Columbia University

Project Report

COMS 6901

Analyzing Effects of Visual Data, Textual Data and Speaker on Student Learning

Author:
Mayank Saxena
ms5736@columbia.edu

Supervisor:
Dr. John Kender

1 Abstract

The aim of this project is to analyze student's understanding and learning of a concept based on where they are looking at. For this purpose, we divide the student's eye-gaze data into four different categories - visual data, textual data, speaker and other. These four categories correspond to where the student is looking at for a particular instance. For eg - the "visual data" category corresponds to the instances when the student is looking at the images and figures in the slide. Similarly, the "textual data" category corresponds to the instances of when the student is looking at the text in the slides. The "speaker" category corresponds to the instances of when the student is looking directly at the speaker while the speaker is delivering content via his voice. And lastly, the "others" category corresponds to the instances when the student is looking at none of the above mentioned three categories. We use these different classes as features and try to find which class has the highest positive and negative correlation with the student's post exam score. We also try to predict using these scores as features whether the student will get an answer right or wrong.

Contents

1 Abstract **1**

2 Introduction **3**

3 Database **3**

4 Related Works **4**

5 Analysis **4**

5.1 Data Cleaning and Preprocessing 4
5.2 Feature Engineering 6
5.3 Correlations 8
5.3.1 Complete Data 9
5.3.2 Infer Type Questions 10
5.3.3 Recall Type Questions 12
5.4 Prediction Models 13
5.4.1 Gradient Boosting Classifier 14
5.4.2 Support Vector Machine 14
5.4.3 2 Layer Neural Network 15
5.5 Results 16

6 Conclusion **16**

7 Future Work **17**

2 Introduction

This project is based on the hypothesis that it is easier to learn and understand new ideas and concepts through images and figures as compared to text. We also try to understand which aspect of the content being presented lead to better retention of concepts and which are positively correlated with student's learning and engagement. We also analyze the effect of looking at the speaker while he/she is presenting the content as compared to looking at the slides. We use all this information to derive conclusions of how lecture slides should be designed and what technique leads to the best understanding of concepts. We use the student's post exam scores as ground truth to find these relationships.

3 Database

The database consists of three different modules titled "Bicycle", "Tarmac" and "Visual Persepctive". These three modules correspond to the topic being taught in the videos. All the three modules have eye-gaze data files of the students which took part in the study. There are eye-gaze data files for 27 students for the "Bicycle" module, for 61 students for the "Road Paving" module and eye-gaze for 30 students for the "Visual Perspective" module.

Module	**Duration (m:s)**	**Eye-gaze data files**
"Bicycle"	4:59	27
"Road Paving"	4:39	61
"Visual Perspective"	4:56	30

The eye-gaze data files contain the following information:

1. CURRENT_FIX_X: Current x-coordinate of eye fixation position
2. CURRENT_FIX_Y: Current y-coordinate of eye fixation position
3. CURRENT_FIXATION: Duration for the current eye fixation

We use the above information to calculate the durations of when the student was looking at the visual data, textual data and at the speaker and thus derive the features required for our prediction model.

4 Related Works

Similar studies have been conducted in the past which show that studying through images and figures leads to better retention of concepts. By reading past literature, we found that all students learned better when picture and text was used alternately [1]. Similar experiments revealed that introducing imagery triggered active learning behaviors and that that student academic engagement was greater when apposite images were applied [2]. [3] tells us that words are abstract and rather difficult for the brain to retain, whereas visuals are concrete and, as such, more easily remembered. [4] suggests that relevant diagrams have the power to significantly enhance learning. [5], [6], [7] tell us that pictures are not only more effortless to recognize and process than words, but also easier to recall. The dual-coding nature of images allows for two independent ways of accessing visual memories, increasing the odds of remembering at least one of them thus it helps in improving students' learning of materials. [8] suggests that visuals help learners grasp concepts easily by stimulating imagination and affecting their cognitive capabilities. In a similar experiment conducted in [9], the results showed that students who had any form of visual aid always did better than students in the control group, students with diagrams, on average, outperformed students with only outlines. We also have studies suggesting that visualizations could be used by teachers as a formative assessment tool to improve student learning and creating a visual explanation is an excellent way to learn [10].

