UNIVERSITY OF GRONINGEN

BACHELOR PROJECT COMPUTING SCIENCE

Instagram Video Classification with convLSTM Neural Network

Authors:
Hwajun LEE *(s3543609)*

Supervisor:
prof. dr. E. TALAVERA MARTÍNEZ
dr. GEORGE AZZOPARDI

July 13, 2021

Instagram Video Classification with convLSTMs Neural Networks

Hwajun Lee

Abstract—Instagram is one of the social media platforms. The number of users continues to increase, and many posts such as videos are being uploaded. The growing amount of video can be a good resource for classifying videos. Video classification is currently a challenging field. The video can be viewed as a continuous image. Continuous images occurring over time can be classified using ConvLSTM neural network. Therefore, we propose to classify Instagram videos using the ConvLSTM neural network. We expect ConvLSTMs neural network to implement video classification effectively compared to other algorithms.

Index Terms—Video Classification, convolutional Long Short Term Memory, convLSTM, Instagram

1 Introduction

Instagram is an image and video-sharing social media platform. At 1 billion active users, Instagram is ranked as the 6th most popular social network in the world[1]. Statistics show the number of Instagram users is still increasing, this indicates just how much video data is being uploaded to Instagram. Video content on Instagram also attracts a higher rate of engagement than still images do, with a higher rate of user engagement on metrics such as likes and comments[2]. The average length of a video on Instagram lies somewhere between 30 to 60 seconds, making them very suitable for ML training data compared to long-form content.

In the field of computer vision, deep learning has shown promising results when applied to different applications such as image classification, speech recognition, and text sentiment classification. Nowadays, information on the Internet is often communicated via multimedia data including texts, audio, images, and videos. Among them, videos contain the highest amount of information parameters, this makes video a challenging field for applied machine learning. In 2015, a model called ResNet began to surpass humans in the task of classifying images with approximately 96% accuracy. Since then, the application of deep learning in image classification has continued to improve. However, the video field is still challenging. While the difficulties of video classification are challenging, the study of video classification can be helpful in various fields and act as the basis for future technological advances for society.

The video data is divided by the modality of data, different types of data. As a result, we divide video into 4 data modalities: Photo-sequences, Audio, Text, and detected concepts by the YOLO network. This research focuses on continuous photography of videos. A brief order in which our research is conducted is as follows. Our research has collected videos from Instagram and uses them for video classification using ConvLSTMs neural networks. The research aims to solve the main question: how to effectively classify the Instagram video.

A video is a collection of sequential images in order. But the approach between the classification of image and video is different since the video is multiple frames with time. There are various methods to classify videos. Among various methods, this research wants to find an effective way to classify Instagram videos. Therefore, video classification is implemented by using ConvLSTM neural network architecture. Although LSTM input data is one-dimensional, ConvLSTM is designed for three-dimensional input data. This can raise the following research question:

How can we effectively classify Instagram Videos with ConvLSTMs neural network?

Advertisers and companies are paying attention to Instagram's potential to influence consumer purchase decisions. Video classification may present new marketing methods for advertisers or companies that may need to identify Instagram trends. To go further, video classification can help advertisers in finding by categorizing videos and influencers that fit their customer's demands. Instagram can use video classification to moderate its content. If a specific environment is not allowed for upload, then Instagram can use automation to flag a video for review. It is expected that our research into the automated classification of Instagram videos could help with any of the aforementioned tasks.

The paper begins by investigating the current related work in Section 4. In Section 5, the main methods and related tools are mentioned. In Section 6, the experiment will be shown with a discussion about the results in Section 7. In Section 8, the research is concluded, and Section 9 will talk about the future work with this research.

2 Related Work

Video can be understood as a series of individual images. The most typical example is a video on social media such as YouTube, Facebook, or Instagram. Especially, the amount of social media video content has been increased in recent years[3]. Therefore, there were many practices and challenges to study video classification as performing image classification a total of N times. Furthermore, video classification has been an active topic of research in computer vision for applications in video behavior recognition[4], video indexing/retrieval[5], prediction of sea surface temperature [6], prediction of nowcasting[7] and autonomous driving[8].

2.1 Convolutional Neural Networks

CNN (Convolutional Neural Networks)[9] was first introduced in Lecun et al.,(1989) to process images more effectively by applying filtering techniques to artificial neural networks, and later in Lecun et al.,(1998) a form of CNN currently being used in deep learning was proposed[10]. The image passes through convolutional layers. Then, filters extract important features from the images. After passing convolutional layers in order, the output is connected to a fully connected network[11].

Karpathy et al.,(2014) suggested a good foundation to integrate the temporal component of videos into CNN (Convolutional Neural Networks) models. CNN is particularly useful for finding patterns to recognize images. CNN learns directly from data and uses patterns to classify images. Karpathy et al. (2014) showed that CNN is effective in speeding up computation and transfer learning with video classification [12].

