Multimodal Deep Regression on TikTok Content Success

Louis Wong
Georgia Tech
lwong64@gatech.edu

Mingyao Song
Georgia Tech
msong41@gatech.edu

Jason Xu
Georgia Tech
jxu623@gatech.edu

Ahmed Salih
Georgia Tech
asalih6@gatech.edu

Abstract

Content creators grapple with the challenge of predicting if their time and investments will translate into increased viewership and audience growth, a task made more complex by the hidden algorithms and unpredictable audience interaction of social media platforms. This research's objective is to architect a model that predicts video success, effectively indicating a video's potential virality. By employing advanced convolution techniques for video encoding, leveraging strides in natural language processing models, we're pushing boundaries in deep video content analysis. We construct a powerful multimodal ensemble model for general video content regression, capable of comprehending human-generated content and accurately predicting its nonlinear relationship elements to determine success. Our preliminary results demonstrate the model's effectiveness in predicting video virality, showcasing the potential of our innovative approach.

1. Introduction

The landscape of digital content creation is continuously evolving, presenting content creators with the challenge of predicting the success of their videos in terms of viewership and audience growth. This challenge is further complicated by the opaque algorithms and unpredictable audience interactions on social media platforms like TikTok. To address this issue and empower content creators, various methods and models have been explored for predicting video success and understanding the factors contributing to content virality.

To address the challenges of interpreting and predicting YouTube viewership, Liu et al. [1] proposed a novel Precise Wide-and-Deep Learning model. This model accurately predicts viewership using unstructured video data and established features while providing precise interpretations of feature effects. In the context of TikTok, where content and user preferences are continually evolving, researchers [2] delved into predicting user participation in TikTok challenges. They introduced a novel deep learning model capable of learning and combining latent user and challenge representations from past videos to predict a user's likelihood of participating in a challenge. Salvador et al. [3] studied on human action recognition and explored the use of attention mechanisms to improve the accuracy and efficiency of video recognition models. In their work, they integrated space and time attention mechanisms into the framework of Vision Transformer network structure for feature extraction from video data. To effectively understand and recommend short videos, a multi-modal fusion framework was proposed [4], integrating features from different modalities to capture inherent relationships. Deep neural networks were employed for feature extraction and fusion to accomplish video understanding and recommendation tasks. An approach to describing videos using multimodal fusion techniques was explored [5]. The research presented a deep neural network that combined visual and textual information at various stages in the network, aiming to learn a joint representation for video description tasks. Additionally, researchers [6] focused on predicting the popularity of images and videos on Instagram. They employed convolutional neural networks and long short-term memory networks to extract spatial and temporal information from images and videos, respectively, and used a regression model to predict popularity. Another work investigated popularity prediction on Instagram using neural networks and regression analysis [7]. The authors explored the predictive power of image composition on Instagram posts by comparing the popularity predictions of neural networks trained on aesthetic value with predictions from regression models using social metadata.

While these work makes significant contributions, research focused on multi-modal fusion encounters the challenge of effectively integrating information from diverse modalities, such as visual and audio data, to achieve a cohesive and informative representation. The success of fusion techniques heavily relies on striking the appropriate balance between these modalities and ensuring their seamless integration.

In multimodal learning, there are two main fusion approaches: early fusion, which concatenates original or extracted features at the input level, and late fusion, which aggregates predictions at the decision level. The performance comparison between early and late fusion is influenced by various factors [8,9], such as the characteristics of the multimodal data, the complexity of the task, the interdependence between modalities, the quality of the features, the architecture of the network, the size of the dataset, and the availability of labeled data. Therefore, there is no definitive answer regarding which fusion approach is universally superior. Each approach can prove to be more effective in specific scenarios. In this study, we have chosen the late fusion approach. Through careful evaluation and analysis of the aforementioned factors specific to our project, we have determined that late fusion offers distinct advantages. By combining predictions at the decision level, we can leverage the strengths of each modality effectively, allowing us to capture complementary information and improve overall performance

In this study, our goal is to predict video virality using a multimodal ensemble model. To achieve our goal, we collected a diverse dataset of approximately 5,000 videos from TikTok, covering a wide range of hashtag topics, such as Sports, Dance, Entertainment, Comedy, Autos, Fashion, Lifestyle, Pets and Nature, Relationships, Society, Informative, and Music. The dataset includes relevant metadata for each video, such as video IDs, TikTok URLs, and video view counts.