5 Analysis

5.1 Data Cleaning and Preprocessing

The first step which was involved in the data cleaning process was the removal of the data files in which the hardware wasn't calibrated properly. This resulted in wrong collection of data and thus these data files needed to be removed. Below is an example of a file in which there was a calibration drift. We can see that many of the eye fixations are out of the frame, a clear indication that there was a drift. One of the possible reasons for this drift could be that the participants moved their heads during video watching. This step resulted in the removal of 5 eye-gaze data files. The next step was to convert all the three videos to their corresponding frames and then finding

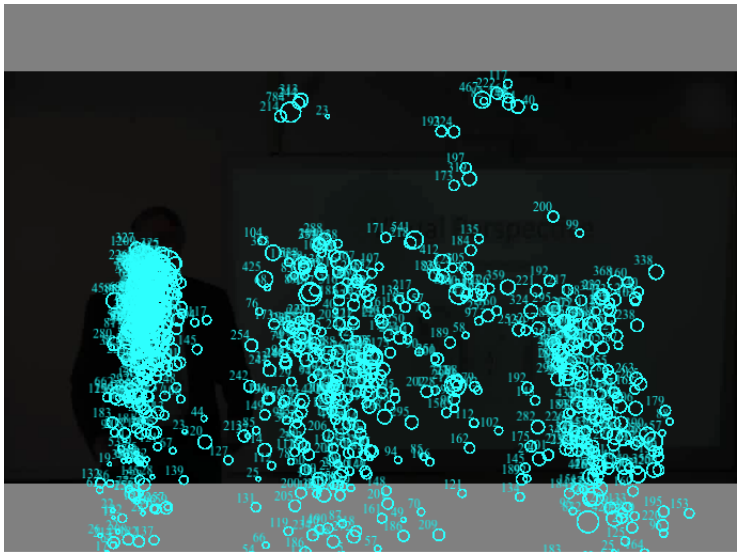
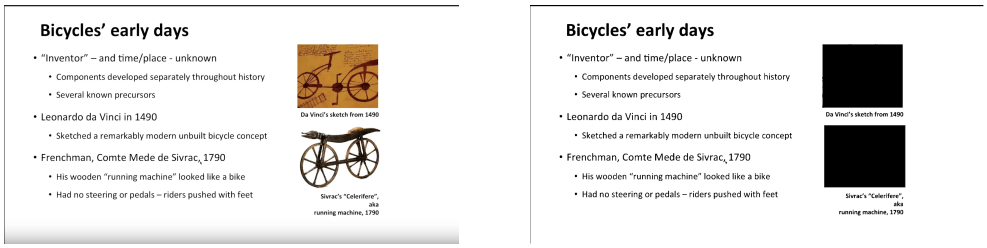


Figure 1: Calibration Drift while Data Collection

out the regions and bounding boxes in the frames for the visual data, textual data and the speaker. The conversion of the video to its corresponding frames was done using the OpenCV library. We also tried using OpenCV to detect the image and text area within the slides, but the results which were returned were inconsistent.



(a) Original Image (b) Visual Data Marked

Figure 2: Visual Data Marked using OpenCV

We then found out the regions for the text, image and speaker areas by manually drawing the bounding boxes around the regions across the different slides in the videos. Figure 3 demonstrates the same.
By finding the x,y coordinates of all the three bounding boxes, we were able