2.2 Recurrent Neural Network and Long Short Term Memory

RNN (Recurrent Neural Network)[13] is a type of artificial neural network in which a hidden node is connected to a directed edge from a circular structure. It is known as a suitable model for processing data that appears sequentially. However, RNNs are known to significantly degrade their ability to learn if there is a distance between the relevant information and the point at which they are used. This is called the "vanishing gradient" problem. To solve the "vanishing gradient" problem of RNN, Hochreiter, and Schmidhuber proposed LSTM (Long Short Term Memory) in 1997 [14][15]. LSTM is one of RNN. The core idea of the LSTM model is a cell state to maintain the state overtime [14][15]. The cell state allows LSTM to add or remove information regulated by structures called gates [8].

Data collected over successive periods are characterized as a Time Series. Therefore, it has been shown that most of the research about handling video with LSTM. The recent paper of Liu et al.(2019) proposed a video image target monitoring algorithm based on RNN-LSMT deep learning [14]. In addition, Zhang et al.(2020) handle the video information with LSTM for analysis of temporal clues. The method proposed by Zhang et al. (2020) is using the LSTM layer to model the long-term dynamic information in video sequences to achieve deep learning on both spatial and temporal clues [8]. Joe et al.(2015) also use LSTM for video classification to reduce time[16].

2.3 ConvLSTMs Neural Network

To implement effective video classification, we suggest ConvLSTMs neural network architecture to classify video for photo sequence cases. ConvLSTM[6] is a variant of LSTM with a convolution operation inside the LSTM cell. The main difference between ConvLSTM and LSTM is the input dimensions. It is not suitable for spatial sequence data such as video since the input data of LSTM is one-dimensional. However, ConvLSTM neural network is suitable for the three-dimensional data type input [11].

There is interesting research using ConvLSTM neural network to precipitation nowcasting [7]. The part that was composed of a single LSTM was replaced with a stacked one of several convolutional LSTMs. Therefore, both temporal and spatial information can be considered, and input, output, and state can all be handled efficiently with 3D tensors. The forecasting model does not use the output of the last Convolutional LSTM cell of the Stacked Convolutional LSTM as the final result. Instead, it receives the state of the Convolutional LSTM at each different level one by one and applies a 1×1 convolution layer to the concatenation result. This seems to be due to the effect of adjusting the output to the same dimension as the input, while simultaneously considering the cell state at different levels. Using convLSTM has the advantage that the input/output and layers state are computed as three-dimensional vectors. So to summarize, this paper concluded that using a new method called convolutional Long Short Term Memory is most optimal for this use case. Ge et al.(2019) also use ConvLSTM to improve the accuracy of recognition by extracting the salient regions of action in video effectively [17]. Parsia

et al.(2020) use ConvLSTM to predict anomaly detection since ConvLSTM neural networks can predict the subsequent video sequence from a given input [18].

2.4 Instagram Video Evaluation

Interestingly, some works are using Instagram as a dataset. Jeong et al.(2017) conducted Instagram image classification with deep learning. This research used AlexNet and ResNet which are CNN. Then it compares the result to check the efficient method of image classification [19]. Another research is automatically detecting image-text mismatching on Instagram with deep learning [20]. It seems that the most of the research with Instagram handles not the video but image data. Therefore, we expect that Instagram video classification will be meaningful and interesting in challenging new fields.

3 Our Proposed Methodology

In this section, we describe how given a set of videos we train a model for their classification based on visual information.

1) Feature extraction
2) ConvLSTM for video classification
   a) What is a convLSTM
   b) How do we implement them
      (Network details and implementation details)

Feature Extraction: Video is a sequence of images with order. This experiment will focus continuous photography, one of the characteristics of the video. Therefore, before starting video classification, the task of extracting images from the video must be done first. It can be extracted images in frames from images to obtain a large amount of image data. A frame is a picture that is sprayed on the screen when a video, movie, or TV delivers a video medium. Each of these pictures changes rapidly at some rate per second, creating a moving video.

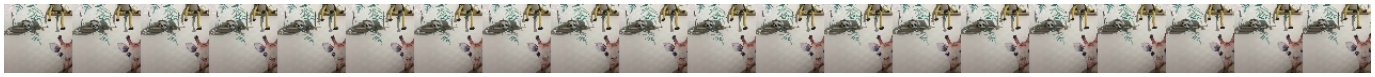
The video dataset is the most important feature of the project. If collecting precise video datasets for each class is successful, the result of the training will be satisfied. The project considers an ideal number of between 300 and 400 videos for each class. In order to use all the frames of the video evenly, 20 frames at regular intervals were extracted from all frames. This paper will call the same intervals frames as equidistance frames. For example, if the total frames of video are 100 and we want to get equidistance 20 frames, every 5 frames steps skip to get 20 frames. In other words, if we have a total of 100 frames of video and we collect 20 frames, we will only use frames 5, 10, 15, 20, ...,95, 100 frames. Although both are extracted 20 frames per video, Figure 1 and Figure 2 show different sequence of frames. Unlike Figure 1, Figure 2 shows 20 frames per video with even distance. It can be seen that each frame also has more differences. Since it will have the effect of learning every frame of the video evenly, the using equidistance 20 frames can train effectively compared to using the first extracted 20 frames. The two results from both extracted first 20 frames and equidistance extracted 20 frames are compared in Section 6.4 and 6.5.1.

