ISSN 2959-6157

Improving Marvel Hero Classification through Dataset Curation

Yu Jiao

Abstract:

PART 1: In the domain of computer vision, this study explores the development of a computer vision model for classifying Marvel superheroes such as Black Widow, Hulk, Iron Man, and Spider-Man. Utilizing a curated dataset sourced from Kaggle, the research emphasizes the critical role of dataset quality in refining model accuracy. Insights gained include adjustments to neural network configurations and leveraging Edge Impulse for enhanced performance. The findings highlight effective strategies for optimizing classification accuracy in complex image recognition tasks.

PART 2:This part explores the application of Bayesian logistic regression to model the relationship between temperature and the probability of O-ring failure. By leveraging Bayesian inference techniques, analyzing historical data to quantify the risk associated with temperature variations and emphasize the importance of probabilistic approaches in safety-critical decision-making.The Space Shuttle Challenger disaster on January 28, 1986, remains a poignant case study in aerospace engineering failure. The investigation concluded that the failure of O-ring seals in cold temperatures led to the tragic loss of the shuttle and its crew.

Keywords:

PART 1:Marvel superheroes, computer vision, dataset curation, neural network configuration, Edge Impulse, image recognition

PART 2:Space Shuttle Challenger disaster, O-ring failure, Bayesian logistic regression, temperature dependency, probabilistic modeling, aerospace engineering, safety assessment

In the realm of computer vision, the classification of Marvel superheroes is a captivating challenge that pushes the boundaries of image recognition algorithms. I have embarked on an exciting project titled “Marvel Hero Classify” ,which aims to develop a computer vision-based model capable of accurately identifying various Marvel superheroes, including Black Widow, Hulk, Iron Man, and Spider-Man.

The project’s initial steps involve setting up the necessary infrastructure, with the operation guide page featuring images as the primary entry point, followed by the classification of a single object on the subsequent page.

Upon initiating data acquisition, the Edge Impulse offers multiple functions, enabling this project to upload and export data. The dataset is downloaded from Kaggle, with computer vision employed to filter the retrieved data, thereby preserving only the image component. Unzip the dataset, and upon inspection, it is divided into two folders: “test” and “train”.

Upload the data within the training folder, consisting of images of Black Widow, Hulk, Iron Man, and Spider-Man in sequence. Constantly monitor the label position to prevent uploading all photos for a single hero. Subsequently, upload the data within the test folder while observing strict protocols.

After the dataset is generated, proceed to create an impulse in the impulse design module, and click “save impulse” upon completion of the selection process.Then enter “Images” and select “RGB” in respect to color depth. Given the abundance of training images and longer overall duration, 20 iterations are chosen as the training frequency. Upon conducting the training, I noticed that the outcome was not as anticipated; several heroes were erroneously identified as Spider-Man, with only the training model for Spider-Man demonstrating acceptable performance.

Consequently, I revisited the dataset and discovered that numerous images depicted multiple heroes, being auto-identified as Spider-Man. This enlightened me to the fact that enhancing the training model’s accuracy hinges greatly on the dataset’s precision, even expand the dataset and accomplish automatic rotation and other functions,the results did not vary significantly.

Howver alter the last neuron in the neural network configuration to 128, and the outcome will significantly improve. The rationale behind this is that a larger number of neurons results in a swifter training process. Consequently, the accuracy rate experiences a marginal enhancement.

Overall, leveraging Edge Impulse for model training and learning facilitated a better understanding of embedded machine learning applications. It offers a wealth of features and support, streamlines the development process, and provides real-time feedback and debugging tools aiding in the rapid and efficient construction of high-performance embedded machine learning applications.

Space Shuttle Challenger Disaster

The Space Shuttle Challenger Disaster remains one of the most tragic and scrutinized events in the history of space exploration. On January 28, 1986, the Space Shuttle Challenger broke apart just 73 seconds into its flight, resulting in the loss of all seven crew members aboard.