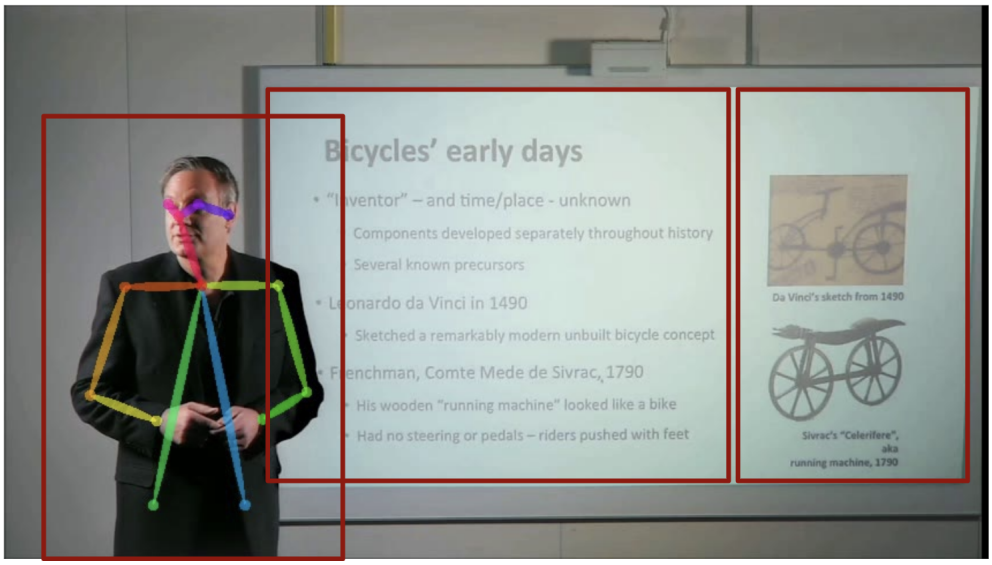


Figure 3: Bounding Boxes around Speaker, Visual Data and Text Data

to classify each eye fixation position as belonging to either of three bounding boxes or belonging to "other" class. We can see that the bounding box for the speaker and the text data overlap which lead to a fixation position being classified into two categories if it belonged to the overlapping region.

5.2 Feature Engineering

After drawing the bounding boxes around the three required regions, we were able to classify each fixation point into atleast one class. Since, in some cases there was an overlap between the text and speaker bounding box, the sum of all the percentages of the frequencies is greater than 100. Figure 4 is a plot demonstrating the percentage wise frequencies of the different classes.

Figure 5 represents the points classified into different classes by their label. The points labelled with white belong to the "speaker" class. Similarly the green, blue and black points correspond to the "text", "visual" and "other" class respectively. In Figure 5, the number of points belonging to each class is only for representational purposes. It is not the final number of points which belonged to each category. This figure was constructed using only a section of the data.

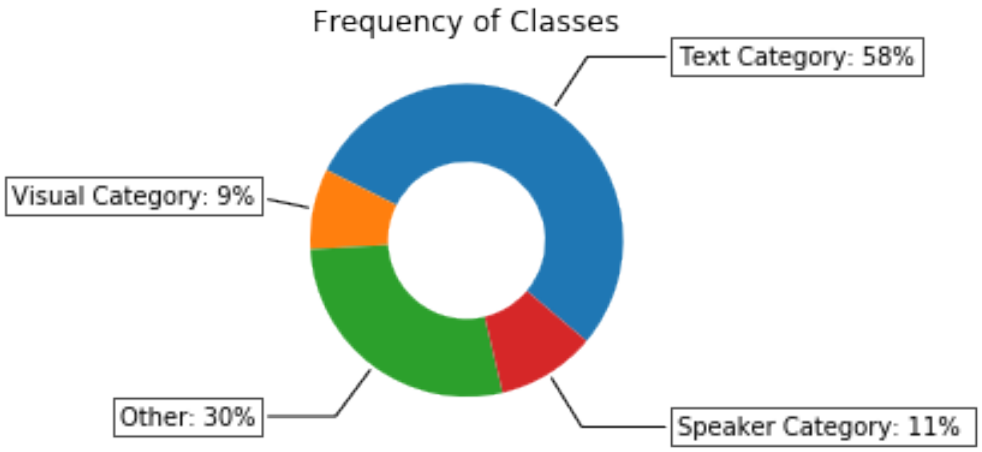


Figure 4: Percentage Wise Frequency Distribution of Classes

Figure 5: Points Classified into Bounding Boxes

We also knew beforehand the frames which contain the required information for each question, thus we were able to calculate the scores for each of the features for each question. The scores are derived as the fraction of the duration for which the student looked at that particular category. Algorithm 1 was followed for calculating the scores for each feature.