ConvLSTM for video classification: The research will focus on architectures that use photo sequences of video and aim to classify videos of Instagram effectively. For this aim, the main architecture will be Convolution Long Short Term Memory (ConvLSTM) neural network. It is expected to see how effectively ConvLSTMs can classify videos through this research. Basic convLSTMs sources are provided as open-source by Keras[21], Tensorflow[22] in Python. The ConvLSTM is a model that recognizes spatial characteristics. This model has the advantage that the input/output and layers state are computed as three-dimensional vectors. This model expects better performance than the existing LSTM and Fully Connected LSTM models [7]. FC-LSTM[6] is a multivariate version of LSTM in which the input, output, and state of a cell are one-dimensional vectors. However, convLSTM takes a completely different approach. It's putting convolution in the LSTM internal operation itself. There are two major differences. First, in convLSTM, the input-memory-output gates and cell input, output, and cell state are all three-dimensional tensors. All elements are different from FC-LSTM, which is a one-dimensional vector. Second, all matrix product was replaced by convolutional operations. This means that the number of weights in each cell can be dramatically lower than FC-LSTM. This is equivalent to the effect when the Fully-Connected Layer is replaced by the convolution layer, which can significantly reduce the number of weights in the model as a whole. Consequently, convLSTM can capture both spatial and temporal meaning simultaneously in the LSTM cell itself.

To prevent overfitting, the project uses cross-validation. A model consisting of training/validation data is highly likely to overfit only the corresponding training data. However, it is possible to create a more generalized model by learning the entire range of data through a cross-validation method.

We divide the data into 3 equal parts. Then, we get $\frac{1}{3}$ as Validation data and $\frac{2}{3}$ as training data and change $\frac{1}{3}$ of the verification data and evaluate the performance. So, there will be a total of 3 performance results. From the three results, the mean accuracy and standard deviation can be calculated.

The result will be evaluated by confusion matrix or classification report, including precision, recall, f1-score, macro average, and weighted average. The experiment will use `'classfication_report'` and

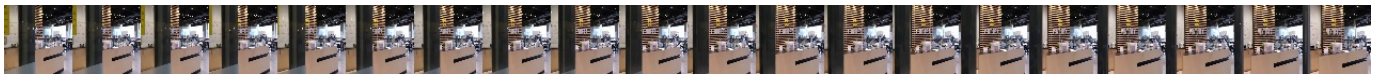


(a) Child's Room.

(b) Restaurant.

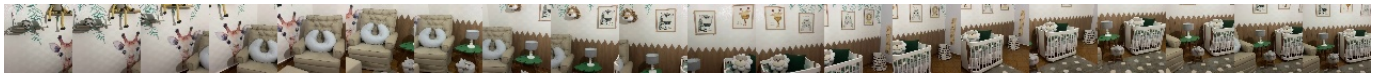
(c) Museum.

(d) Dining Room.



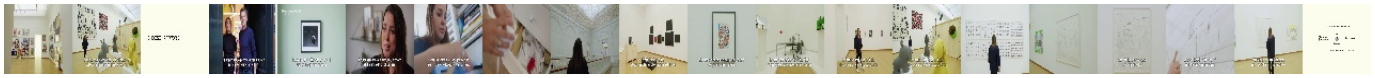
(e) Office.

Fig. 1: Non equidistance 20 extracted frames



(a) Child's Room.

(b) Restaurant.



(c) Museum.

(d) Dining Room.

(e) Office.

Fig. 2: Equidistance 20 extracted frames

'confusion_matrix' function which is provided by Scikit-learn. The metrics package in the Scikit-learn provides 'classfication_report' to obtain precision, recall, macro, and weighted average. This command determines the precision, recall, and f1-score of each class viewed of a positive class, respectively, and evaluates the performance of the entire model with its average value. If the prediction results of the classification model are correct and if positive or negative is predicted, it is called True. It is called False if the prediction results are incorrect and if positive or negative is predicted.

	Predicted True	Predicted False
Actual True	True Positive (TP)	False Negative (FN)
Actual False	False Positive (FP)	True Negative (TN)

TABLE 1: Binary Confusion Matrix.

Precision is the ratio of what the model classifies as True to be true.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

The recall is the ratio of what the model predicts to be true among the actual True.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

Macro average is a simple average. As an extension of the Macro Average, Weighted Average calculates the average by weighting the number of data corresponding to each class.

This research also uses the confusion matrix to evaluate the result. The Confusion Matrix is a table of the results counting whether the actual class matches the predicted class. The actual class is represented by a row, and the predicted class is represented by a column. In the last steps, the result of the experiments will be analyzed and discussed.

4 Experiments

In this section, the experiment will be discussed with the Instagram video dataset, set-up, and implementation details.

4.1 Instagram Video Dataset

Thankfully, the Python code to collect and scrap the video data was provided by Keiko Angela Nicolasky of the University of Groningen. This code helps us collect videos of the desired category using hashtags. The videos collected in this way are divided into a total of 5 classes. The 5 classes are following: Office, Child's Room, Dining Room, Restaurant, and Museum.

