Process Flow

Team : 10011

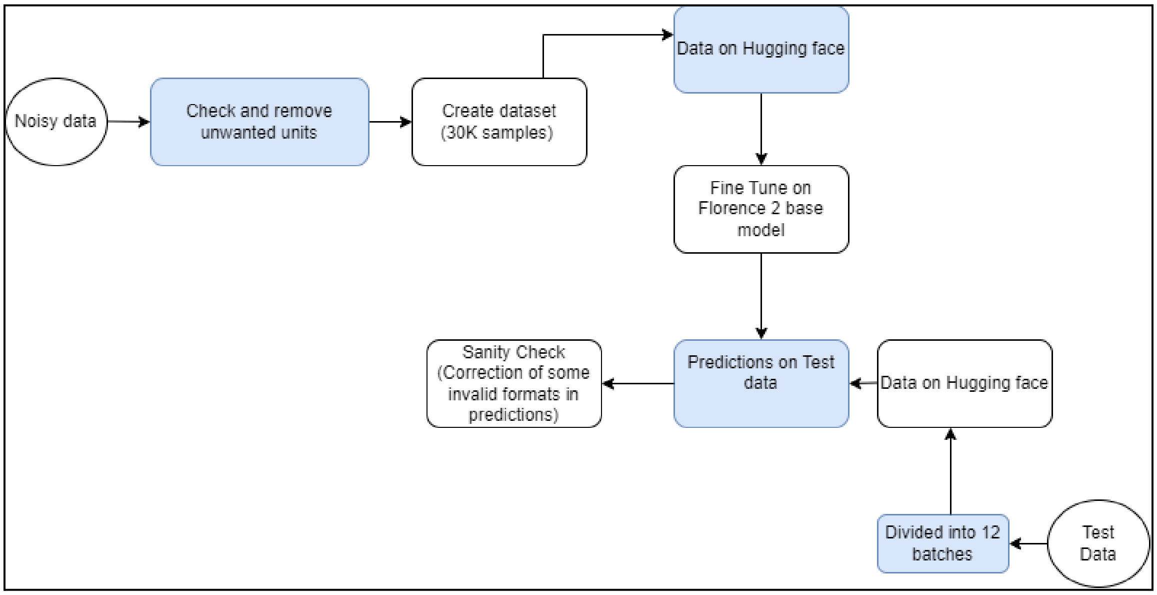
Noisy data
Check and remove unwanted units
Create dataset (30K samples)
Data on Hugging face
Fine Tune on Florence 2 base model
Predictions on Test data
Sanity Check (Correction of some invalid formats in predictions)
Data on Hugging face
Divided into 12 batches
Test Data

Figure : Flow diagram

□ **Handling Noisy Data :** Remove entity_value rows which are not present in entity unit map (constants.py). Also included all the kinds of possible entities for creating a subset as using the whole training data did not seem to be feasible.
Example : '10.0 horsepower', '2 bulbs' etc.
Notebook : 1_cleaning_data.ipynb

□ **Data Preparation (Dataset downloaded & processed on Hugging Face) :** Created dataset one with 30k samples using the train dataset provided as finetuning on the whole wasn't feasible. The data included all the parsable entity values which were in the constant.py. So we had all the data which was the possible output to the test.csv. We spent a fair amount of time to create a good pipe line for dataset and models fetching so we pushed all the images in the PIL format along with other columns of the dataset to hugging face private datasets which are accessible only using hf token.. We created around 13 datasets on hugging face 11 for test dataset and 2 for the fine tuning purpose.
Notebook : 2_upload_data.py