Algorithm 1 Calculating Feature Scores for Every Question

Result: Feature Scores

```
while ∃ question do
    initialize: text = 0, visual = 0, speaker = 0, other = 0
     find frame_start and frame_end
     if fixation_point ∈ text_bounding_box then
         text + = current_fixation_duration
    else
         if fixation_point ∈ visual_bounding_box then
             visual + = current_fixation_duration
         else
         end
         if fixation_point ∈ speaker_bounding_box then
             speaker + = current_fixation_duration
         else
             other + = current_fixation_duration;
         end
    end
end
return text, visual, speaker, other
```

The final step involved normalizing the scores so that the sum of all the scores for each feature summed to 1 for every question. This step was necessary since the number of total frames which needed to be processed was different for every question.

5.3 Correlations

We use the Karl Pearson Correlation to determine the degree and direction of the relationship between the variables in our model. It is obtained via a Least-Squares fit and a value of 1 represents a perfect positive relation-ship, -1 a perfect negative relationship, and 0 indicates the absence of a relationship between variables.
It is defined as below:

$$\implies \rho = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

$$\implies \rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}}$$

We conduct the analysis for the correlations in three parts. First, we try to do find the correlations between the features and the post exam score for the complete data (across all the modules and question types). Second, we find the relationship between the features and scores across all the modules but only for a particular question type (i.e. for infer and recall type of questions). We conduct the analysis in this manner because we expect that the difficulty level of the question varies according to the type of questions. Literature suggests that image data is more useful for remembering concepts and thus it would lead to a better performance for recall type of questions.

5.3.1 Complete Data

Figure 6 and Table 1 show the Karl Pearson Coefficient of Correlation for the complete data. We can see that there is a negative correlation between the post exam score field (answer) and the in_text as well as the in_image field. This negative correlation makes our assumption void. However, we do see a positive correlation between the post exam score and the in_speaker field. By finding out the p-values while testing for the significance of these correlation coefficients we find that the former two coefficients are not significant whereas the result for the relationship between the post exam score and in_speaker field holds significant. We assume that a p-value below 0.05 is significant.