Videos are collected by using hashtags. For example, if we want to collect office videos, the hashtag will be #officedeco or #offcelife. Our dataset consisted of 199 videos of Child's Room, 215 for Dining room, 232 for Museum, 178 for Office, and 183 for Restaurant. Table 2 shows the number of videos in each class. The office has the smallest number of data with 178 videos and the largest number of data with 232 museums. The five data classes are slightly imbalanced. Imbalanced data is data that does not have a uniform proportion of data occupied by each class in the data and is biased. To improve the performance of classification algorithms, all classes will be under-sampled to maintain a balance between class data. In other words, it controls the number of video data in different classes to match the number of data in the office that has the smallest number of video data. After checking the distribution of the data, we will go through the process of matching the class with the higher distribution to the class size of the lower distribution.

Class	Number of Videos
Office	**178**
Child's Room	199
Museum	232
Dining Room	215
Restaurant	182

TABLE 2: Number of video dataset per class.The bold number represents the lowest number of video dataset.

Class	Number of Videos
Office	178
Child's Room	178
Museum	178
Dining Room	178
Restaurant	178

TABLE 3: Actual Number of used video dataset per class.

4.2 Experiment Set-up

To train the video classification model, the general steps are followed:

1) Constants Definition
2) Extract Frames from the Video
3) Create Data
4) Define the architecture of model
5) Train the model and evaluate

The program needs a constant definition for the height, weight of images, and the number of frames which we want to extract from video. The experiment defines image height and weight with 64. This is the dimension of each frame of videos. After defining the constants, the video will be extracted as frames. The number of frames can be changed. Higher sequences of images will give better results. However, lots of sequences of frames will be expensive. In the main experiment, 20 frames from video are used. From the extracted frames, sequence of images, the data will be created.

The most important thing in machine learning is to get the highest accuracy. To improve accuracy, complex layer architectures are built such as dense layer and dropout layers. Adding more layers causes overfitting. It means over-fitting learning data by machine learning. In other words, the training model learns up to the noise that exists in the training set, and the

accuracy is low in the test set. It may be enough to explain the present well, but the necessary information is not the data we already know, but the new data we will know. But if none of the new data comes together, the system could just be useless. The more complex the model and the less data there is, the more likely it is to overcharge. With a lot of data, overcharging is less likely to occur even with complex models. However, the number of videos on Instagram is limited for each class, so this is the only way to collect all the videos possible.

To solve or prevent over-fitting, four methods can be suggested: Simplify Layer, Add Normalization Layer, Raise Dropout ratio, Implement Cross Validation. Based on four methods, the model architecture layer is created by adding a normalization layer, using one convLSTM layer, and raising dropout ratio to 0.5. The Figure 3 shows the final model layer architecture. Flatten layer is the layer changing 2-dimensional image to 1-dimensional image. The dropout layer is the layer to prevent over-fitting and a method of randomly removing partial neurons.

Since all data sets are used for evaluation, it is also a method to be applied when data sets are insufficient. Because the data that has been removed separately from the validation set is also recycled for learning, if the entire data is too small to divide by learning/validation at once, we can divide the data several times as above and compare the model performance for each cross-validation. In addition, since one result is derived by integrating K performance results, a more general model performance evaluation is possible. A model consisting of one training/validation data is highly likely to over-fit only the corresponding training data. However, a more generalized model can be created by learning the entire range of data through the cross-validation method that is divided several times and evaluating the performance with the validation data. This experiment divides the data into 3 equal parts with a cross-validation method.

When the model is trained, it uses an optimizer. Optimizer is a word that is used a lot in mathematics. It is sometimes referred to as a ”repair plan” or ”repair plan” problem or ”optimization problem” in physics or computers, as a representation of the energy of a system modeled on a function in mind. In deep learning, optimization can be said to make the learning speed fast and stable[23]. Adam is currently the most widely used algorithm for learning deep neural networks and generally demonstrated the best learning performance. In this work, we choose the Adam optimizer that performs best in prior study and in this work when setting up the optimizer to stabilize learning [23].

After defining the model architecture, the model will be trained. After training, the result is evaluated

