Successful Youtube video identification using multimodal deep learning

Lucas de Souza Rodrigues[1], Kenzo Sakiyama[2], Leozitor Floro de Souza[2], Edson Takashi Matsubara[1], Bruno Nogueira[1]

[1] Universidade Federal de Mato Grosso do Sul, Brazil
lucas.rodrigues@ifms.edu.br; leozitor@gmail.com; edsontm@facom.ufms.br; bruno@facom.ufms.br
[2] Universidade de São Paulo, Brazil
kenzosakiyama@usp.br

Abstract. Text from titles and audio transcriptions, image thumbnails, number of likes, dislikes, and views are examples of available data in a YouTube video. Despite the variability, most standard Deep Learning models use only one type of data. Moreover, the simultaneous use of multiple data sources for such problems is still rare. To shed light on these problems, we empirically evaluate eight different multimodal fusion operations using embeddings extracted from image thumbnails and video titles of YouTube videos using standard Deep Learning models, ResNet-based SE-Net for image feature extraction, and BERT to NLP. Experimental results show that simple operations such as sum or subtract embeddings can improve the accuracy of models. The multimodal fusion operations in this dataset achieved 81.3% accuracy, outperforming the unimodal models by 3.86% (text) and 5.79% (video).

Categories and Subject Descriptors: I.2.6 [**Artificial Intelligence**]: Learning; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.11 [**Artificial Intelligence**]: Vision and Scene Understanding

Keywords: multimodal, fusion, deep learning

1. INTRODUCTION

When content creators publish a new video, they must include metadata about their publication, such as title, thumbnail, and description. The success of a publication on social video networks is correlated with the quality of this metadata [Zhou et al. 2010]. In particular, the video's thumbnail and title must capture the audience's attention to generate new views [Carta et al. 2022]. A common practice among these content creators is to change the title and thumbnails from time to time and evaluate whether a new arrangement draws more attention and clicks to this video [Teng et al. 2018]. Finding compelling thumbnails and titles has become critical for YouTube content creators.

Similar problems have already been tackled in the literature. The study conducted in [Song et al. 2016] evaluates YouTube thumbnails to select visually attractive images using a clustering algorithm using stillness as a distance metric. Stillness uses visual aesthetic features such as color, texture, contrast, exposure, and composition. When looking for studies that use YouTube video titles, the study conducted in [Kalra et al. 2019] uses titles and descriptions to classify videos from travel, science, food, manufacturing, history, and art. The authors used the standard bag-of-words representation and Random Forest as the learning algorithm. The study [Islam et al. 2021] uses YouTube video titles to predict exaggerated titles in Bengali language. The authors used bag-of-words on a tfidf to represent the titles in a table format to feed standard machine learning algorithms, decision trees, random forest, Naive Bayes, logistic regression, multi-layer perceptron [Witten and Frank 2002], and convolutional neural network [Goodfellow et al. 2016].

Taking this context into account, several questions arise at this point. What impact does the title have on this brand? Can the same question be asked for video thumbnails? Or is it the right

combination of both? The literature has not addressed these issues considering the proposed modeling problem. From a machine learning point of view, a possible attempt to answer these questions can be made by building models with a representation layer by merging data from multiple sources to generate a richer representation of the model. In this article, we propose the implementation of a multimodal model that combines the independent characteristics of unimodal modalities through textual and visual data flows. We then merge the modalities to predict successful YouTube videos based on the number of views. Figure 1 shows the steps of the proposal in this study.