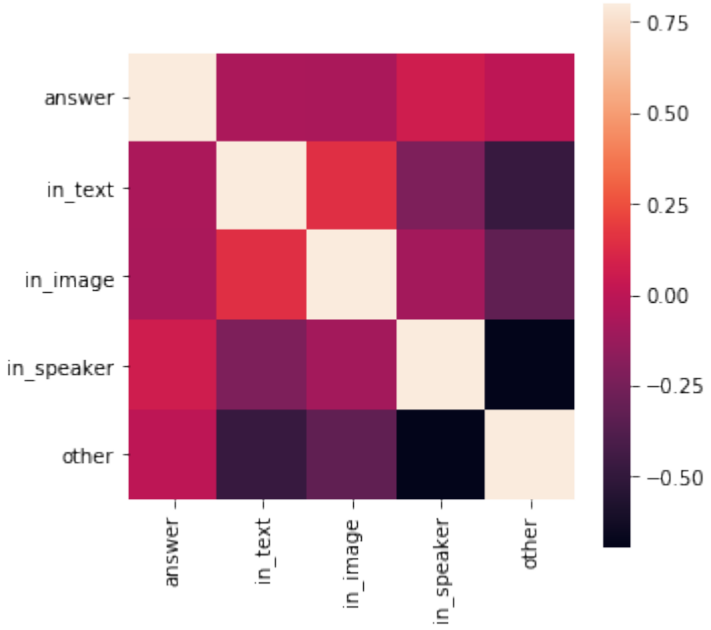


Figure 6: Correlation plot for all questions

	answer	**in_text**	**in_image**	**in_speaker**	**other**
answer	1.000	-0.060	-0.067	0.066	0.001
in_text	-0.060	1.000	0.150	-0.220	-0.478
in_image	-0.067	0.150	1.000	-0.090	-0.328
in_speaker	0.066	-0.220	-0.090	1.000	-0.696
other	0.001	-0.478	-0.328	-0.696	1.000

Table 1: Correlations between features for all questions

	answer	**in_text**	**in_image**	**in_speaker**	**other**
answer	0	0.935	0.338	0.038	0.078

Table 2: p-values for significance testing

5.3.2 Infer Type Questions

We conduct a similar analysis to find the relationship between the post exam scores and the features for only the "infer" type of questions. We again

see that there is a negative correlation between the score and in_text and in_image features and a positive correlation with the in_speaker feature. The p-values for the correlation coefficients are again insignificant for the in_text and in_image fields and it is significant for the in_speaker field.

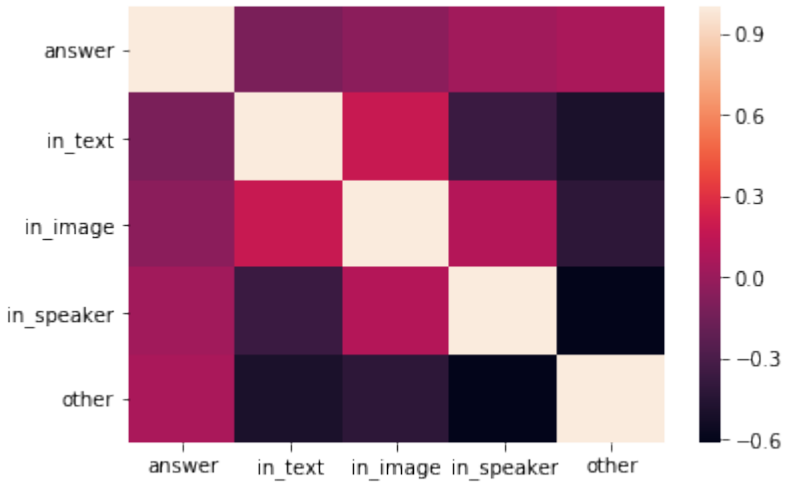


Figure 7: Correlation plot for infer type of questions

	answer	**in_text**	**in_image**	**in_speaker**	**other**
answer	1.000	-0.113	-0.046	0.036	0.065
in_text	-0.113	1.000	0.179	0.365	-0.493
in_image	-0.046	0.179	1.000	0.102	-0.419
in_speaker	0.036	-0.365	0.102	1.000	-0.611
other	0.065	-0.493	-0.419	-0.611	1.000

Table 3: Correlations between features for infer type questions

	answer	**in_text**	**in_image**	**in_speaker**	**other**
answer	0	0.585	0.138	0.026	0.062

Table 4: p-values for significance testing

5.3.3 Recall Type Questions

In the third component of our analysis we try to find the above mentioned relationships only for the "recall" type of questions. We again get the same results for the correlation coefficients as well as the p-values. One difference however we do see between the "recall" type of questions and "infer" type of questions is that the negative correlation is less strong for the in_image field as compared to the in_text field. This result in a way supports our initial hypothesis of visual data being a better method for learning as compared to textual data.