The cause of the disaster was determined to be the failure of an O-ring seal in one of the solid rocket boosters. The O-rings, which were meant to prevent hot gases from escaping during the launch, failed due to the unusually cold temperatures on the morning of the launch. This failure allowed hot gases to breach the rocket's external fuel tank, leading to the catastrophic explosion.

For the Space Shuttle Challenger Disaster problem, a suitable probabilistic model to capture the relationship between temperature and the probability of O-ring failure is a logistic regression model.

Logistic Regression Model:

Model Formulation:

Let (T) represent the temperature (in Fahrenheit) and (F) represent the binary outcome variable indicating O-ring failure (1 for failure, 0 for success). We assume that the probability of O-ring failure, (p), follows a logistic function:

[p(T) = \frac{1}{1 + e^{-z}}]

where (z) is a linear combination of temperature and model parameters:

[z = \beta_0 + \beta_1 \cdot T]

Here, (\beta_0) is the intercept parameter, and (\beta_1) is the coefficient associated with temperature.

Bayesian Formulation:

In a Bayesian framework, we assign prior distributions to the model parameters (\beta_0) and (\beta_1) and use Bayes' theorem to update these priors based on observed data, yielding posterior distributions for the parameters.

Priors:

(\beta_0 \sim Normal(\mu_0, \sigma_0^2))

(\beta_1 \sim Normal(\mu_1, \sigma_1^2))

Likelihood:

(F_i \sim Bernoulli(p(T_i))) for each observation (i)

Posterior:

(\beta_0, \beta_1 | {T_i, F_i} \sim ?) (to be determined using MCMC methods)

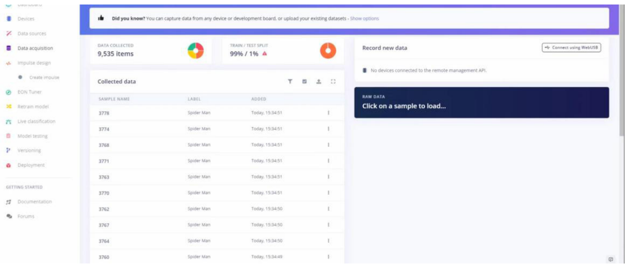
Implementation in PyMC:

The logistic regression model can be implemented in PyMC using the pm.glm.GLM class, which provides a convenient interface for specifying generalized linear models. We'll specify the model formula, likelihood distribution, and prior distributions for the model parameters. This code defines a Bayesian logistic regression model using PyMC. It takes temperature data and O-ring failure data as input, and then uses Markov Chain Monte Carlo (MCMC) sampling to obtain estimates of the model parameters from the posterior distribution. Finally, it outputs summary statistics of the posterior distribution, including the mean, standard deviation, and other relevant information for each parameter.

Based on the Bayesian logistic regression model, there is a moderate likelihood of O-ring failure at a temperature of 75°F. However, the margin of error in this projection is relatively substantial, as exemplified by the expansive credible interval. This underscores the necessity of taking uncertainty into account when making predictions, particularly in critical situations such as engineering decisions that involve safety.

References

Kaggle dataset: [https://www.kaggle.com/datasets/hchen13/marvel-heroes]



Did you know? You can capture data from any device or development board, or upload your existing datasets - Show options

DATA COLLECTED
9,535 items

TRAIN / TEST SPLIT
99% / 1%

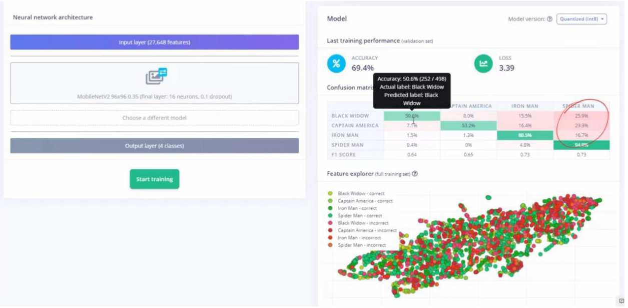
Record new data

No devices connected to the remote management API.