Our approach involves data preparation and multimodal deep regression modeling. The data preparation process includes scraping video data from TikTok, audio extraction, and meticulous organization of visual tensors for efficient integration into the model. For audio analysis, we leverage the open-source Whisper model to transcribe audio content and obtain audio embeddings that capture the semantic meaning of the audio. The heart of our model lies in the visual embeddings generated through an unsupervised pretraining process using a ConvLSTM Autoencoder. This process encodes the context of the video into compact and informative embeddings that retain essential spatial and temporal features. Subsequently, the visual and audio embeddings are concatenated and fed into a Transformer-based regression model for multimodal analysis. The late fusion technique combines the visual and audio data, enabling the model to learn the semantic and nonlinear relationships that contribute to video creator success. The Transformer model, with its self-attention layers and feed-forward neural networks, captures complex patterns and relationships within the data.

This research can benefit various stakeholders in the social media ecosystem. Content creators, including influencers, advertisers, and artists, can gain valuable insights into the potential success of their videos before investing time and resources. The predictive ability can help creators can optimize their content strategies, increase their audience reach, and potentially experience viral success. This research will empower creators to make data-driven decisions, improve their content's impact, and enhance their overall presence on social media platforms.

In the following sections, we will delve into the specifics of our implemented model, including its training process and the results showcasing its performance.

(5 points) What did you try to do? What problem did you try to solve? Articulate your objectives using absolutely no jargon.

(5 points) How is it done today, and what are the limits of current practice?

(5 points) Who cares? If you are successful, what difference will it make?

(5 points) What data did you use? Provide details about your data, specifically choose the most important aspects of your data mentioned here. You don't have to choose all of them, just the most relevant.

2. Approach

2.1. Data collection and preparation

The data collection process involved scraping video data from TikTok using a custom-built Selenium scraper running on a Chromium browser. The scraper was designed to randomly collect videos across various hashtag topics, ensuring a diverse representation of content. The selected hashtag topics included Sports, Dance, Entertainment, Comedy and Drama, Autos, Fashion, Lifestyle, Pets and Nature, Relationships, Society, Informative, and Music. This approach aimed to capture a wide range of video content to ensure the model's generalizability. In total, the data collection effort yielded approximately 5,000 videos, each in .mp4 format, resulting in a substantial dataset with a total size of 32.7 GB. To facilitate further analysis and model training, each video was assigned a unique video ID tag for easy reference and organization. Alongside the video files, the dataset also includes JSON data containing relevant metadata for each video, such as the video ID tag and its corresponding TikTok URL. Additionally, the video view count on TikTok was also recorded as an essential metric for evaluating video creator success.

Data preparation is a critical aspect in the training of multimodal deep regression models, encompassing several pivotal steps, such as data loading, preprocessing, and splitting, to achieve optimal outcomes. Initially, audio extraction from video datasets is performed, with the extracted audio files thoughtfully organized into an audio directory. Simultaneously, the visual data undergoes meticulous processing, involving frame skip, shrink scale adjustments, and normalization to enhance consistency and compatibility. The resultant processed visual tensors are then meticulously stored in their respective directories, poised for seamless integration into subsequent training and validation stages, paving the way for an efficient and robust multimodal deep regression model. To maintain a clean and organized workflow, a virtual environment is set up outside the project folder to prevent conflicts with other packages. We leverage appropriate utilities for loading and processing the data, including audio extraction, transcribing, and obtaining embeddings. data transformations such as normalization and frame skipping are applied to enhance the data quality. The dataset is split into training and validation sets to evaluate the model's performance effectively. Data loaders are created to enable efficient batching during the training process, leading to smoother and more effective training of the model.

2.2. Model

The video is first broken down into visual and audio tracks, with our primary focus on the video visual. The video visual will undergo an autoencoder process, utilizing a convolution-based network architecture, ConvLSTM Autoencoder, to unsupervised pre-training from scratch. This process encodes the context of the video into embedding vectors. Subsequently for the audio, we leverage a pretrained model, the open-source Whisper, to create a transcript for the audio, supplementing the project. Afterward, visual embeddings and audio embeddings are respectively extracted from the visual branch and the audio branch, which are then be concatenated and input into a Transformer-based regression model. The aim is for this model to learn the semantic and non-linear relationships needed to predict video creator success metrics, such as video views. Finally, we establish a baseline using a less complex model and compare it with our main implementation to evaluate its success.

