Reuters News Topic Classification

Hao Yin

About me

- Data Analyst in Fintech Startup Cmin
- Master in Data Science at WPI
- Bachelor in Finance at Jinan University

Goal

Tag the news with various topics with NLP technologies

Getting to know the data -- Reuters 21578

1. Dataset Introduction
2. Data Preprocess
3. Data Visualization

Reuters 21578

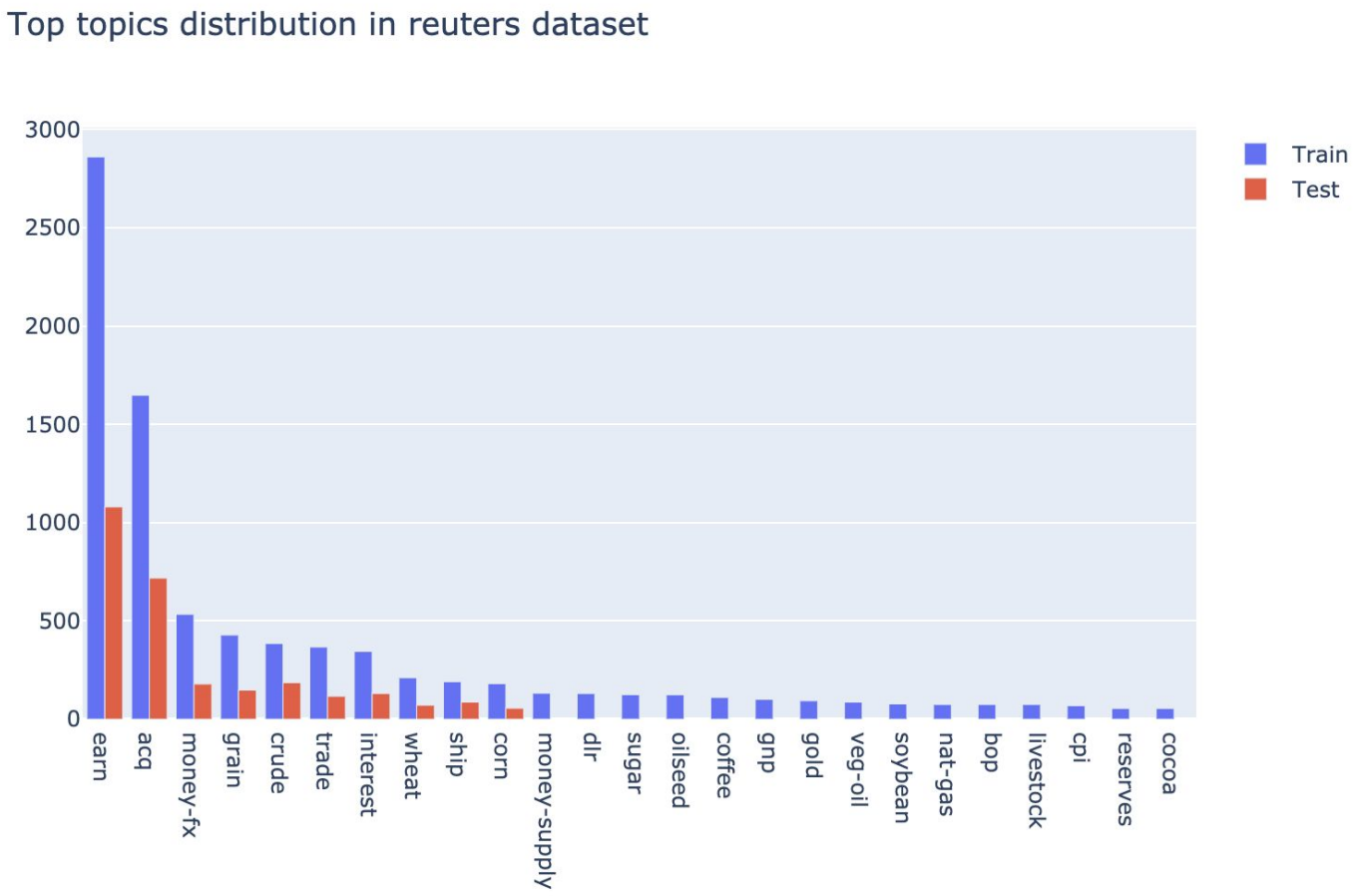
- **10000** news in 22 sgm data files
- **135** topics
- The **Modified Apte Split**
  - 9303 training set
  - 3299 testing set
  - Split by 1987/4/7

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" NEWID="1">
 <DATE>26-FEB-1987 15:01:01.79</DATE>
 <TOPICS> <D>cocoa</D> </TOPICS>
 <PLACES> <D>el-salvador</D> <D>usa</D> <D>uruguay</D> </PLACES>
 <PEOPLE></PEOPLE>
 <ORGS></ORGS>
 <EXCHANGES></EXCHANGES>
 <COMPANIES></COMPANIES>
 <UNKNOWN>C T f0704reute u f BC-BAHIA-COCOA-REVIEW 02-26 0105</UNKNOWN>
 <TEXT>
   <TITLE>BAHIA COCOA REVIEW</TITLE>
   <DATELINE>SALVADOR, Feb 26 -</DATELINE>
   <BODY>Showers continued throughout the week in the Bahia cocoa zone,
         alleviating   the drought since early January and improving prospects for
         …
         after carnival which ends midday on February 27. Reuter</BODY>
 </TEXT>
</REUTERS>
```

Data Preprocessing

- Parse SGM file
  - Extract topics
  - Extract news title and body
- Build **Training and Testing** data based on the **modified apte split**
  - Flatten topics

	id	topics	texts
0	4005	interest	u.s. economic data key to debt futures outlook...
1	4005	retail	u.s. economic data key to debt futures outlook...
2	4005	ipi	u.s. economic data key to debt futures outlook...
3	4012	earn	bank of british columbia 1st qtr jan 31 netope...
4	4014	earn	restaurant associates inc <ra> 4th qtr jan 3sh...


Top topics distribution in reuters dataset
3000
2500
2000
1500
1000
500
0
earn
acq
money-fx
grain
crude
trade
interest
wheat
ship
corn
money-supply
dlr
sugar
oilseed
coffee
gnp
gold
veg-oil
soybean
nat-gas
bop
livestock
cpi
reserves
cocoa
Train
Test

Problem Statement

- Multiclass classification
- Challenges include dealing with over 100 classes, imbalanced dataset, topics overlapping

Research

Text Categorization with Support Vector Machines [1]

Approaches	Data Preprocess	Criteria	Best Result
Tf-idf + SVM	- