

Case Study Rubric - ASL Digit Translation

DS XXXX

Due: Month Day, Year

Submission format:

- Upload link to Github repository.

Individual Assignment

General Description: Submit to canvas a link to your case study repository.

Preparatory Assignments – Everything in your course thus far, but especially those related to sentiment analysis.

Why am I doing this?

This assignment serves as a way to familiarize yourself with using convolutional neural networks (CNN) to process and categorize image data. By the end of this assignment, you should have an understanding of how image data are prepared for testing, normalized, and categorized using packages like TensorFlow. You will also have created a working machine learning model that can be extended in the future to translate additional static ASL signs and/or phrases.

What am I going to do?

Over the course of this assignment, you will use the script and data provided to analyze a dataset of 2180 different ASL digit signs in order to classify each image as one of ten digits using the TensorFlow package. Your final deliverable will include the output from your CNN analysis, as well as the various statistical tests and visualizations you will produce over the course of the main script.

Tips for success:

- Be sure to read and understand the documentation behind TensorFlow - this is a high-level machine learning package, which means that it will benefit you to understand each individual step of the process in case something goes wrong when you use it to analyze the image data.
- Validate your results by looking at the data yourself, in order to ensure that everything is working as it should be and there aren't any obvious bugs in your code.
- Name your output files and any additional variables you create so that an outsider to your project would be able to understand what they contain/mean.
- Consult with your course instructors and fellow students if you have any questions.
- Remember - no model is perfect, so don't stress if your results aren't 100% accurate.

How will I know I have Succeeded? You will meet expectations on this case study when you follow the criteria in the rubric below.

Formatting	Github repository submitted via link on Canvas containing: <ul style="list-style-type: none">- A README.md file documenting the steps you took to complete the analysis and any sources you used along the way- A license file (MIT preferred)

	<ul style="list-style-type: none"> - Data folder containing the data you used - Scripts folder containing the scripts you used - Output folder, containing any exploratory data analysis, output visualizations, or statistical test output you generated as part of the process
README.md	<p>.md file in the base folder of the repository containing:</p> <ul style="list-style-type: none"> - Your name, class, and date - Software, platform, and any packages used - Map of your Github repo, including all folders and documents contained within, with descriptions. - Full documentation of steps to reproduce the results you generated, including (but not limited to) data uploading, data cleaning, train/test separation, the various components of the main script, and generation of outputs/visualizations. The goal here is to make it so that an outsider to your project could read this section and easily replicate your results on their own machine. - Any sources you used during this project, cited in standard IEEE format (link).
Data	<p>A folder containing a link to the data that you used for this project (provided in the Github repo as a link, which we recommend keeping as the linked dataset is quite large). Also include any additional data you used at any point during this assignment.</p>
Scripts	<p>A folder containing the main script that you used for this project. Also include any additional scripts or other code you used at any point during this assignment. This will include:</p> <ul style="list-style-type: none"> - Load all packages necessary for your code - Loading the data from the linked Github repo: we recommend cloning this data to a Google Drive folder so you can access it from Google Colab more easily - Cleaning and splitting the data into training and testing sets - Building and training your CNN model using TensorFlow or a similar package - Generating a confusion matrix for your results - Evaluating the model using various metrics, including accuracy, precision, recall, and more - Generating a visualization for the training and validation accuracy over each epoch of your CNN model <p>Be sure to comment on each stage of your script so that an outsider could follow along with no prior knowledge of your project.</p>
Output	<p>A folder containing all the exploratory data analysis, model output, visualizations, and other products of the scripts you used over the course of the project. This should include, at minimum:</p> <ul style="list-style-type: none"> - EDA demonstrating the composition of the dataset (num. of categories, num. of participants in each category)

	<ul style="list-style-type: none"> - Table showing the results of your CNN analysis (confusion matrix) - Table with precision, recall, F-1 score, and other metrics if relevant - Chart visualizing training and validation accuracy over the epochs of your CNN model <p>Be sure to use clear and accurate names for your files.</p>
--	--

Acknowledgements: Thank you to Prof. Rasero for providing the rubric structure. Additionally, thank you to Zoe Averill and Ravza Aykan from previous DS courses for providing examples of similar assignments.