2.2.1 Visual Embedding

To generate compact and informative visual embeddings, an unsupervised pretraining method using a ConvLSTM Autoencoder is employed. The Autoencoder architecture consists of ConvLSTM layers, which are a combination of Convolutional and Long Short-Term Memory (LSTM) layers. The ConvLSTM layers are used to process the video frames, capturing both the spatial features (through the convolutional layers) and the temporal dependencies (through the LSTM layers). This allows the Autoencoder to retain important contextual information from the video data while reducing the dimensionality to create compact embeddings.

During training, the Autoencoder takes video frames as input and tries to reconstruct them at the output. The difference between the original and reconstructed frames is quantified using the Mean Squared Error (MSE) loss function. Through backpropagation, the Autoencoder adjusts its weights and biases to minimize this reconstruction error, thus learning to capture meaningful patterns in the video data. The trained Autoencoder is evaluated thoroughly over multiple epochs to ensure that it learns robust and meaningful embeddings. In each epoch, the Autoencoder is tested on both training and validation data, and the loss values are recorded to monitor the training progress. The embeddings generated by the Autoencoder represent the visual context of the video. These embeddings condense the raw visual data into a more informative and compact representation, which is crucial for downstream visual data analysis.

2.2.2 Audio Embedding

Audio embeddings are obtained using the open-source Whisper model through unsupervised pretraining. Whisper is a powerful automatic speech recognition (ASR) system developed by OpenAI to convert spoken language into text. First, the audio is extracted from the video dataset, and then Whisper transcribes the audio dialog, producing textual transcripts that capture essential information from the spoken content. These textual outputs serve as audio embeddings, effectively encapsulating the semantic meaning and characteristics of the audio. Utilizing the pre-existing Whisper model for audio embedding generation offers several advantages. The Whisper model is already trained on extensive speech data, making it effective in understanding diverse spoken language patterns. This saves the effort and time required to train an ASR system from scratch, making the process more efficient. The resulting audio embeddings play a pivotal role in augmenting audio analysis and comprehension capabilities, facilitating downstream tasks that demand a profound understanding of the audio content.

2.2.3 Transformer-based Ensemble Model

Transformer-based ensemble model is employed to perform multimodal deep regression by combining visual and audio data using late fusion technique. The ensemble model is composed of two main components: visual Transformer model and audio Transformer model. Both models utilize the Transformer architecture, which consists of multiple self-attention layers and feed-forward neural networks. These models take visual and audio embeddings as inputs, respectively, and process them using the Transformer layers to capture complex patterns and relationships within the data. Positional encodings are added to the input data, providing positional information to the Transformer model. The positional encodings are calculated using sine and cosine functions with varying frequencies.

The ensemble model is trained on the combined data using Mean Squared Error (MSE) loss and the Adam optimizer for a specified number of epochs. The trained model is evaluated on the validation set, and the MSE between the predicted values and ground truth values is calculated, providing an indication of the model's performance. The model generates plots to visualize the training and validation loss during the training of the Ensemble Model. Finally, the model inspects and compares the ground truth and predicted values for a random sample from the validation set, and the MSE value between the two sets of values is obtained.

2.3. Loss Function

The model is trained using the mean squared error (MSE) loss function, which serves as a measure of the discrepancy between the model's predicted values and the actual ground truth video creator success metrics. The primary objective is to minimize this discrepancy and achieve a high level of accuracy in predicting the success metrics. The MSE loss function calculates the average squared difference between the predicted values and the actual values. By squaring the differences, it penalizes larger errors more severely, emphasizing the importance of accurate predictions across all data points. MSE loss function is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where n is the total number of samples in the dataset. y_i represents the actual value of the success metric for the i^{th} video creator. $\hat{y}_i$ represents the predicted value of the success metric for the i^{th} video creator.

2.4. Implementation

We utilized PyTorch (version xxx) and Python 3.xxx to implement our model. Hyperparameters.....

(10 points) What did you do exactly? How did you solve the problem? Why did you think it would be successful? Is anything new in your approach?