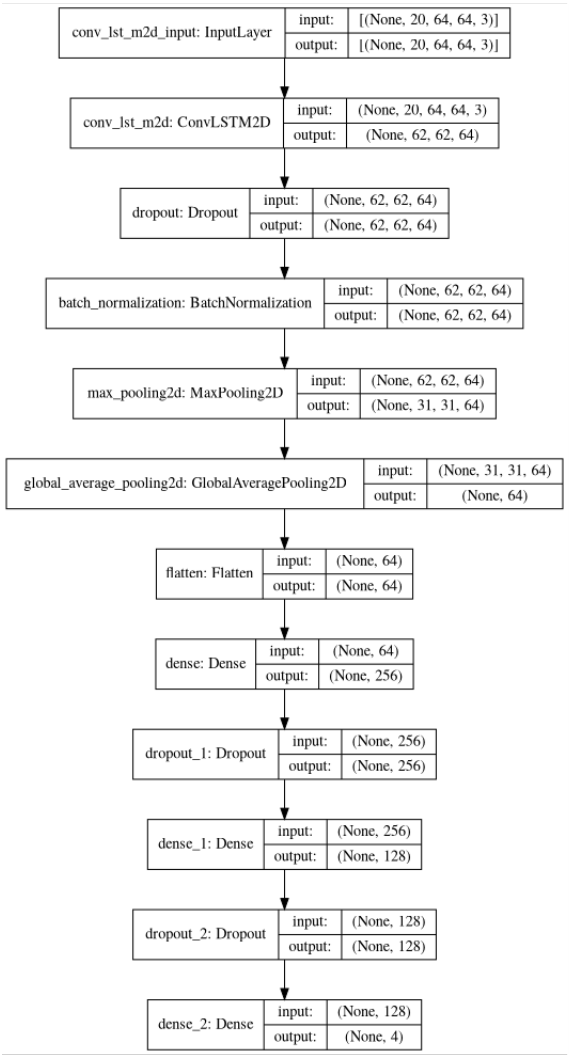


Fig. 3: Model Layer Architecture.

with a confusion matrix and a classification report that included precision, and recall.

4.3 Main experiments

This experiment is conducted with equidistance extracted 20 frames. The entire dataset is randomly split into 3 folds. Then, it will fit the model. After repeating this process 3 times, it gives 3 performance results of accuracy as Table 4. The average accuracy of these three is 0.409 which is about 41% and the standard deviation is 0.0174.

`classification_report` function from Skicit-learn helps to evaluate the result and obtain precision, recall, macro and weighted average. Macro and

	Accuracy
1	0.404
2	0.3906
3	0.4324
Average Accuracy	0.409
Standard Deviation	0.0174

TABLE 4: Results of Performance.

Weighted Average show the same value because we use an equal 178 video data for all classes. Table 5 shows that the recall of the restaurant is the highest value compared to other classes. 0.55 of restaurants recall means that 55% of actual restaurant videos are predicted as restaurants. Recall value of Dining room shows also 0.52 which is higher than 0.5. It means 52% of Dining Rooms videos are correctly predicted as Dining Room.

	precision	recall	f1-score
Dining Room	0.36	0.52	0.42
Office	0.37	0.40	0.38
Restaurant	0.49	**0.55**	0.52
Museum	0.43	0.29	0.35
Child's Room	0.49	0.34	0.40
Accuracy			0.42
Macro avg	0.43	0.42	0.42
weighted avg	0.43	0.42	0.42

TABLE 5: Classification Report. The bold number of Restaurant represents the highest value of actual prediction.

This is the confusion Matrix of the experiment. The Confusion Matrix is a table of the results of counting whether the original class of the target matches the class predicted by the model. The actual class is represented by a row, and the predicted class is represented by a column. For example, 52% of Dining Room videos are predicted correctly as Dining Room and 55% of restaurants videos are predicted correctly as Restaurant. This visualized confusion matrix shows a good-looking result in which the diagonal lines from the top left to the bottom right are expressed in dark colors compared to others. This matrix shows a pretty diagonal line from left top to right bottom. According to this table, restaurants and dining rooms are the most well predicted.

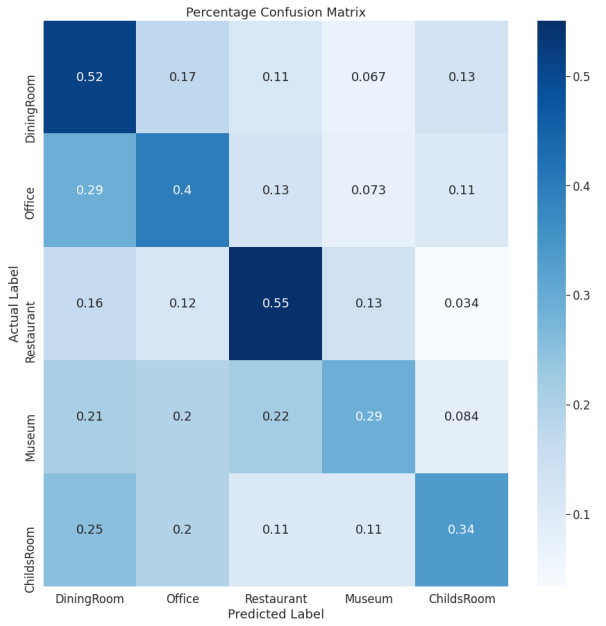


Fig. 4: Confusion Matrix of result from equistance 20 extracted frames.

4.4 Auxiliary Experiment

As auxiliary experiments, we implement classification with various condition such as extracted first 20 frames, different number of videos dataset, different number of frames and different number of layer. Auxiliary experiments will be useful for comparing and understanding the results of the main experiment.

4.4.1 Non equidistance 20 frames

This experiment is run to observe whether video classification results using frames extracted from video would show better results than video classification results using frames extracted from scratch.

Class	Number of Videos
Office	**178**
Child's Room	199
Museum	232
Dining Room	215
Restaurant	182

TABLE 6: Number of video dataset per class.

The mean accuracy of 3 result is 0.2966 which is about 30% and the standard deviation is 0.0401. Non-euqidistance extracted frames shows 11% lower accuracy than equidistance extracted frames experiment result.

Precision of most classes shows lower than 6.4 main experiment. In addition, except Restaurant and Museum, recall of other classes get lower result compared to 6.4 classification report.

Class	Number of Videos
Office	178
Child's Room	178
Museum	178
Dining Room	178
Restaurant	178

TABLE 7: Number of used video dataset per class.

	Accuracy
1	0.3266
2	0.3232
3	0.2399
Average Accuracy	0.2966
Standard Deviation	0.0401

TABLE 8: Results of performance with non equidistance frames.

According to Figure 5, despite 29% of Dining Room videos are predicted as Dining Room, 31% of Dining Room videos are predicted as Museum. Only 15% of Office videos are predicted as Office and 44% of Child's Room videos are predicted as Museum. Entirely, the most of the classes are predicted as Museum except Restaurant. This is not a successful result of confusion matrix. Comparing the accuracy,

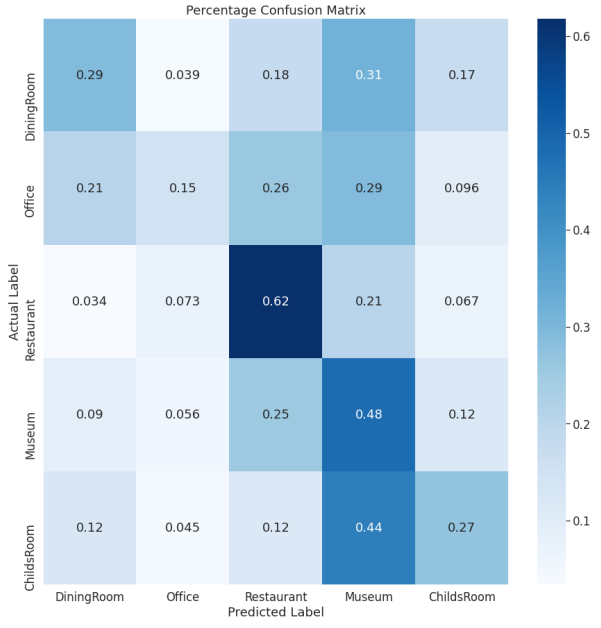


Fig. 5: Confusion Matrix of result from non-equistance 20 extracted frames.

precision, recall, and confusion matrix of the two experiments (Section 6.4 and Section 6.5.1), experiments

	precision	recall	f1-score
Dining Room	0.39	0.49	0.44
Office	0.41	0.15	0.21
Restaurant	0.43	**0.62**	0.51
Museum	0.28	**0.48**	0.35
Child's Room	0.37	0.27	0.31
Accuracy			0.36
Macro avg	0.38	0.36	0.34
Weighted avg	0.38	0.36	0.34

TABLE 9: Classification Report with non-equidistance frames.

with frames extracted at equal intervals in most parts show better results. Using frames extracted at equal intervals allows learning of the entire video evenly, while learning to use non-equidistance frames does not learn everything the video shows as a whole and can be learned unintentionally.

4.4.2 *Complex model architecture*

Now, one more convLSTM is added to the previous model. This experiment is implemented with same number, 178, of video per classes to compare the result. We expect higher accuracy than the previous result from more complex model architecture since this experiment have more complex architecture layer.

Class	Number of Videos
Office	**178**
Child's Room	199
Museum	232
Dining Room	215
Restaurant	182

TABLE 10: Number of video dataset per class.

Class	Number of Videos
Office	178
Child's Room	178
Museum	178
Dining Room	178
Restaurant	178

TABLE 11: Actual Number of used video dataset per class.

The average of the accuracy of these three is 0.345 which is about 35%. However, this is 6% lower than 178 video per class classification. The standard deviation is 0.0354 as Table12.

