnet for the 2020s*. arXiv preprint arXiv:2201.03545.

Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. (2022) Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Liu Z, Wang Y, Vaidya S, Ruehle F, Halverson J, Soljačić M, et al. (2024) *Kan: Kolmogorov-Arnold networks*. arXiv preprint arXiv:2404.19756.

Lopes AT, Aguiar ED, Souza AFD, Oliveira-Santos T (2017) Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. *Pattern Recognition* 61: 610–628.

Lulu (2023) *Load MNIST database in Matlab, lulu's blog* [Online; accessed 12. Nov. 2023].

Luong MT, Pham H, Manning CD (2015) *Effective approaches to attention-based neural machine translation.* arXiv preprint arXiv:1508.04025.

McNeely-White D, Beveridge JR, Draper BA (2020) Inception and Resnet features are (almost) equivalent. *Cognitive Systems Research* 59: 312–318.

Mei W, Weihong D (2018) *Deep face recognition: A survey*. CoRR, abs/1804.06655.

Min E, Chen R, Bian Y, Xu T, Zhao K, Huang W, et al. (2022) *Transformer for graphs: An overview from architecture perspectiv*e. arXiv preprint arXiv:2202.08455.

Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, 1–10. IEEE.

Ng HW, Nguyen VD, Vonikakis V, Winkler S (2015) Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 443–449.

O'Hara S and Draper BA (2011) *Introduction to the bag of features paradigm for image classification and retrieval*. arXiv, abs/1101.3354.

Passalis N and Tefas A (2017) Learning neural bag-of-features for large-scale image retrieval. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47(10): 2641–2652.

Pelosin F, Jha S, Torsello A, Raducanu B, Weijer J (2022) Towards exemplar-free continual learning in vision transformers: an account of attention, functional and weight regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3820–3829.

Phuong M and Hutter M (2022) *Formal algorithms for transformers*. arXiv preprint arXiv: 2207.09238.

Pinkus A (1999) Approximation theory of the mlp model in neural networks. *Acta Numerica* 8: 143–195.

Raheel A, Majid M, Alnowami M, Anwar SM (2020) Physiological sensors based emotion recognition while experiencing tactile enhanced multi-media. *Sensors* 20(14): 4037.

Ren S, Sun J, He K, Zhang X (2016) Deep residual learning for image recognition. In *CVPR*, 4.

Salazar J, Liang D, Nguyen TQ, Kirchhoff K (2019) *Masked language model scoring*. arXiv preprint arXiv:1910.14659.

Scherhag U, Raghavendra R, Raja KB, Gomez-Barrero M, Rathgeb C, Busch C (2017) On the vulnerability of face recognition systems towards morphed face attacks. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, 1–6.

Schetinin V, Jakaite L, Nyah N, Novakovic D, Krzanowski W (2018) Feature extraction with GMDH-type neural networks for EEG-based person identification. *International Journal of Neural Systems* 28(6): 1750064.

Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing* 27(6): 803–816.

Shang C, Li H, Meng F, Wu O, Qiu H, Wang L (2023) Incrementer: Transformer for class-incremental semantic segmentation with knowledge distillation focusing on old class. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7214–7224.

Simonyan K, Zisserman A (2014) *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556.

Struhl S (2015) *Practical text analytics: Interpreting text and unstructured data for business intelligence*. Kogan Page Publishers.

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. (2015) Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Taylor WL (1953) "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly* 30(4): 415–433.

Tolstikhin IO, Houlsby N, Kolesnikov A, Beyer L, Zhai X, Unterthiner T, et al. (2021) Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems* 34.

Uglov J, Jakaite L, Schetinin V, Maple C (2008) Comparing robustness of pairwise and multiclass neural-network systems for face recognition. *EURASIP Journal on Advances in Signal Processing* (Dec): 468693.

Vaca-Rubio CJ, Blanco L, Pereira R, Caus M (2024) *Kolmogorov-Arnold networks (kans) for time series analysis*. arXiv preprint arXiv:2405.08790.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. (2017) *Attention is all you need*. CoRR, abs/1706.03762.

Verma P and Berger J (2021) *Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions*. arXiv preprint arXiv:2105.00335.

Whitehill J and Omlin CW (2006) Haar features for facs au recognition. In *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE.

Wu Y, Rabe MN, Hutchins DL, Szegedy C (2022) *Memorizing transformers*. arXiv preprint arXiv:2203.08913.

Wu Z, Sun QKJ, Shum HY (2011) Scalable face image retrieval with identity-based quantization and multireference reranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(10).

Xie Z, Jiang P, Zhang S (2017) Fusion of lbp and hog using multiple kernel learning for infrared face recognition. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, 81–84.

Xiong R, Yang Y, He D, Zheng K, Zheng S, Xing C, et al. (2020) On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 10524–10533. PMLR.

Yadlowsky S, Doshi L, Tripuraneni N (2023) *Pretraining data mixtures enable narrow model selection capabilities in transformer models*. arXiv preprint arXiv:2311.008 71v1

Yu L, Simig D, Flaherty C, Aghajanyan A, Zettlemoyer L, Lewis M (2023) *Megabyte: Predicting million-byte sequences with multiscale transformers*. arXiv preprint arXiv: 2305.07185.

Yu W, Yang K, Bai Y, Xiao T, Yao H, Rui Y (2016) Visualizing and comparing Alexnet and VGG using deconvolutional layers. In *Proceedings of the 33rd International Conference on Machine Learning*.

Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, et al. (2021) *Big bird: Transformers for longer sequences*. arXiv preprint arXiv:2007.14062.

Zhang Y, Jin R, Zhou ZH (2010) Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* 1: 43–52.