MODEL CARD

LanguageDetector

MODEL DETAILS

A lightweight model (315kB in size) for predicting the language of a user's input text.

Model Type
An embedding-based neural network classification model.

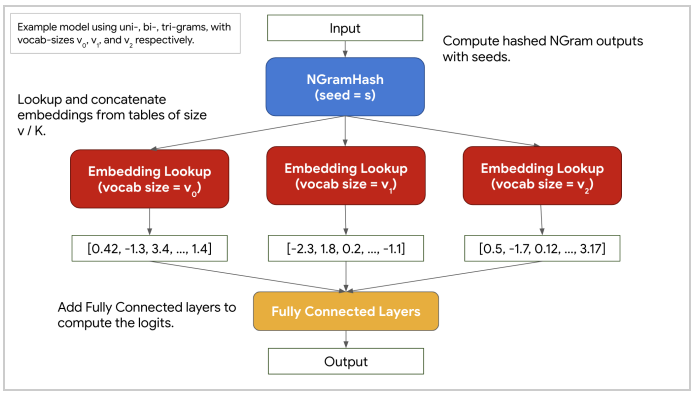
Input
String in UTF-8 format.

Output(s)
A list of predictions represented by pairs, with each pair consisting of:

- An ISO 639-1 language / locale code (e.g. "en" for English, "uz" for Uzbek, "ja-Latn" for Japanese (romaji)) as a string.
- The probability for the prediction as a float.

MODEL SPECIFICATIONS

Model Architecture
An embedding-based neural network classification model that has been optimized using techniques such as hashing and efficient embedding storage and lookup operations.



AUTHORS

Who created this model?
Learn2Compress Team, Google Research
Conceptual Understanding of Learning Architectures Team, Google Research

Who provided the model card?
MediaPipe Team, Google

DATE

May 10, 2023

LICENSED UNDER

Apache License, Version 2.0

Intended Uses

APPLICATION

- Predicting the language of an input text.
- Language identification preprocessing for downstream applications such as smart reply or optical character recognition (OCR).

DOMAIN AND USERS

- MediaPipe Tasks (e.g. SmartReply, OCR)

OUT-OF-SCOPE APPLICATIONS

Not appropriate for:

- Human life-critical decisions, including financial, contractual, legal or medical decisions.
- Predicting languages that are not one of the 110 preset languages (see “Prediction Limitations”).
- Multilingual text, as there is no guarantee that the model will predict every language present in the text or even assign the highest probability to the dominant language.

Limitations

PRESENCE OF ATTRIBUTES

The model supports 110 languages (ISO 639-1 language codes[1] are given below) and cannot predict any languages outside of that list.

TRADE-OFFS

This model’s size and latency have been optimized for on-device use-cases so it may not perform as well as a larger model intended for server-side use.

Ethical Considerations

HUMAN LIFE

This model is not intended for human life-critical decisions. The primary intended application is in language identification.

PRIVACY

The model was trained on anonymized web data translated by professional translators.

BIAS

The model has more data for popular languages like English and French.

[1] The suffix “-Latn” means the Latin alphabet is used.

Training Factors and Subgroups

TRAINING DATA

- All datasets were stripped of personally identifiable information before being used to train this model.

ENVIRONMENTS

The model can predict languages from each of the following geographic subregions (based on the United Nations geoscheme with mergers):

- Australia and New Zealand
- Melanesia, Micronesia, Polynesia
- Europe
- Central Asia
- Eastern Asia
- Southeastern Asia
- Southern Asia
- Western Asia
- Caribbean
- Central America
- South America
- Northern America
- Northern Africa
- Eastern Africa
- Middle Africa
- Southern Africa
- Western Africa

Evaluation metrics

Model Performance Measures

Latency

Average latency per inference call on an input text of 350 tokens measured on a Pixel2 device.

F1

The harmonic mean of precision and recall.

Precision

True positive rate among top language predictions.

Recall

True positive rate among ground truth language predictions.

Evaluation results

DATASET

The test data consists of multilingual text samples from en.wikinews.org, distributed under a Creative Commons v2.5 license.

EVALUATION RESULTS

Detailed evaluation is presented in the table below.

Latency (Pixel2)	F1	Precision	Recall
90 µs	0.9781	0.9893	0.9673