RAW DATA
Click on a sample to load...

Collected data

SAMPLE NAME	LABEL	ADDED
3778	Spider Man	Today, 15:34:51
3774	Spider Man	Today, 15:34:51
3768	Spider Man	Today, 15:34:51
3771	Spider Man	Today, 15:34:51
3763	Spider Man	Today, 15:34:51
3770	Spider Man	Today, 15:34:51
3762	Spider Man	Today, 15:34:50
3767	Spider Man	Today, 15:34:50
3764	Spider Man	Today, 15:34:50
3760	Spider Man	Today, 15:34:49



Neural network architecture

Input layer (27,648 features)

MobileNetV2 96x96 0.35 (final layer: 16 neurons, 0.1 dropout)

Output layer (4 classes)

Model

Model version: Quantized (int8)

Last training performance (validation set)

ACCURACY
69.4%

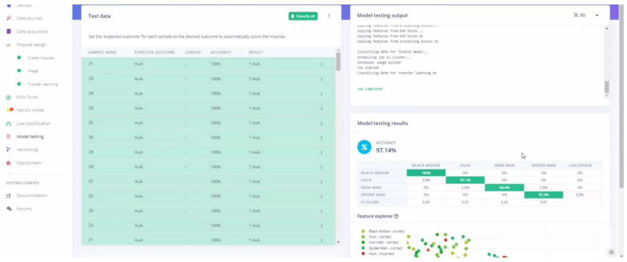
LOSS
3.39

Accuracy: 50.6% (252 / 498)
Actual label: Black Widow
Predicted label: Black Widow

Confusion matri

		PTAIN AMERICA	IRON MAN	SPIDER MAN
BLACK WIDOW	50.6%	8.0%	15.5%	25.9%
CAPTAIN AMERICA	7.1%	53.2%	16.4%	23.3%
IRON MAN	1.5%	1.3%	80.5%	16.7%
SPIDER MAN	0.4%	0%	4.8%	94.8%
F1 SCORE	0.64	0.65	0.73	0.73

Feature explorer (full training set)



Test data

Set the 'expected outcome' for each sample to the desired outcome to automatically score the impulse.

SAMPLE NAME	EXPECTED OUTCOME	LENGTH	ACCURACY	RESULT
31	Hulk	-	100%	1 Hulk
23	Hulk	-	100%	1 Hulk
24	Hulk	-	100%	1 Hulk
33	Hulk	-	100%	1 Hulk
32	Hulk	-	100%	1 Hulk
30	Hulk	-	100%	1 Hulk
29	Hulk	-	100%	1 Hulk
28	Hulk	-	100%	1 Hulk
27	Hulk	-	100%	1 Hulk
25	Hulk	-	100%	1 Hulk
26	Hulk	-	100%	1 Hulk
22	Hulk	-	100%	1 Hulk
21	Hulk	-	100%	1 Hulk

Model testing output

```
Copying features from processing blocks...
Copying features from DSP block...
Copying features from DSP block OK
Copying features from processing blocks OK

Classifying data for float32 model...
Scheduling job in cluster...
Container image pulled!
Job started
Classifying data for Transfer learning OK


Job completed
```

Model testing results

ACCURACY
97.14%

	BLACK WIDOW	HULK	IRON MAN	SPIDER MAN	UNCERTAIN
BLACK WIDOW	100%	0%	0%	0%	0%
HULK	2.9%	97.1%	0%	0%	0%
IRON MAN	0%	2.8%	94.4%	2.8%	0%
SPIDER MAN	0%	0%	0%	97.2%	2.8%
F1 SCORE	0.99	0.97	0.97	0.97	

Feature explorer

Feynman, R. P. (1986). Report of the Presidential Commission on the Space Shuttle Challenger Accident. Washington, DC: Government Printing Office.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Bayesian formulation used in Bayesian logistic regression.