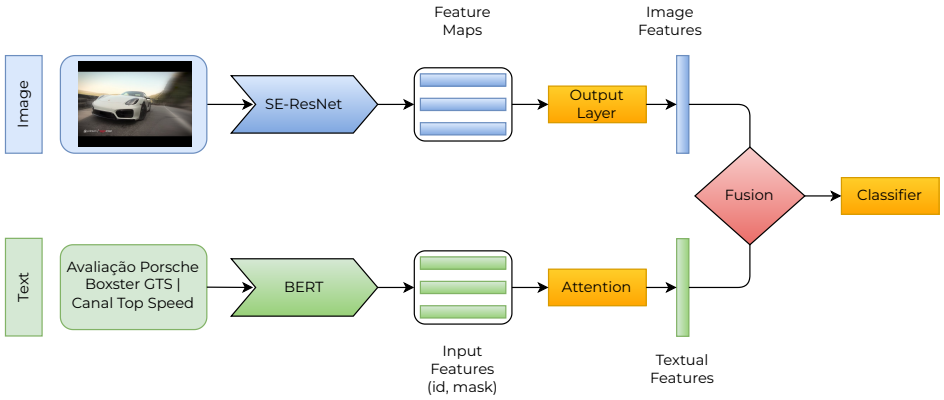

Fig. 1. Multimodal Fusion with BERT and SE-ResNet

To describe our multimodal deep learning approach, we used data from the popular Brazilian motorsport YouTube channel (former Top Speed, currently named *Mobiauto*)[1]. This YouTube channel has around 800k subscribers. The main contributions of this work are the following: (i) Experimental evaluation of BERT and SE-ResNet using titles and video thumbnails to predict successful video posts; (ii) Experimental evaluation of eight fusion methods; (iii) Construction of a joint representation or fusion layers that can merge incoming representations of modalities; (iv) General recommendations for predicting successful videos using thumbnails and titles as input.

2. MULTIMODAL LEARNING

Data can be found in various formats, such as spreadsheets, images, texts and videos. Multimodal deep learning consists of techniques that seek to unify different data modalities. According to [Ramachandram and Taylor 2017], deep architectures offer the flexibility to implement multimodal merging as early, middle, or late merging. Figure 2 illustrates traditional types of multimodal fusion that span multiple application domains and allow for the combination of multiple data sources. This section briefly reviews the fundamental methods/algorithms used in multimodal learning, we also list works described in the literature that configure some of the methods of multimodal learning.

Early fusion - The first approach consists of combining data of different modalities before applying a neural network model. This method is known as entry-level fusion, and it applies to raw data or pre-processed data obtained from different modalities. In this scenario, the neural networks observe the combined data and learn the modalities' complementary strengths and weaknesses. A classic example of early fusion is the work of [Poria et al. 2015], using vectors of combined features of textual, visual, and audio modalities to train a classifier based on multi-kernel learning for sentiment analysis.

Intermediate fusion - The intermediate fusion technique builds a shared representation of data from multiple modalities by merging the intermediate features obtained by separate machine learning models. The idea behind this technique is the fusion of representations from different modalities in a

[1] `https://www.youtube.com/user/CanalTopSpeed/`

single layer. The work [Joze et al. 2020] presents a simple neural network module to take advantage of the knowledge of multiple modalities in convolutional neural networks. The study uses different hierarchy levels of features, enabling slow modality fusion.

Late fusion - Late fusion uses unimodal decision values and fuses them with a fusion mechanism. This utilizes the individual potentials of unimodal classifiers. According to [Ramachandram and Taylor 2017], this fusion architecture is often favored because errors from multiple classifiers tend to be uncorrelated and the method is feature independent. When the data's dimensionality is significant concerning its modalities, late fusion is a more straightforward and flexible approach. Recent works employ late fusion for deep multimodal learning [Liu et al. 2018; Gadzicki et al. 2020; Trong et al. 2020] and show more promising results. Therefore, we proceeded with this study using this approach in our experiment.

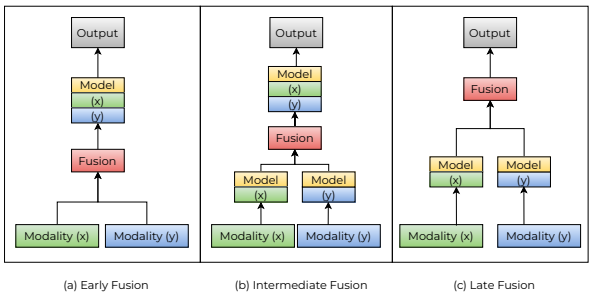

Fig. 2. Fusion Models for Multimodal Learning

3. PROPOSAL

This section will describe the steps used to build the proposed multimodal architecture. In this work, we proposed learning a joint representation using image and text data modalities present in YouTube videos through deep learning models. In order to measure video success, we establish the threshold in 100k views (12.5% of subscribers). After this mark, according to the video analyst, the video is more likely to monetize better and be successful for the audience of this channel. This limit is also associated with the balance and number of examples available in the dataset, there are a total of 462 videos with approximately 50% of these above 100k views. Another factor that prevents us from treating this work as a regression problem is that the videos have temporality and are not seasonal. We initially modeled this problem using the temporality of views over a 30-day time frame, but the variation in results was not significant for the experiments. Thus, we modeled the problem as a binary classification problem (positive $>$ 100k views and negative $<$ 100k views).

It should be noted that most previous work has focused on detecting visually attractive images in YouTube videos to generate video thumbnails. However, a visually appealing image does not necessarily mean that the video will be monetized or successful. According to our domain expert, even videos with poorly constructed and unattractive thumbnails can be quite successful and generate many views. Thus, to predict the success of a video, more information may need to be analyzed than just the attractiveness of the thumbnail.

3.1 Data Collection

The dataset for our experiments was obtained by an expert in video content and currently a motorsport YouTuber. Raw data were exported from Youtube[2] using the tool provided by the proprietary Google[3]: Youtube Analytics. It is important to note that the Youtube Analytics information is only available to the content creator himself. Therefore, the use of the data depends on the creator's authorization. In this work, we had access to the data of only one Youtuber.

[2]`http://www.youtube.com`
[3]`http://www.google.com`

Dataset construction - The extracted data have various information about the content of the Youtube platform (video title, description, publication time, views, CTR, watch time, subscribers, likes, dislikes, and comments). In our experiments, we used the video title to extract textual features. As image data input, we used thumbnails, whose term is used by graphic designers and photographers for a small representation of a larger image. We extracted the correspondent video thumbnail for each video uploaded by the content creator. This representation makes it easier and faster to look at or manage a group of videos on the online platform.

Pre-processing of data - We extracted information from around 500 Youtube videos from the content creator studied during the data collection step. We pre-processed the data, removing examples with missing data. In summary, 462 examples were used for the experiments in our study. The examples were separated into two classes in our approach (220 = non-successful and 242 = successful). The images were pre-processed as described in section 4 and we fed the text data into the model without further pre-processing.

3.2 Modalities embedding

Deep learning models can learn data representation [LeCun et al. 2015]. We use the term embeddings to represent the vector extracted from these model's final layers in this work. To extract high quality features, we choose state-of-the-art feature extractors for each modality. The selected algorithms to compute the embeddings of our approach are the following:

Squeeze-and-Excitation Networks (SE-Net) - In computer vision, the common approach to feature extraction is to use Convolutional Neural Networks (CNN's) to extract features (embeddings) from the input images. The CNN's are commonly pre-trained in a large corpus of images and the extracted features are used as inputs of linear feed forward neural networks. In this work, we used a ResNet [He et al. 2016] based SE-Net [Hu et al. 2018] to extract image features. The SE-Net's adaptively recalibrates channel-wise feature responses by explicitly modelling inter dependencies between channels. Using the SE-ResNet50 model [Wightman 2019], pre-trained on Imagenet, we extracted image embeddings of size 1×1000.

Bidirectional Encoder Representations from Transformers (BERT) - Similar to computer vision, current state-of-the-art classifiers in natural language processing use powerful Transformer encoders to extract features from text data [Devlin et al. 2018]. The Transformer model is capable of generating context-aware representations for text tokens, utilizing a powerful attention framework [Vaswani et al. 2017]. We choose `Bertimbau-base` as our text feature extractor [Souza et al. 2020]. The chosen model was pre-trained in a large brazilian-portuguese corpus, and we used the first token (the "[CLS]" token) hidden-state (a 1×768 tensor) as our text embeddings. This hidden-state representation is often used for sequence classification tasks.

3.3 Fusion

When considering multimodal literature using YouTube, the study [Yu and Shi 2020] concatenates feature embeddings of frames, titles, descriptions, and audios to select visually attractive images. The concatenated feature is processed by context gating similar to the self-attention modules of the transformer [Vaswani et al. 2017] and submitted to fully connected layers. In addition to concatenation in our work, we propose the use of seven arithmetic operations to merge textual and visual data, as we believe that the use of simple operations can have significant results regarding concatenation.