(5 points) What problems did you anticipate? What problems did you encounter? Did the very first thing you tried work?

Important: Mention any code repositories (with citations) or other sources that you used, and specifically what changes you made to them for your project.

3. Experiments and Results

3.1. Experiment Setup

ABCDE

3.2. Results

ABCDE

(10 points) How did you measure success? What experiments were used? What were the results, both quantitative and qualitative? Did you succeed? Did you fail? Why? Justify your reasons with arguments supported by evidence and data.

Important: This section should be rigorous and thorough. Present detailed information about decision you made, why you made them, and any evidence/experimentation to back them up. This is especially true if you leveraged existing architectures, pre-trained models, and code (i.e. do not just show results of fine-tuning a pre-trained model without any analysis, claims/evidence, and conclusions, as that tends to not make a strong project).

4. Conclusion and Future Improvements

You are welcome to introduce additional sections or subsections, if required, to address the following questions in detail.

(5 points) Appropriate use of figures / tables / visualizations. Are the ideas presented with appropriate illustration? Are the results presented clearly; are the important differences illustrated?

(5 points) Overall clarity. Is the manuscript self-contained? Can a peer who has also taken Deep Learning understand all of the points addressed above? Is sufficient detail provided?

(5 points) Finally, points will be distributed based on your understanding of how your project relates to Deep Learning. Here are some questions to think about:

What was the structure of your problem? How did the structure of your model reflect the structure of your problem?

What parts of your model had learned parameters (e.g., convolution layers) and what parts did not (e.g., post-processing classifier probabilities into decisions)?

What representations of input and output did the neural network expect? How was the data pre/post-processed? What was the loss function?

Did the model overfit? How well did the approach generalize?

What hyperparameters did the model have? How were they chosen? How did they affect performance? What optimizer was used?

What Deep Learning framework did you use?

What existing code or models did you start with and what did those starting points provide?

Briefly discuss potential future work that the research community could focus on to make improvements in the direction of your project's topic.

5. Work Division

Please add a section on the delegation of work among team members at the end of the report, in the form of a table and paragraph description. This and references do **NOT** count towards your page limit. An example has been provided in Table 1.

Student Name	Contributed Aspects	Details
Louis Wong	ABCDE	ABCDE
Mingyao Song	ABCDE	ABCDE
Jason Xu	ABCDE	ABCDE
Ahmed Salih	ABCDE	ABCDE

Table 1. Contributions of team members.

6. Miscellaneous Information

The rest of the information in this format template has been adapted from CVPR 2020 and provides guidelines on the lower-level specifications regarding the paper's format.

6.1. Language

All manuscripts must be in English.

6.2. Paper length

Papers, excluding the references section, must be no longer than six pages in length. The references section will not be included in the page count, and there is no limit on the length of the references section. For example, a paper of six pages with two pages of references would have a total length of 8 pages.

6.3. The ruler

The LaTeX style defines a printed ruler which should be present in the version submitted for review. The ruler is provided in order that reviewers may comment on particular lines in the paper without circumlocution. If you are preparing a document using a non-LaTeX document preparation system, please arrange for an equivalent ruler to appear on the final output pages. The presence or absence of the ruler should not change the appearance of any other content on the page. The camera ready copy should not contain a ruler. (LaTeX users may uncomment the `\cvprfinalcopy` command in the document preamble.) Reviewers: note that the ruler measurements do not align well with lines in the paper — this turns out to be very difficult to do well when the paper contains many figures and equations, and, when done, looks ugly. Just use fractional references (e.g. this line is 095.5), although in most cases one would expect that the approximate location will be adequate.

6.4. Mathematics

Please number all of your sections and displayed equations. It is important for readers to be able to refer to any particular equation. Just because you didn't refer to it in the text doesn't mean some future reader might not need to refer to it. It is cumbersome to have to use circumlocutions like "the equation second from the top of page 3 column 1". (Note that the ruler will not be present in the final copy, so is not an alternative to equation numbers). All authors will benefit from reading Mermin's description of how to write mathematics: http://www.pamitc.org/documents/mermin.pdf.

