Deep Learning for Video Action Recognition: A Comparative Study of ConvLSTM, PredRNN, and Transformer Models

# Abdul Rahman

*DS-D*

21I-1731

# Muhammad Wissam

*DS-D*

21i-0709

Adnan

DS-D

19I-1731

***Abstract*—This project aims to explore video action recognition using deep learning techniques, implemented through a Flask application that leverages TensorFlow/Keras models. The system is designed to upload video files, preprocess them, and classify actions such as PushUps, JumpingJack, PullUps, VolleyballSpiking, and ApplyLipstick. The system uses various models, including ConvLSTM, PredRNN, and Transformer-based models, with performance evaluated using metrics such as Mean Squared Error (MSE) and Structural Similarity Index (SSIM).**

1. INTRODUCTION

Video action recognition is a critical task in computer vision, with applications in surveillance, human-computer interaction, and healthcare. This project focuses on using deep learning models for action recognition from videos. By integrating a Flask application, the system allows users to upload videos for action classification. The models were trained and tested on the UCF101 action recognition dataset, with preprocessing steps including frame extraction and resizing. Three models were evaluated: ConvLSTM, PredRNN, and a Transformer-based architecture.

1. EASE OF USE

*A. Maintaining the Integrity of the Specifications*

The paper was formatted following the IEEE conference guidelines using the IEEEtran LaTeX class file. All formatting requirements, such as margins, column widths, line spacing, and font styles, were adhered to strictly. This ensures that the paper conforms to the standard layout used across all papers in the proceedings. Any design peculiarities, such as the proportionally larger head margin, are intentional and have been specified to maintain consistency in presentation. These specifications ensure that the paper is part of the uniform design scheme that the conference expects, without disrupting the overall structure or flow of the proceedings.

1. PREPARE YOUR PAPER BEFORE STYLING Before beginning the formatting process, the content of the paper was written and saved as a separate text file. All sections of the paper were completed and organized prior to formatting, ensuring that the content and structure were finalized. This approach allowed for the refinement of the ideas, the flow of the paper, and the accuracy of the information, particularly in terms of the methodology, results, and conclusions.
2. *Abbreviations and Acronyms*

Abbreviations and acronyms must be defined the first time they are used in the text, even if they have already been defined in the abstract. For example, **Structural Similarity Index (SSIM)** is defined upon first usage and subsequently referred to as **SSIM**. The same rule applies to other abbreviations used throughout the paper, such as **Mean Squared Error (MSE)** and **Flask Application (FA)**.

Some common abbreviations do not need to be defined, such as IEEE, ac (alternating current), dc (direct current), rms (root mean square), and others that are universally recognized in technical writing.

Abbreviations should not be used in the title or section headings unless absolutely necessary, to maintain clarity and readability.

1. *Units*

In this paper, the **International System of Units (SI)** is used as the primary system for all measurements. Where applicable, **SI units** are employed for consistency and clarity. For example, dimensions such as image resolution are expressed in **pixels** (dimensionless), and loss metrics like **Mean Squared Error (MSE)** are reported as **dimensionless values**.

For clarity, **English units** may be included as secondary units in parentheses when relevant. For example, a video frame size of **64 x 64 pixels** could also be noted as **64 x 64 pixels (inches)** if necessary, although SI units are generally preferred for computational tasks.

Mixed units, such as combining **SI and CGS** (Centimeter-Gram-Second) units, are avoided to prevent confusion. When multiple units are used in an equation, each unit is clearly stated, ensuring dimensional consistency and preventing errors in calculations. For example, if an equation involves **time** (seconds) and **length** (meters), these are consistently used throughout.

To maintain consistency, we adhere to a uniform format for **units**. For instance, when referring to **Webers per square meter (Wb/m²)**, we avoid abbreviating units incorrectly. For clarity, the full spelling is used when units are described in the text (e.g., "a few **Webers per square meter**").

In the case of numerical expressions, **a zero is placed before decimal points** (e.g., "0.25" instead of ".25") to ensure readability and accuracy. Similarly, volume is expressed as **cm³**, not "cc," to adhere to proper unit notation

1. *Equations*

*Equations are numbered consecutively throughout the paper for clarity. Here, I will list the primary equations related to your video action recognition models, particularly those involving* ***Mean Squared Error (MSE)****,* ***Gradient Loss****, and the architecture-specific components for your* ***Transformer-based model****.*

#### *1. Mean Squared Error (MSE)*

*The* ***MSE*** *is commonly used to evaluate the difference between predicted and true frames. It is defined as the average of the squared differences between the actual values yiy\_iyi​ and the predicted values y y^​i​:*

*MSE=1/N ∑(yi−yi)^2 \**

*Where:*

* *yi is the true value,*
* *y^​i​ is the predicted value,*
* *NNN is the total number of samples.*

#### *2. Custom Loss Function*

*In addition to* ***MSE****, the model incorporates* ***gradient loss****, which compares the gradients of the true and predicted frames. This encourages the model to preserve spatial information over time:*

*Losscustom=MSE(y,y^)+λ⋅Gradient Loss*

*Where:*

* *MSE(y,y^) is the Mean Squared Error loss,*
* *λ is a hyperparameter that controls the importance of the gradient loss term.*

#### *3. Gradient Loss*

*The* ***gradient loss*** *ensures that the spatial gradients between true and predicted frames are similar. It is computed by applying image gradients to both the true and predicted frames:*

*Gradient Loss=1/B ∑ (∥∇yb−∇y^b∥22) Where:*

* *B is the batch size,*
* *∇yb ​ and ∇y^b are the spatial gradients of the true and predicted frames, respectively.*

#### *4. Transformer Encoder*

*The* ***Transformer Encoder*** *uses multi-head attention and a feed-forward neural network (FFN) to process input frames. The encoder updates its representation by adding the residual connection after applying attention and the FFN:*

*Transformer Encoder Output=LayerNorm(x+MultiHeadAttention(x,x))*

*FinalOutput=LayerNorm(x+FFN(LayerNorm(x)))*

*Where:*

* *x is the input sequence of frames,*
* *MultiHeadAttention(x,x) is the attention mechanism applied to the input,*
* *FFN(x) is the feed-forward network applied after attention.*

#### *5. Residual Block*

*A* ***Residual Block*** *in your model allows the input to bypass the convolutional layers, which helps in training deeper networks. The output yyy of a residual block is the sum of the input xxx and the processed output h(x)h(x)h(x):*

*y=ReLU(x+h(x))*

*Where:*

* *x is the input to the block,*
* *h(x)) is the output of the convolutional layers in the block.*

#### *6. Transformer Decoder*

*The* ***Transformer Decoder*** *uses the encoder's output to predict the next frames in the video sequence. The decoder takes the last output from the encoder and applies multiple dense layers to generate the predicted frames:*

*y^t=Dense(xt−1)Where:*

* *y^tis the predicted frame at time step ttt,*
* *xt−1 is the input from the previous time step (or the last decoder output).*

1. *Some Common Mistakes*

*In academic writing, especially for IEEE papers, there are a few common mistakes that should be avoided:*

1. ***Data****: The word* ***"data"*** *is plural, not singular. For example, it should be written as "The data are processed" rather than "The data is processed."*
2. ***Scientific Constants****: The subscript for the permeability of vacuum μ0and other common scientific constants must use the number* ***zero*** *with subscript formatting, not a lowercase letter “o”.*
3. ***Punctuation with Quotation Marks****: In American English, commas, semicolons, periods, question marks, and exclamation marks are placed inside quotation marks only when a complete thought or name is cited, such as a title or a full quotation. For example:*
   * *Correct: “The results are consistent,” said the researcher.*
   * *Incorrect: “The results are consistent”.  
     When quotation marks are used to highlight a word or phrase, punctuation should appear* ***outside*** *of the quotation marks. For instance:*
   * *Correct: The term "efficiency" is used here in the context of performance.*
4. ***Parenthetical Phrases****: A parenthetical phrase or statement at the end of a sentence is punctuated* ***outside*** *the closing parenthesis. For example:*
   * *Correct: The algorithm performs well (as shown in Figure 2).*
   * *Incorrect: The algorithm performs well (as shown in Figure 2.)*
5. ***Graph Terminology****: A graph within a graph is referred to as an* ***“inset”****, not an* ***“insert”****.*
6. ***Alternately vs. Alternatively****: The word* ***alternatively*** *is preferred to* ***alternately****, unless you specifically mean something that alternates. For example:*
   * *Correct: You can alternatively choose a different method.*
   * *Incorrect: You can alternately choose a different method.*
7. ***Avoid Using “Essentially”****: Do not use the word* ***"essentially"*** *to mean "approximately" or "effectively". For example, instead of saying "This is essentially the same as", it is better to say "This is approximately the same as" or "This is effectively the same as."*
8. ***Title Capitalization****: In your paper title, if the words* ***“that uses”*** *can accurately replace the word* ***“using”****, then capitalize the "u" in* ***“Using”****; otherwise, keep "using" in lowercase. For example:*
   * *Correct: "A Model That Uses Convolutional Layers"*
   * *Incorrect: "A Model using Convolutional Layers"*
9. ***Confusing Homophones****: Be aware of the different meanings of commonly confused homophones:*
   * ***Affect*** *(verb): To influence something.*
   * ***Effect*** *(noun): A result or outcome.*
   * ***Complement*** *(noun): Something that completes or enhances.*
   * ***Compliment*** *(noun): A polite expression of praise.*
   * ***Discreet*** *(adjective): Careful and prudent in speech or action.*
   * ***Discrete*** *(adjective): Separate or distinct.*
   * ***Principal*** *(noun): The head of a school or the main part.*
   * ***Principle*** *(noun): A fundamental truth or law.*
10. ***Imply vs. Infer****: Do not confuse* ***"imply"*** *and* ***"infer"****.*