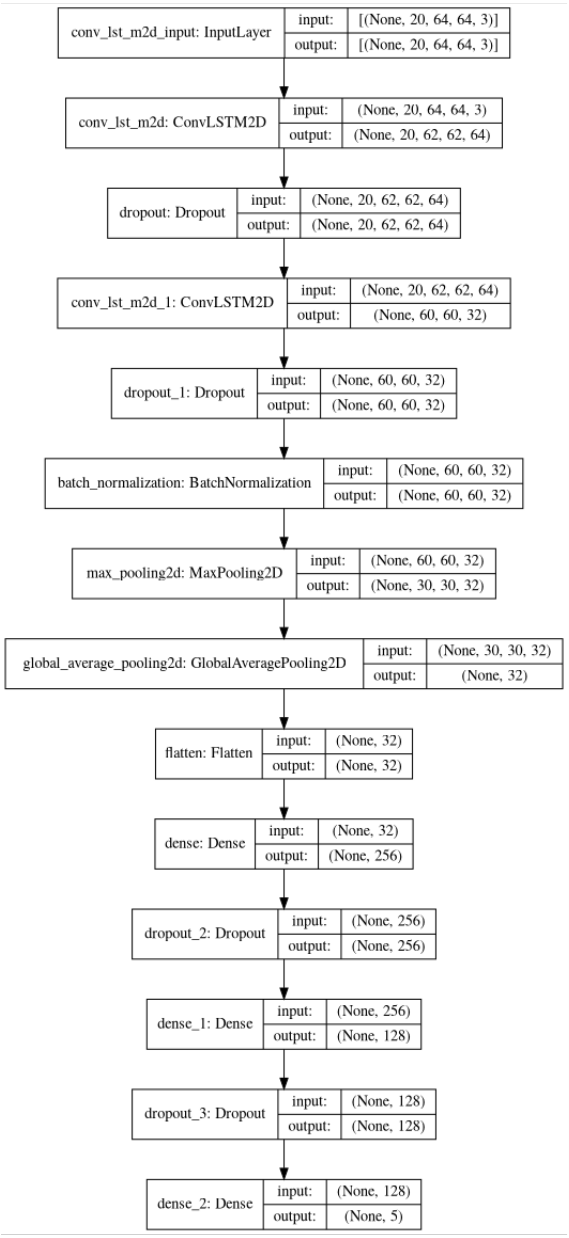


Fig. 6: 2 convLSTM Model Layers Architecture.

	Accuracy
1	0.3064
2	0.3367
3	0.3919
Average Accuracy	0.345
Standard Deviation	0.0354

TABLE 12: Results of performance from 2 convLSTM layers.

From the following classification report results, precision of Restaurant, recall of Museum, and recall of Children's room values are higher than 6.4 experiment.

	precision	recall	f1-score
Dining Room	0.29	0.30	0.29
Office	0.34	0.21	0.26
Restaurant	**0.63**	0.37	0.46
Museum	0.34	**0.45**	0.39
Child's Room	0.39	**0.31**	0.46
Accuracy			0.38
Macro avg	0.40	0.38	0.37
Weighted avg	0.40	0.38	0.37

TABLE 13: Classification Report of 2 convLSTM layers.

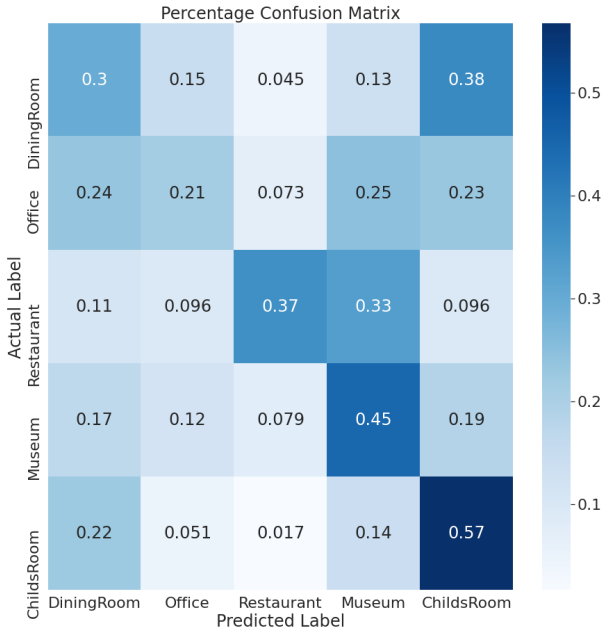


Fig. 7: Confusion Matrix of result with 2 convLSTM layers.

The Figure 7 confusion matrix with 2 convLSTM layers shows a less satisfactory result compared to

using 1 convLSTM layer. 30% of the Dining Room are predicted to be the Dining Room. However, 38% of the dining rooms are predicted as child's room class which is more than predicted as dining room. Moreover, Office videos are predicted as Dining room, Office, Museum, and Child's room with similar possibility. The Figure 7 evaluates that experiments with 2 convLSTMs layer did not perform video classification using data from each class properly.

We assume that a higher number of weights to train with a low number of samples is the reason. When weights are large, accuracy is more sensitive to small noises in the input data. So, when a small amount of noise is propagated through a network with large weights, it produces a much different value than a network with small weights.

5 Results and Discussion

In this paper, the environmental categorization of Instagram video was studied using the ConvLSTM deep learning model. We presented three video-classification methods: Equidistance extracted frames methods, Non-equidistance extracted frames methods and using 2 convLSTM layers. According to Table14, extracted frames at intervals (equidistance frames) with the ConvLSTM model were used to show 41% accuracy. This is better than using non-equidistance frames and using two convLSTM layers.

Experiment	Average Accuracy	Standard Deviation
Equidistance	**0.409**	0.0174
Non-equidistance	0.2966	0.0401
2 convLSTM Layers	0.345	0.0354

TABLE 14: Results from each experiment. The bold number represents the highest accuracy.

However, some videos received from the Instagram API do not have accurate descriptions. For example, Figure 8 is a screen capture of a video of the office. Despite the video taking place in an office, a cat takes up more than half of the frame.