Finally, you may feel you need to tell the reader that more details can be found elsewhere, and refer them to a technical report. For conference submissions, the paper must stand on its own, and not *require* the reviewer to go to a techreport for further details. Thus, you may say in the body of the paper "further details may be found in [5]". Then submit the techreport as additional material. Again, you may not assume the reviewers will read this material.

Sometimes your paper is about a problem which you tested using a tool which is widely known to be restricted to a single institution. For example, let's say it's 1969, you have solved a key problem on the Apollo lander, and you believe that the CVPR70 audience would like to hear about your solution. The work is a development of your celebrated 1968 paper entitled "Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties", by Zeus *et al*.

You can handle this paper like any other. Don't write "We show how to improve our previous work [Anonymous, 1968]. This time we tested the algorithm on a lunar lander [name of lander removed for blind review]". That would be silly, and would immediately identify the authors. Instead write the following:

> We describe a system for zero-g frobnication. This system is new because it handles the following cases: A, B. Previous systems [Zeus et al. 1968] didn't handle case B properly. Ours handles it by including a foo term in the bar integral.
>
> ...
>
> The proposed system was integrated with the Apollo lunar lander, and went all the way to the moon, don't you know. It displayed the following behaviours which show how well we solved cases A and B: ...

As you can see, the above text follows standard scientific convention, reads better than the first version, and does not explicitly name you as the authors. A reviewer might think it likely that the new paper was written by Zeus *et al*., but cannot make any decision based on that guess. He or she would have to be sure that no other authors could have been contracted to solve problem B.

FAQ

Q: Are acknowledgements OK?
A: No. Leave them for the final copy.

Q: How do I cite my results reported in open challenges?
A: To conform with the double blind review policy, you can report results of other challenge participants together with your results in your paper. For your results, however, you should not identify yourself and should not mention your participation in the challenge. Instead present your results referring to the method proposed in your paper and draw conclusions based on the experimental comparison to other results.

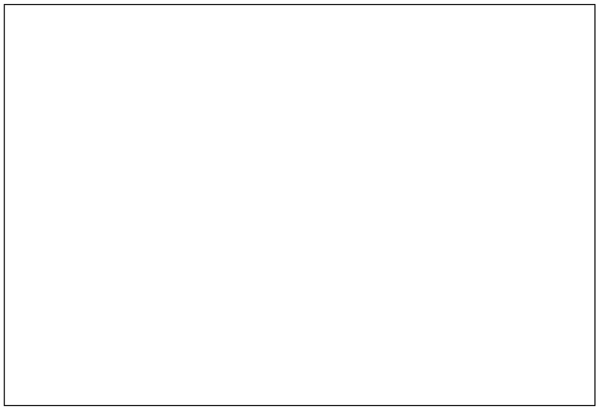

Figure 1. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

6.5. Miscellaneous

Compare the following:

`$conf_a$`	$conf_a$
`$\mathit{conf}_a$`	conf_a

See The TEXbook, p165.

The space after *e.g.*, meaning "for example", should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using "et alia", shortened to "*et al.*" (not "*et. al.*" as "*et*" is a complete word.) However, use it only when there are three or more authors. Thus, the following is correct: " Frobnication has been trendy lately. It was introduced by Alpher [1], and subsequently developed by Alpher and Fotheringham-Smythe [2], and Alpher *et al*. [3]."

This is incorrect: "... subsequently developed by Alpher *et al*. [2] ..." because reference [2] has just two authors. If you use the `\etal` macro provided, then you need not worry about double periods when used at the end of a sentence as in Alpher *et al*.

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [2, 1, 4] to [1, 2, 4].

6.6. Formatting your paper

All text must be in a two-column format. The total allowable width of the text area is $6\frac{7}{8}$ inches (17.5 cm) wide by