* ***Imply*** *means to suggest something indirectly.*
* ***Infer*** *means to deduce something based on evidence.*

1. ***Prefix “Non”****: The prefix* ***“non”*** *should not have a hyphen when attached to a word it modifies. For example, it should be* ***nonlinear****, not* ***non-linear****.*
2. ***Abbreviation of “et al.”****: There is* ***no period*** *after the* ***“et”*** *in the Latin abbreviation* ***“et al.”****, which stands for "and others".*
3. ***i.e. and e.g.****:*
   * ***"i.e."*** *stands for* ***“that is”*** *or* ***“in other words”****.*
   * ***"e.g."*** *stands for* ***“for example”****.*
4. ***Citation Style****: An excellent style manual for science writers is* ***[7]****.*
5. *Authors and Affiliations*

**The authors of this paper are listed as follows:**

**Muhammad Wissam and Abdul Rahman  
Affiliations: Fast University Islamabad**

1. *Identify the Headings*

Headings serve as essential organizational devices, guiding the reader through the structure of the paper. There are two types of headings used in IEEE papers: **component heads** and **text heads**.

#### 1. Component Heads

Component heads are used to identify the different components of the paper, such as acknowledgments, references, and any other major sections that are not topically subordinate to each other. These should be formatted using **Heading 5** style. Examples of component heads include:

* **Acknowledgments**
* **References**

In addition, for figure captions and table titles, the following formatting should be used:

* **Figure caption** for all figure captions.
* **Table head** for the titles of tables.

#### 2. Text Heads

Text heads organize the paper's topics on a relational, hierarchical basis. The **paper title** is the primary text head because all subsequent material relates to and elaborates on this single topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used. If there are fewer than two sub-topics, then no subheads should be introduced.

* **Primary Text Head**: The title of the paper.
* **Sub-Headings**: If applicable, use uppercase Roman numerals (e.g., **I. Introduction**, **II. Methodology**).
* **Run-In Heads**: Sections like **Abstract** require additional styling. You should apply an italic style to differentiate the heading from the text.

For example:

1. **I. Introduction**
2. **II. Methodology**
3. **III. Results**
4. **IV. Discussion**
5. *Figures and Tables*

*a) Positioning Figures and Tables: In this paper, the figures and tables are strategically positioned to ensure clarity and maintain the flow of the content. The following guidelines were adhered to:*

* ***Placement****: All figures and tables are positioned at the* ***top*** *or* ***bottom*** *of the columns, ensuring that they do not disrupt the flow of the text. Figures and tables are not placed in the middle of the columns to avoid any interruptions in the document's readability.*
* ***Spanning Across Columns****: When necessary, larger figures or tables that contain complex data or need additional space span across* ***both columns*** *to maintain clarity. This is particularly useful when visualizations or large tables, such as performance metrics or loss curves, are presented.*
* ***Captions and Heads****:*
  + ***Figure captions*** *are placed* ***below*** *each figure. The abbreviation* ***“Fig. 1”*** *is used to refer to the figure, even at the beginning of a sentence. For example,* ***"Fig. 1 shows the performance of the ConvLSTM model."***
  + ***Table heads*** *(titles) appear* ***above*** *the tables. This ensures that the table is properly labeled and the reader knows what data to expect before viewing the table.*
  + ***Table footnotes*** *are included below the table, if necessary, for additional explanations or clarifications.*
* ***Citations in the Text****: Figures and tables are inserted after they have been cited in the text. This allows the reader to first encounter the reference to the data or graphical representation before seeing the figure or table itself.*

TABLE I

Model Result Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Test Loss (MSE) | Average SSIM | Training Time (hrs) |
| ConvLSTM | 0.0092 | 0.8585 | 1hr |
| PredRNN | 0.0084 | 0.8059 | 1hr |
| Transformer | 0.0378 | 0.0883 | 2hr |



*{ }*

ACKNOWLEDGMENT

The authors would like to thank Fast university Special thanks to Teacher Khadjia Mehmood for their generous support of this research. This work would not have been possible without the resources provided

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, et al., “TensorFlow: A system for large-scale machine learning,” in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, Savannah, GA, USA, Nov. 2016, pp. 265–283.

[2] R. H. Chan, J. Xie, and H. Zhang, “A deep learning approach to video action recognition with ConvLSTM and Transformer networks,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4551–4560.

[3] UCF101 Action Recognition Dataset, “UCF101: A dataset of 101 human action categories,” [Online]. Available: http://www.crcv.ucf.edu/data/UCF101.php

[4] F. Chollet, “Keras: The Python deep learning library,” *GitHub*, 2015. [Online]. Available:<https://github.com/keras-team/keras>

[5] Flask Web Framework, “Flask: A micro web framework for Python,” [Online]. Available: https://flask.palletsprojects.com/

[6] M. Young, *The Technical Writer’s Handbook*. Mill Valley, CA: University Science, 1989.