Let us consider $\mathbf{z}^{img} = f(\mathbf{x}^{img})$ a mapping from an image $\mathbf{x}^{img}$ to an embedding space $\mathbf{z}^{img} \in \mathrm{R}^d$ where d denotes the dimension of this image mapping. Taking text as input, let $\mathbf{z}^{text} = g(\mathbf{x}^{text})$ be the mapping from a text $\mathbf{x}^{text}$ to an embedding space $\mathbf{z}^{text} \in \mathrm{R}^n$ where n denotes the dimension of this text mapping.

Therefore, we define the concatenation fusion operation as:

$$Concatenate(\mathbf{z}^{text}, \mathbf{z}^{img}) = \mathbf{z}^{text} \frown \mathbf{z}^{img} \tag{1}$$

Note that the concatenation allow the operation of two embeddings with different dimensions. When a operation requires same number of dimensions we define a fully connected layer $fc : \mathrm{R}^d \rightarrow \mathrm{R}^n$. For the purpose of this study we defined $\mathbf{rz}^{img} = fc(\mathbf{z}^{img})$.

The other fusion operations explored in this study are defined in following equations:

$$Division(\mathbf{z}^{text}, \mathbf{rz}^{img}) = \sum_{i=1}^{n} \frac{z_i^{text}}{rz_i^{img}} \quad (2)$$

$$Maximum(\mathbf{z}^{text}, \mathbf{rz}^{img}) = \max_{1 \leq i \leq n} \begin{cases} z_i^{text}, & \text{if } z_i^{text} > rz_i^{img}. \\ rz_i^{img}, & \text{otherwise.} \end{cases} \quad (3)$$

$$Minimum(\mathbf{z}^{text}, \mathbf{rz}^{img}) = \min_{1 \leq i \leq n} \begin{cases} z_i^{text}, & \text{if } z_i^{text} < rz_i^{img}. \\ rz_i^{img}, & \text{otherwise.} \end{cases} \quad (4)$$

$$Multiplication(\mathbf{z}^{text}, \mathbf{rz}^{img}) = \sum_{i=1}^{n} z_i^{text} * rz_i^{img} \quad (5)$$

$$Subtraction(\mathbf{z}^{text}, \mathbf{rz}^{img}) = \sum_{i=1}^{n} (z_i^{text} - rz_i^{img}) \quad (6)$$

$$Sum(\mathbf{z}^{text}, \mathbf{rz}^{img}) = \sum_{i=1}^{n} (z_i^{text} + rz_i^{img}) \quad (7)$$

$$Power(\mathbf{z}^{text}, \mathbf{z}^{img}) = (\mathbf{z}^{text} \frown \mathbf{z}^{img})^2 \quad (8)$$

Note that for the operations Division, Maximum, Minimum, Multiplication, Subtraction, Sum and Power, the input embeddings' dimensions must be the same and in this work $n = 768$ and $d = 1000$.

3.4 Classification

Some fusion operations require embeddings with the same dimensions; others do not. In this configuration, the fully connected layers must be compatible with the related operation after the fusion. In our setup, the image embeddings z^{img} have 1000 dimensions, while the text embedding z^{text} has 768. The concatenation is a straightforward process and results in a vector of 1768 dimensions. However, the element-wise operations require dimension compatibility.

To resize the embeddings, we use a fully connected layer, rz, resizing output embeddings of the image model ($1000_{dim} \rightarrow 768_{dim}$). After the fusion, we applied another fully connected layer to generate the output estimation, and at the end, we defined the loss function. The following section will compare the multimodal model with basis SOTA architectures using only images or text features to evaluate our method's effectiveness.

4. EXPERIMENTAL EVALUATION

We conducted eight experiments with the following objectives: analyze the various fusion methods, predict successful videos, and statistically compare the differences of the multimodal and unimodal models. Initially, we ran separately the unimodal models, then we merge your embeddings with the operations described in Section 3.3. Following the the standard process in the literature, we randomly resized and cropped the images to 224×224 for training and evaluation [Wu et al. 2015; Purushwalkam and Gupta 2016]. The experiments were performed on individual GPUs Nvidia Tesla M40 and Tesla K80 with memory usage limited to 24GB.