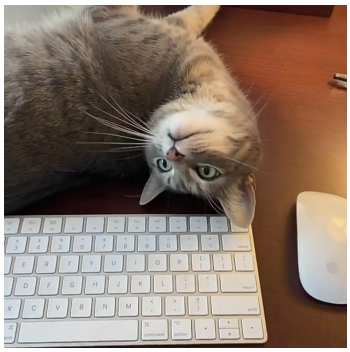
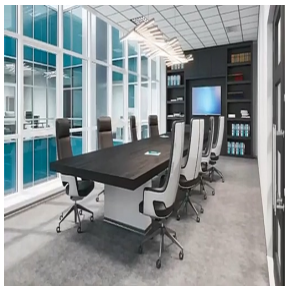
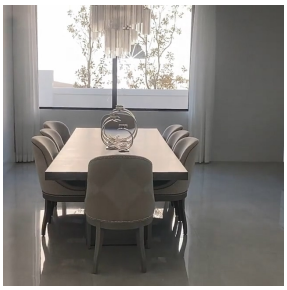


Fig. 8: Screen Capture of Office video.

Also, some videos have low inter-class variance. For example, like Figure 9, screen captures of a dining room and office, are very similar. Both screen captures have a big long desk and multiple chairs.



(a) Office.



(b) DiningRoom

Fig. 9: Office and Dining Room video Screen Captures.

Experiments were influenced by data from these uncertain and similar videos. The confusion matrix of Figure 4 clearly shows satisfactory results, but the accuracy is 41%, less than 50%. Better performance in video classification would have been achieved if video data represented each class well and had clear differences. Furthermore, the number of video clips from Instagram public accounts is very limited. Most of the videos on private accounts were not available because they were excluded from the API as a result of personal protection. The set of training data is relatively small, the result would be better if we collect more video data per class.

6 Conclusion

There are two key takeaways from the experiment results. First, we can see how frames should be extracted for video classification consisting of continuous images. Second, using 1 layer of convLSTM gives better results than using 2 layers of convLSTMs. The reason is that the data is small, but the number of weights has increased, which makes it sensitive to small noise. Also, the accuracy has not yet reached more than 50%. This comes as a result of some videos in our training data not being categorized properly. Nevertheless, the Figure 7 confusion matrix showed that most of the video data are properly predicted.

7 Future Work

This study may lead to better results by creating more complex layer structures in the future. However, more video data is needed to use layers of more complex structures. Also, this experiment used 20 frames that were extracted from one video. It is possible to experiment with extracting more than 20 frames such as 40, 60frames per video in the future. There is also a way to

increase the others such as the number of video classes. This experiment uses five classes to implement video classification. However, the experiment can also be run by extending it to more environmental video classes.

REFERENCES

[1] Jasmine Enberg. Global instagram users 2020, Dec 2020.

[2] The ultimate breakdown of social video metrics for every platform, Jul 2020.

[3] Mehtab Afzal, Nadir Shah, and Tufail Muhammad. Web video classification with visual and contextual semantics. *International Journal of Communication Systems*, 32(13), 2019.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features, 2005.

[5] Video indexing and retrieval. *Multimedia Technology for Applications*, 2009.

[6] Sihun. Jung, Young Jun. Kim, Sumin. Park, and Jungho. Im. Prediction of sea surface temperature and detection of ocean heat wave in the south sea of korea using time-series deep-learning approaches, 2020.

[7] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *CoRR*, abs/1506.04214, 2015.

[8] Lei Zhang and Xuezhi Xiang. Video event classification based on two-stage neural network. *Multimedia Tools and Applications*, 79(29-30):21471–21486, 2020.

[9] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, December 1989.

[10] Gradientbased learning applied to document recognition. *Intelligent Signal Processing*, 2009.

[11] Alexandre Xavier. An introduction to convlstm, Apr 2019.

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[13] Jitendra Kumar, Rimsha Goomer, and Ashutosh Kumar Singh. Long short term memory recurrent neural network (lstm-rnn) based workload forecasting model for cloud datacenters. *Procedia Computer Science*, 125:676–682, 2018. The 6th International Conference on Smart Computing and Communications.

[14] Feng Liu, Zhigang Chen, and Jie Wang. Video image target monitoring based on rnn-lstm. *Multimedia Tools and Applications*, (4):4527–4544, 2018.

[15] Sepp Hochreiter and Schmidhuber Jurgen. *Long short term memory*. Inst. fur Informatik, 1995.

[16] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4694–4702, 2015.

[17] Hongwei Ge, Zehang Yan, Wenhao Yu, and Liang Sun. An attention mechanism based convolutional lstm network for video action recognition. *Multimedia Tools and Applications*, 78(14):20533–20556, 2019.

[18] Smita Parsai and Sachin Mahajan. Anomaly detection using long short-term memory. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020.

[19] Nokwon Jeong and Soosun Cho. Instagram image classification with deep learning, 2017.

[20] Yui Ha, Kunwoo Park, Su Jung Kim, Jungseock Joo, and Meeyoung Cha. Automatically detecting image–text mismatch on instagram with deep learning. *Journal of Advertising*, 50:1–16, 01 2021.

[21] Keras Team. Simple. flexible. powerful.

[22]

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.