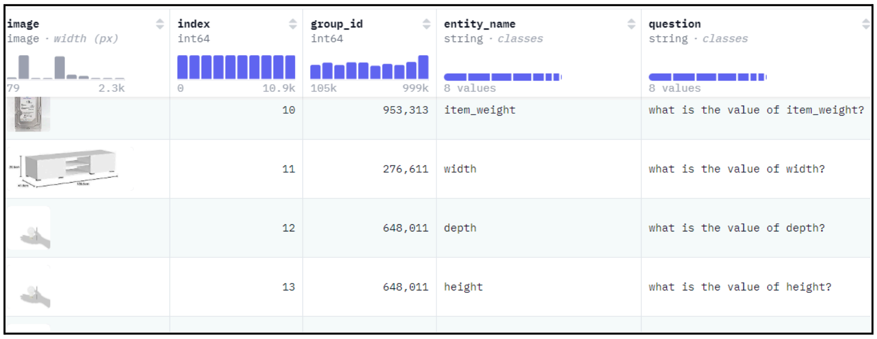

image	index	group_id	entity_name	question
image · *width (px)*	int64	int64	string · *classes*	string · *classes*
79 2.3k	0 10.9k	105k 999k	8 values	8 values
	10	953,313	item_weight	what is the value of item_weight?
	11	276,611	width	what is the value of width?
	12	648,011	depth	what is the value of depth?
	13	648,011	height	what is the value of height?

☐ **Fine tuning using Florence 2 :** We utilized the Florence-2 base-ft model for our fine-tuning experiment. This model is part of the Florence family of vision-language models and was loaded from "microsoft/Florence-2-base-ft" using the Hugging Face Transformers library. The model is of 0.23B parameters.
We fine tuned it using the prompt: **"<DocVqa>What is the value of {entity_name}?"**

We used a custom dataset "Sarvesh2003/ml_challenge_30k_train" from Hugging Face that we pushed earlier. The dataset was split into train, validation, and test sets:
Train set: 24,000 samples (80% of the original dataset)
Validation set: 3,000 samples (10% of the original dataset)
Test set: 3,000 samples (10% of the original dataset)
Hyperparameters:

- Number of epochs: 15
- Batch size: 6
- Learning rate: 1e-6
- Optimizer: AdamW

We leveraged NVIDIA L40 for the finetuning process from the lightning ai platform.
After each epoch we saved the model and also pushed it to hugging face.
Here while fine tuning the we didn't frozen any parameters and kept all updatable..
We observed a significant reduction in training loss over the course of the fine-tuning process:

- After 1 epoch: Training loss was 0.67 and validation loss was 0.2748
- After 6 epochs: Training loss decreased to 0.13 and validation decreased to 0.2345

Finetuned Model : https://huggingface.co/Sarvesh2003/florence_ft_base_Surya6
Reference : https://huggingface.co/blog/finetune-florence2
Notebook : 3_fine_tuning_florence_ml_challenge.ipynb

☐ **Predictions on test data :** Divided the test data into 12 batches because the testing set was having 1.3 lakh samples. So for faster processing we chose this approach.
Notebook : 4_Predictions_on_test_data.ipynb
Code to concat all the 12 subsets of the test.csv prediction:
5_concat_all_predicted_datasets.py

☐ **Post processing on predictions :** Some predicted values were having invalid format which was not acceptable. Example : '[2.0, 240.0] volt', '[10.0] watt', '[20.0 20.0] watt', '13.23.0 centimetre'.
We handled it using regular expressions and some other formulations.
Notebook : 6_post_processsing.ipynb

In conclusion, we fine-tuned the Florence-2 model for extracting entity values from images using a custom dataset. The process began with handling noisy data by removing irrelevant entity-value pairs and creating a 30K sample dataset based on parsable entity values. This dataset was uploaded to Hugging Face, where the Florence-2 base model was fine-tuned using a custom prompt. Training was done over 6 epochs with a batch size of 6 and learning rate of 1e-6 on the NVIDIA L40 lightning ai platform. The test data, containing 1.3 lakh samples, was divided into 12 batches and pushed to hugging face for faster predictions. Finally, post-processing was done to correct invalid formats in predictions using regular expressions.