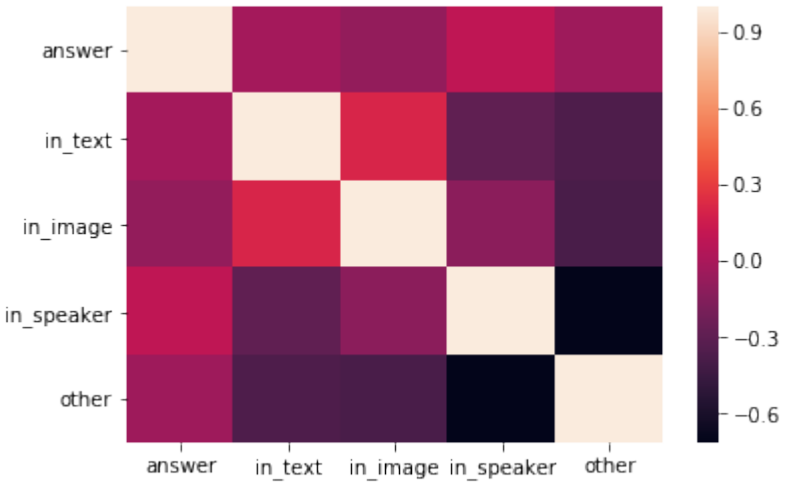


Figure 8: Correlation plot for recall type of questions

	answer	**in_text**	**in_image**	**in_speaker**	**other**
answer	1.000	-0.017	-0.278	0.089	-0.043
in_text	-0.017	1.000	0.207	-0.288	-0.362
in_image	-0.078	0.207	1.000	-0.118	-0.382
in_speaker	0.089	-0.288	-0.118	1.000	-0.714
other	-0.043	-0.362	-0.382	-0.714	1.000

Table 5: Correlations between features for recall type questions

	answer	in_text	in_image	in_speaker	other
answer	0	0.779	0.088	0.021	0.066

Table 6: p-values for significance testing

5.4 Prediction Models

We tried out multiple classification models with different hyperparamters. Figures 9 and 10 show the accuracies and loss of these models.