4.1 Models and Training Procedure

Our training protocol follows parts of [Ramachandram and Taylor 2017], which utilizes late fusion with the aggregation of decisions from multiple classifiers, each trained on separate modalities. The models did not have their layers frozen during training and their output dimensions were changed as described in the section 3.3. We performed a learning rate range test [Smith 2017] to find the highest learning rate that minimizes loss and does not cause it to explode. The learning to rate found was used as the upper bound of the one cycle policy [Smith and Topin 2019] used for training. The one cycle learning rate policy changes the learning rate after each batch. We used 10-fold cross-validation to evaluate the studied models, and each experiment used 100 epochs per fold. The metrics were evaluated at the end of the 100 epochs. Finally, each result is the average of the 10-folds. The AdamW [Loshchilov

and Hutter 2017] optimizer was selected for the experiments with learning rate (3e-5 <n <3e-2) and weight decay (1e-4).

Table I shows the results in terms of the chosen metrics (Cross entropy Loss, F1, Recall, Precision, Accuracy, and ROC-AUC). First, the results between text (BERTimbau) and image (SE-ResNet50) indicate that titles are more discriminative than the thumbnails in our setup. Second, our analysis reveals that late fusion using sub, sum, concat, max and min have higher performance than unimodal models. Pow, multi and div show lower performance than unimodal models.

Table I. Metrics of the analyzed models, using 10-folds cross validation.

Models	**Loss**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
BERTimbau	0.572 ± 0.052	0.774 ± 0.040	0.779 ± 0.039	0.774 ± 0.040	0.772 ± 0.041	0.774 ± 0.040
SE-ResNet50	0.588 ± 0.046	0.755 ± 0.067	0.758 ± 0.067	0.755 ± 0.066	0.754 ± 0.066	0.755 ± 0.066
Fusion	**Loss**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Sub	0.536 ± 0.068	0.813 ± 0.067	0.819 ± 0.068	0.813 ± 0.067	0.812 ± 0.067	0.813 ± 0.067
Sum	0.529 ± 0.068	0.809 ± 0.067	0.815 ± 0.068	0.809 ± 0.067	0.807 ± 0.067	0.809 ± 0.067
Concat	0.537 ± 0.065	0.794 ± 0.078	0.799 ± 0.080	0.795 ± 0.078	0.793 ± 0.078	0.795 ± 0.078
Max	0.540 ± 0.063	0.793 ± 0.041	0.798 ± 0.044	0.793 ± 0.040	0.793 ± 0.041	0.793 ± 0.040
Min	0.530 ± 0.067	0.791 ± 0.070	0.802 ± 0.076	0.790 ± 0.071	0.789 ± 0.071	0.790 ± 0.071
Pow	0.629 ± 0.079	0.702 ± 0.095	0.709 ± 0.102	0.699 ± 0.097	0.696 ± 0.098	0.699 ± 0.097
Multi	0.690 ± 0.016	0.531 ± 0.124	0.532 ± 0.149	0.526 ± 0.120	0.508 ± 0.130	0.526 ± 0.120
Div	0.697 ± 0.006	0.493 ± 0.050	0.478 ± 0.141	0.492 ± 0.040	0.402 ± 0.058	0.492 ± 0.040

The higher performance is achieved using the *Subctration* operation. Figure 3 shows the metrics adopted among the 10 folds used in the experiment. We can observe that the Fusion operation of *Subctration* shows higher values in accuracy, precision, recall, f1 and ROC-AUC than BERT and SE-ResNet50.

