A new style of teaching: Exploring the benefits of visual language learning

Yuna Shono
Affiliation
City, State, Country

ABSTRACT

This project was designed to investigate the efficiency and likelihood of people learning a new language through means of an AI that would help translate objects in day to day life to a language that the user would want to learn. This specific project would help to benefit a better understanding of languages and enable the user to have better connection and communication skills with others. The overall approach taken to achieve this goal came in the form of two parts: object detection and natural language processing (NLP). The first part contained training the AI on certain categories of objects which would then be put through the second part of the AI, the translation. The user would also be able to interact with the AI by inputting their own data and selecting a certain language they would like to learn. The AI was seen to have around a 35% accuracy rate in determining the correct image but had a few errors when encountering the translation side of things. Regardless, some important conclusions were that the AI was successful in being able to translate languages and would certainly help people in their language learning endeavors.

INTRODUCTION

It is a common problem for many people to easily become discouraged in their language learning path. To combat this, an AI was created to help make language learning more fun by having an interactable image detection translator app. This language learning had to be done visually as learning languages is best through visual cues. A research study in 2014 states "it not only exposes learners to authentic language use (Lin 2014) but the combination of different input modes such as imagery and audio may also stimulate various aspects of second language learning such as comprehension or vocabulary" (Perez). To make this possible, a supervised classification AI has been used to show the accuracy of translating an image from image data alone. The output would provide the user with the translation of that image in the language the user wishes to learn. The aim of this whole project is to give the user a motivational push to learn the language.

BACKGROUND

When looking for evidence to help support this problem and why it is important, many articles came up with the idea that having a visual learning environment would be easier for people to pick up languages. As stated in an article called "Embodied cognition and language learning in virtual environments" by Yu-Ju Lan, "the avatar-based embodied motions are sufficient and strong enough to originate the essential internal mediation in learners' brains and consequently have an effect on language comprehension and acquisition." While this article mainly talks about the effects of having people control an avatar and learning better by having more visual cues as to what a person is trying to say, it does tie into the connection in the brain between language comprehension and visual motion. This would mean that having an AI that presents the user with the opportunity to learn languages through visual cues would help more with language acquisition. While the research aspect of looking at how the brain works with language acquisition is interesting, it's worth it to note if other people have created something similar to a visual language learning environment.

One example of this visual language learning environment is Rosetta stone. This is a language learning app that has many interactive games and listening comprehensions for the learner to complete. Through my own personal experience, Rosetta Stone has presented me with multiple ways to learn a language in a more efficient way and made it easier for me to comprehend that language. For their visual learning, Rosetta stone has implemented a system called "Seek & Speak" which has the user point the iPhone camera at an object and see the vocabulary word in a chosen language, then practice a conversation using the word (Venturebeat).

While this game is a fun addition to Rosetta Stone, it is not entirely easy to use in a situation where the user wants to just know the word. Instead of having to seek out the object and practice using it in a conversation, it would be much better to let the user learn the word and hence use that in their day to day lives instead of trying to consistently implement it in a sentence every time the app is used.

DATASET

The dataset used for the project was cifar10. This was done because otherwise the dataset would be too big and would take too long to process. The type of data consisted of images with 10 different labels (namely: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck) and 5000 images for each label. This would mean there were 50000 images in total for the whole dataset. The training and testing datasets were split evenly 50/50 which was probably not the best idea as having more training data could be better for the testing data.

This choice was done to maximize limited compute time and be able to run the training in a reasonable time frame (<1 hr). The image data was all the same size to ensure that there was no error in the training so all images were scaled down to 32x32 pixels. The image dataset itself however had a different size based on the type of model used. The transformer model (final model used) used the same amount of images in the dataset. However, the baseline model used 10000 images less than the original dataset size and the neural networks only used 10000 images. It could've been better however to assess how the AI did on certain categories of images and then update how many images are in the dataset based on the accuracy measurement for each category.

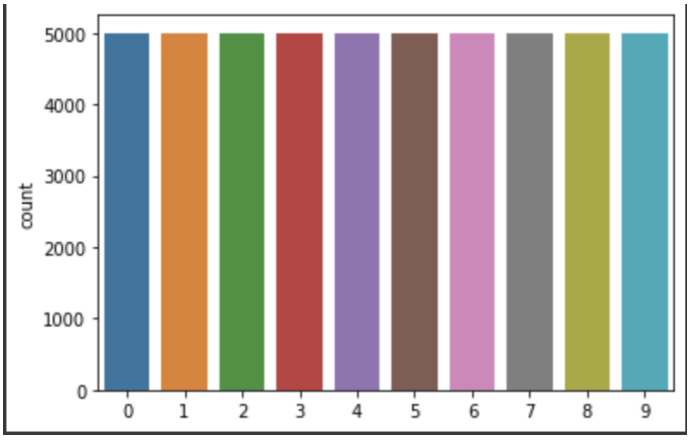

Figure 1: Dataset in graph form. Number of images v number of categories

METHODOLOGY/MODELS

There were multiple methods used to solve this research problem involving testing different models to find the one with the best accuracy as well as testing out a new type of model to see if its learning process was much better than the other more basic models.