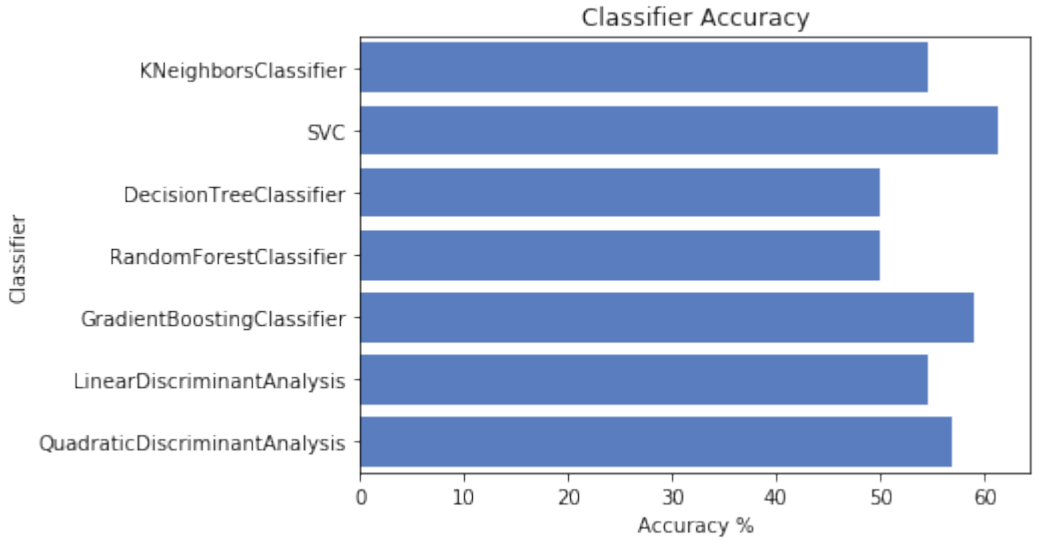


Figure 9: Accuracies of Various Classification Models

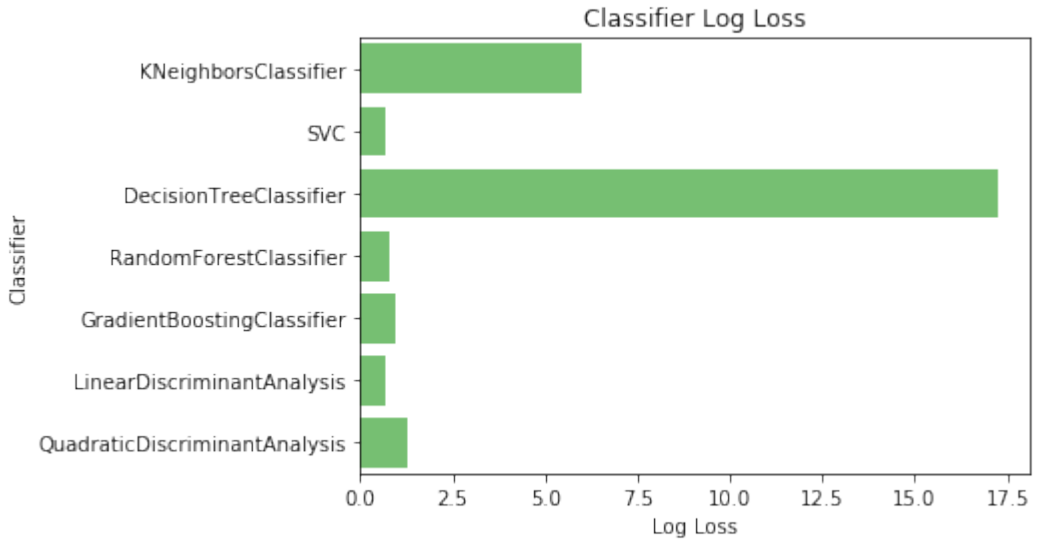


Figure 10: Log loss of Various Classification Models

We got the best results using the following three models:

1. Gradient Boosting Classifier
2. Support Vector Machine

5.4.1 Gradient Boosting Classifier

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Table 7 shows the Mean Square Error and the Mean Absolute Error for the Gradient Boosted Classifier Model.

	Mean Square Error	Mean Absolute Error
Gradient Boosting Classifier	0.331	0.374

Table 7: Errors for Gradient Boosting Classifier Model

5.4.2 Support Vector Machine

SVMs are supervised learning models that analyze data used for classification and regression analysis. Given a set of labelled training examples, a SVM

builds a model that assigns new examples to one category or the other, by using a decision boundary for the classes. Table 8 shows the Mean Square Error and the Mean Absolute Error for the Support Vector Machine Model with two different kernels.

	Mean Square Error	Mean Absolute Error
SVM (kernel = 'poly')	0.271	0.331
SVM (kernel = 'RBF')	0.229	0.353

Table 8: Errors for Support Vector Machine Model

5.4.3 2 Layer Neural Network

Neural networks are a framework for many different machine learning algorithms to work together and process complex data inputs. They learn to perform tasks by considering examples, generally without being programmed with any task-specific rules. We tried Neural Networks with various optimizers and got the best accuracy with the SGD optimizer. Figure 11 shows the accuracies of various Neural Network architectures which we tried.