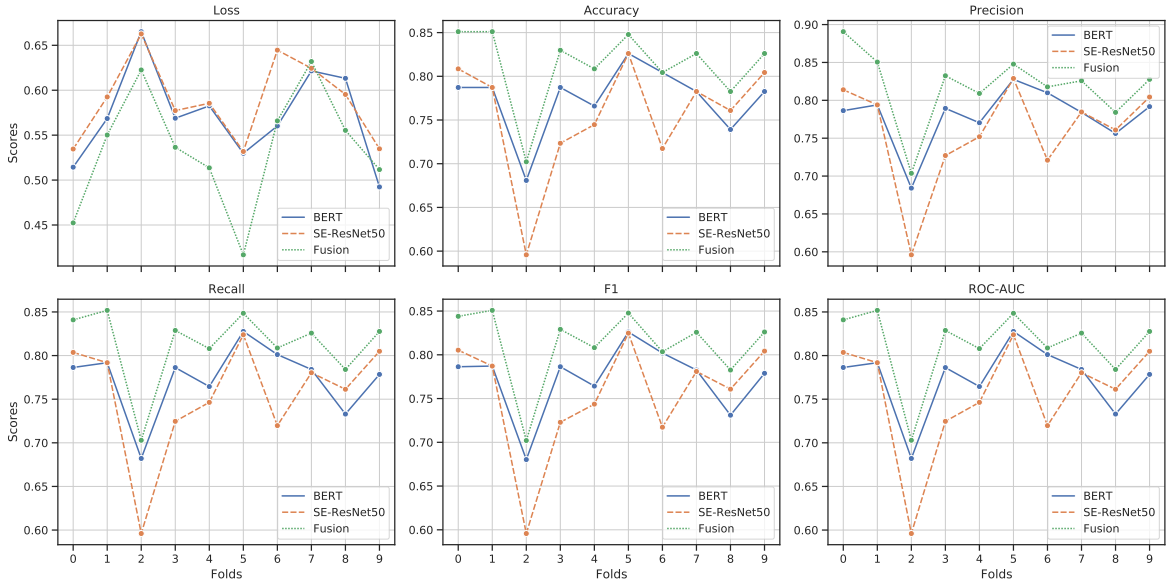

Fig. 3. Loss, Accuracy, Precision, Recall, F1 and ROC-AUC for validation phases, with the number of 100 epochs for the BERT, SE-ResNet50 and Fusion = *subtraction* models.

Intuitively, Late Fusion power comes from the junction of the embeddings, compensating for the weaknesses of unimodal models. On one hand, by exploiting the multimodals models, arithmetic operations boosts the accuracy of the classifiers learned by fusion. However, some operations are not as effective when linking embeddings of various modalities(ie, Division, Multiplication and Power).

The results indicate that interleaving various modalities leads to a robust performance over the entire spectrum of problems.

Gain Comparison - Considering a non-skewed dataset and due to its ease of understanding, we compared the difference of the prevision results by means of accuracy. Table II shows the difference between unimodal vs. multimodal results.

Table II. Cross-validation Results Accuracy

Model	Accuracy (%)	STD	BERT	SE-ResNet50
BERTimbau Base	**77.43**	0.040	-	-
SE-ResNet50	**75.51**	0.067	-	-
Sub	**81.30**	**0.067**	**3.86**	**5.79**
Sum	**80.86**	**0.067**	**3.42**	**5.35**
Concat	**79.37**	**0.078**	**1.94**	**3.86**
Max	**79.35**	**0.041**	**1.91**	**3.84**
Min	**79.14**	**0.070**	**1.70**	**3.63**
Pow	70.16	0.095	-7.27	-5.35
Multi	53.14	0.124	-24.29	-22.37
Div	49.29	0.050	-28.15	-26.22

The proposal reached 81.3% of accuracy using the combination of subtraction with a gain of 3.86% over the textual model and 5.79% for the vision model. Our hypothesis for this result is that the use of late fusion through arithmetic operations is often favored because errors from multiple classifiers tend to be uncorrelated, and the method is featured independent [Ramachandram and Taylor 2017]. Another interesting result is when we compare max and sum operations. Max operation resembles max-pooling operation, and the sum operation occurs in many different deep learning architectures (skip connections and positional encoding of BERT). The results indicate that summing the embeddings can improve the results when compared with max-operation. We perform a significance test using T-test [Kim 2015] with a value of $p = 0.05$ comparing the sub with unimodal approaches (BERTimbau and SE-ResNet50). Sub is significantly different with SE-ResNet50 but not with BERTimbau. Therefore sub fusion operation improved the visual modality significantly.