The data preprocessing part was done in multiple steps: reshaping the data, printing out the data as gray images instead of color for better accuracy, and finally training and testing the data using train test split. The data came out to be in a 32 by 32 pixel and in a normal color. The gray images did not seem to provide much of a difference in terms of accuracy.

Much of the data was also cut down in terms of the amount of images used since training the data into the model multiple times over meant that it would take some time to process. The images were also flattened to train it on the model.

After the data preprocessing process, basic baseline models were used and tested to ensure that the data was being processed smoothly as well as looking at the mean absolute error and the accuracy score.

The baseline models tested included the KNN model (K Nearest Neighbors), the logistic regression model, and the decision tree classifier. These three models were the bare basics of the AI that were used to test this data. KNN model is a model that essentially determines what group a datapoint is in by looking at the data points around it. I chose KNN to be one of the first testing models because its accuracy is normally quite high when dealing with classification and also the dataset had many data points but not that many classification segments meaning that it would be easier for KNN to classify data points into the right categories. The second baseline model used was the logistic regression model. This model models the probability of an event taking place and was used for this particular problem because the model is made for data that needs a high accuracy prediction rate and a classification solution. The third baseline model used was the decision tree classifier. This model predicts the value of a target by learning simple decision rules inferred from the data's features. This model was used because it is good with classification. In terms of accuracy, logistic regression produced the best results with a score of 40.33%.

A simple convolutional neural network was tested next. It had 4 layers each with different filters, a kernel size of 3, stride of 1, "same" padding, and three of the layers contained "relu" activation while the last one contained "sigmoid".

To make the accuracy even better when testing and training the model, we decided to use a transfer learning model called VGG 16. This model is already trained on previous data and will use that training to help determine the categories for each of the images after being trained again on our current data set. This model had 9 different layers: 1 base model, 1 global pooling average 2D, 4 dense layers and 3 drop out layers. This model produced a validation accuracy of 0.3901 or 39%.

The mean absolute error represents the magnitude of difference between the prediction of an observation and the true value of that observation while the accuracy score just represents how accurate the model is at pairing up the right image with the right label.

For the language translation part of the problem, we decided to hardcode each of the languages into the model. This would ensure accuracy and would be less intensive computing wise. There is also an easier module to replace this down the line with web based translation services. In this case, it was also easier to hardcode because there were not many categories to hardcode into a different language and only five languages were chosen to make the translation part easier.

RESULTS

For the KNN model, the accuracy was 35.79% for 1 nearest neighbor, 34.22% for 3 nearest neighbors and 35.25% for 5 nearest neighbors. The decision tree classifier had an accuracy of 18.98% and the logistic regression model had an accuracy of 40.33%. The VGG16 model got an accuracy score of 30.35% and a validation accuracy of 39.01%.

CONCLUSION

To conclude, this project involved making an AI which could both process images and also be able to translate the labels of these images into different languages. There were some limitations to this process of applying it into real life. The amount of categories used for the pictures weren't a lot and they were mainly on very basic options that could be seen as quite distinct from each other. The dataset was also massive so it did take some time for it to train on certain models and I would spend more time waiting for the code to compile/train rather than actually assessing the behavior of the model.

I think the model performed alright. I was expecting the VGG16 model to have a higher accuracy but it ended up being the logistic regression model that had the highest accuracy. This may be because of the dataset that was used. The baseline models were trained on a dataset that had gray images and so it may have given the models an advantage on that type of image dataset. However, the dataset the VGG16 model was trained on was images that were in color.

The future steps for this project would be to implement a translation AI that is not hardcoded into the system and also training that AI model with the image AI model.

ACKNOWLEDGMENTS (ACK. CLAUSE TITLE)

I would like to thank my mentor for guiding me throughout this whole research project and I would also like to thank my parents for their constant support monetarily.

REFERENCES

"Video and Language Learning." Taylor & Francis, https://www.tandfonline.com/doi/full/10.1080/09571736.2019.1629099.

Lan, Yu-Ju, et al. "Embodied Cognition and Language Learning in Virtual Environments - Educational Technology Research and Development." SpringerLink, Springer US, 25 Aug. 2015, https://link.springer.com/article/10.1007/s11423-015-9401-x.

Horwitz, Jeremy. "Rosetta Stone for IPhone Adds AI to Identify Objects for Live Translations." VentureBeat, VentureBeat, 22 Jan. 2019, https://venturebeat.com/2019/01/22/rosetta-stone-for-iphone-adds-ai-to-identify-objects-for-live-translations/.

"CIFAR-10-PYTHON." CIFAR-10 and CIFAR-100 Datasets, https://www.cs.toronto.edu/~kriz/cifar.html.