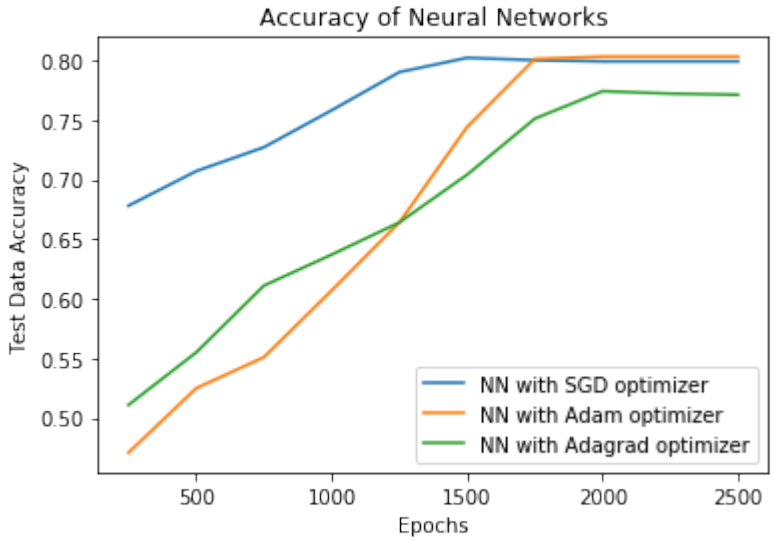


Figure 11: Accuracies of Neural Networks with different optimizers

	Mean Square Error	Validation Loss
Neural Network (optimizer = 'sgd')	0.201	0.221
Neural Network (optimizer = 'adam')	0.197	0.230
Neural Network (optimizer = 'adagrad')	0.221	0.255

Table 9: Errors for Neural Network

5.5 Results

We achieved the highest accuracy on the test set using a 2 layer Neural Network with a 'SGD' optimizer function. The performance of other classification models has been tabulated above and it can be seen that amongst the other models, the gradient boosting classifier model and the support vector machine model performed the best. Since it is an ensemble based method, it builds the model in a stage-wise fashion and hence it generalizes better as compared to other models which leads to a higher test-set accuracy. The complete code for this project can be found on https://github.com/mayank26saxena/visual-textual-data-analysis

6 Conclusion

Firstly, we can conclude that student's had an improved post exam score in the instances when they spent more time looking at the speaker as compared to the visual and textual data for a particular question. There was a relatively weaker negative correlation between the in_image score and the post exam score. Similarly, there was a relatively stronger negative correlation between the in_text score and the post exam score. We also saw that there was no difference in the sign of these correlations when we considered the difference in the type of questions (inference and recall). However, there was a relatively stronger correlation between post exam scores and in_image score for recall type of questions.
We also see that there exists a much more stronger positive correlation between the in_speaker score and the student's post exam score. Thus, we can conclude that looking directly at the speaker while he/she is delivering the lecture content is a much more efficient method to learn than by looking at the content in the slides.
Secondly, we can also show that we can predict with nearly 80% accuracy

whether the student will get a particular question right or wrong from the features we have engineered. This shows that the features we have derived are strong predictors for our model. After trying out a variety of classification models, a neural network model with 2 layers and an SGD optimizer gave us the highest classification accuracy on the test dataset.

7 Future Work

There is a lot of scope for future work in this study. To test the effectiveness of textual data and visual data in lecture slides, we can use two kinds of lecture slides - one which has a majority of text and the other which has a majority of visual content. Comparing the post exam scores of the student's in both cases, we will be able to measure the goodness of our hypothesis. Further data collection and experimentation could be done to add more data to the training dataset and improve the classification accuracy which we have achieved.

References

[1] http://www.oby.no/students-learn-best-reading-text-watching-message-film/?lang=en

[2] https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/24815/3/Roberts_document.pdf

[3] https://www.psychologytoday.com/us/blog/get-psyched/201207/learning-through-visuals

[4] https://oli.cmu.edu/wp-content/uploads/2012/05/Davenport_2008_Framework_For_Designing_Relevant_Representations.pdf

[5] http://www.indiana.edu/~pcl/rgoldsto/courses/dunloskyimprovinglearning.pdf

[6] https://www.shiftelearning.com/blog/bid/350326/studies-confirm-the-power-of-visuals-in-elearning

[7] https://journal.lib.uoguelph.ca/index.php/perj/article/download/3137/3473/0

[8] https://elearningindustry.com/visual-learning-6-reasons-visuals-powerful-aspect-elearning

[9] http://blogs.edweek.org/teachers/teaching_now/2015/06/visual-diagrams-help-students-take-notes.html

[10] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5256450/