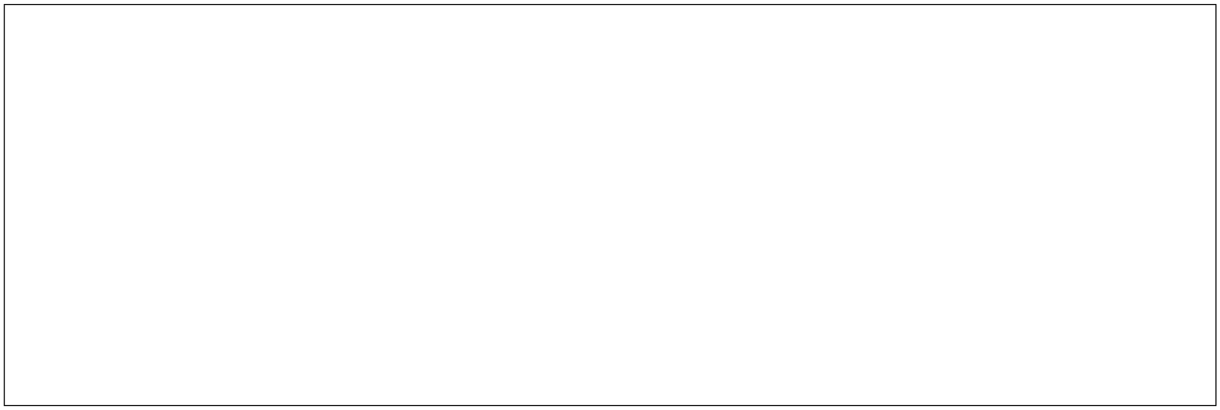

Figure 2. Example of a short caption, which should be centered.

$8\frac{7}{8}$ inches (22.54 cm) high. Columns are to be $3\frac{1}{4}$ inches (8.25 cm) wide, with a $\frac{5}{16}$ inch (0.8 cm) space between them. The main title (on the first page) should begin 1.0 inch (2.54 cm) from the top edge of the page. The second and following pages should begin 1.0 inch (2.54 cm) from the top edge. On all pages, the bottom margin should be 1-1/8 inches (2.86 cm) from the bottom edge of the page for 8.5×11-inch paper; for A4 paper, approximately 1-5/8 inches (4.13 cm) from the bottom edge of the page.

6.7. Margins and page numbering

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high.

6.8. Type-style and fonts

Wherever Times is specified, Times Roman may also be used. If neither is available on your word processor, please use the font closest in appearance to Times to which you have access.

MAIN TITLE. Center the title 1-3/8 inches (3.49 cm) from the top edge of the first page. The title should be in Times 14-point, boldface type. Capitalize the first letter of nouns, pronouns, verbs, adjectives, and adverbs; do not capitalize articles, coordinate conjunctions, or prepositions (unless the title begins with such a word). Leave two blank lines after the title.

AUTHOR NAME(s) and AFFILIATION(s) are to be centered beneath the title and printed in Times 12-point, non-boldface type. This information is to be followed by two blank lines.

The ABSTRACT and MAIN TEXT are to be in a two-column format.

MAIN TEXT. Type main text in 10-point Times, single-spaced. Do NOT use double-spacing. All paragraphs should be indented 1 pica (approx. 1/6 inch or 0.422 cm). Make sure your text is fully justified—that is, flush left and flush right. Please do not place any additional blank lines between paragraphs.

Figure and table captions should be 9-point Roman type as in Figures 1 and 3. Short captions should be centred.
Callouts should be 9-point Helvetica, non-boldface type. Initially capitalize only the first word of section titles and first-, second-, and third-order headings.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**) should be Times 12-point boldface, initially capitalized, flush left, with one blank line before, and one blank line after.

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**) should be Times 11-point boldface, initially capitalized, flush left, with one blank line before, and one after. If you require a third-order heading (we discourage it), use 10-point Times, boldface, initially capitalized, flush left, preceded by one blank line, followed by a period and your text on the same line.

6.9. Footnotes

Please use footnotes[1] sparingly. Indeed, try to avoid footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence). If you wish to use a footnote, place it at the bottom of the column on the page on which it is referenced. Use Times 8-point type, single-spaced.

7. References

[1] Jiaheng Xie, Yidong Chai, and Xiao Liu. ”Unbox the Black-Box: Predict and Interpret YouTube Viewership Using Deep Learning.” Journal of Management Information Systems, 2023, 541-579.

[2] Lynnette Hui Xian Ng, John Yeh Han Tan, Darryl Jing Heng Tan, Roy Ka-Wei Lee. Will you dance to the challenge? Predicting user participation of TikTok challenges.

[1] This is what a footnote looks like. It often distracts the reader from the main flow of the argument.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 2. Results. Ours is better.

In Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM'21), The Hague, 2021, 356–360. New York: ACM.