5. CONCLUSION

In this study, we evaluated several multimodal fusion methods for predicting Youtube views using binary classes. We started by evaluating simple unimodal models using single modalities of data (text or images), followed by experiments with different fusion methods. The results show that some multimodal fusion operations (subtraction, sum, concatenation, maximum, and minimum) perform better than unimodal deep learning. Fusion approaches show improvements between 1.70% and 5.79% in accuracy. Based on the results, it seems feasible to use multimodal deep learning to improve unimodal deep learning models on the proposed task.

The main limitations are related to the availability of the YouTube data. As we depend on the availability of data, a more in-depth study would require more data and, consequently, partnerships with more than one content creator. Another problem is that, since each video corresponds to an example in the dataset, to have access to larger datasets we would need YouTubers with hundreds of published videos.

This discussion shows us that, as future work, it is possible to extract information from different domain sources and process and analyze data using multimodal Deep Learning. Finally, this study can be used as a template for other content creators (from Youtube or other platforms). Provides a methodology for estimating the quality of video titles and thumbnails. Even though our multimodal classifiers perform better than unimodal classifiers, there is still much room for improvement in future work. In the future, we intend to use fusion techniques with self-attention resources along with state-of-the-art models.

REFERENCES

Carta, S., Giuliani, A., Piano, L., Podda, A. S., and Recupero, D. R. Vstar: Visual semantic thumbnails and tags revitalization. *Expert Systems with Applications* vol. 193, pp. 116375, 2022.

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

Gadzicki, K., Khamsehashari, R., and Zetzsche, C. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, pp. 1–6, 2020.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 770–778, 2016.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7132–7141, 2018.

Islam, M., Ria, N. J., Masum, A. K. M., and Ani, J. F. Performance comparison of multiple supervised learning algorithms for youtube exaggerated bangla titles classification. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, pp. 1–6, 2021.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L., and Koishida, K. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13289–13299, 2020.

Kalra, G. S., Kathuria, R. S., and Kumar, A. Youtube video classification based on title and description text. In *2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. IEEE, pp. 74–79, 2019.

Kim, T. K. T test as a parametric statistic. *Korean journal of anesthesiology* 68 (6): 540–546, 2015.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature* 521 (7553): 436–444, 2015.

Liu, K., Li, Y., Xu, N., and Natarajan, P. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Poria, S., Cambria, E., and Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. pp. 2539–2544, 2015.

Purushwalkam, S. and Gupta, A. Pose from action: Unsupervised learning of pose features based on motion. *arXiv preprint arXiv:1609.05420*, 2016.

Ramachandram, D. and Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* 34 (6): 96–108, 2017.

Smith, L. N. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 464–472, 2017.

Smith, L. N. and Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*. Vol. 11006. SPIE, pp. 369–386, 2019.

Song, Y., Redi, M., Vallmitjana, J., and Jaimes, A. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. pp. 659–668, 2016.

Souza, F., Nogueira, R., and Lotufo, R. Bertimbau: pretrained bert models for brazilian portuguese. In *Brazilian conference on intelligent systems*. Springer, pp. 403–417, 2020.

Teng, E., Falcão, J. D., Huang, R., and Iannucci, B. Clickbait: click-based accelerated incremental training of convolutional neural networks. In *2018 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, pp. 1–12, 2018.

Trong, V. H., Gwang-hyun, Y., Vu, D. T., and Jin-young, K. Late fusion of multimodal deep neural networks for weeds classification. *Computers and Electronics in Agriculture* vol. 175, pp. 105506, 2020.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* vol. 30, 2017.

Wightman, R. Pytorch image models. *GitHub repository*, 2019.

Witten, I. H. and Frank, E. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record* 31 (1): 76–77, 2002.

Wu, R., Yan, S., Shan, Y., Dang, Q., and Sun, G. Deep image: Scaling up image recognition. *arXiv preprint arXiv:1501.02876* 7 (8), 2015.

Yu, Z. and Shi, N. A multi-modal deep learning model for video thumbnail selection. *arXiv preprint arXiv:2101.00073*, 2020.

Zhou, R., Khemmarat, S., and Gao, L. The impact of youtube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. pp. 404–410, 2010.