[3] Qiliang Chen, Hasiqidalatu Tang, and Jiaxin Cai. "Human Action Recognition Based on Vision Transformer and L2 Regularization." In Proceedings of the 2022 11th International Conference on Computing and Pattern Recognition (ICCPR '22), 2022, 224-228.

[4] Daya Guo, Jiangshui Hong, Binli Luo, Qirui Yan, Zhangming Niu. "Multi-modal representation learning for short video understanding and recommendation." In 2019 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2019, 687-690.

[5] Jin Q., Chen J., Chen S., Xiong Y., and Hauptmann A. "Describing videos using multi-modal fusion." In ACM MM, 2016.

[6] Massimiliano Viola, Luca Brunelli, and Gian Antonio Susto. "Instagram Images and Videos Popularity Prediction: a Deep Learning-Based Approach." Università degli Studi di Padova, Padova, IT.

[7] Qian C., Tang J., Penza M., and Ferri C. "Instagram Popularity Prediction via Neural Networks and Regression Analysis," 2017, 2561-2570.

[8] Snoek, C. G. M., Worring, M., and Smeulders, A. W. M. "Early versus late fusion in semantic video analysis." In Proceedings of the Annual ACM International Conference on Multimedia, Singapore, 2005, 399-402.

[9] Gadzicki, K.; Khamsehashari, R.; Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In Proceedings of the 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 2020; pp. 1–6.

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example [4]. Where appropriate, include the name(s) of editors of referenced books.

7.1. Illustrations, graphs, and photographs

All graphics should be centered. Please ensure that any point you wish to make is resolvable in a printed copy of the paper. Resize fonts in figures to match the font in the body text, and choose line widths which render effectively in print. Many readers (and reviewers), even of an electronic copy, will choose to print your paper in order to read it. You cannot insist that they do otherwise, and therefore must not assume that they can zoom in to see tiny details on a graphic.

When placing figures in LaTeX, it's almost always best to use \includegraphics, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
                {myfile.eps}
```

References

[1] FirstName Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002. 6

[2] FirstName Alpher and FirstName Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003. 6

[3] FirstName Alpher, FirstName Fotheringham-Smythe, and FirstName Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004. 6

[4] Authors. The frobnicatable foo filter, 2014. Face and Gesture submission ID 324. Supplied as additional material fg324.pdf. 6, 8

[5] Authors. Frobnication tutorial, 2014. Supplied as additional material tr.pdf. 6

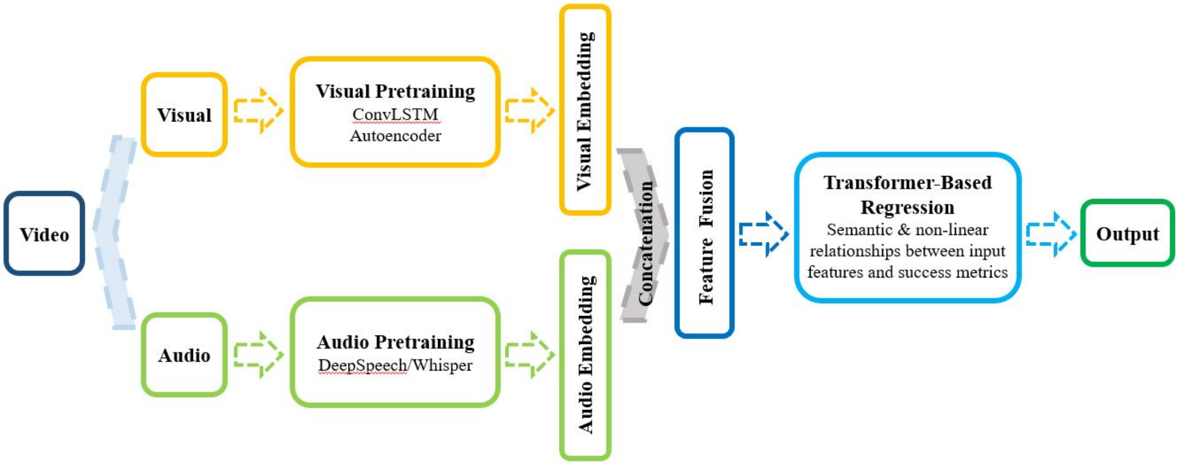

Figure 3. Illustration of the model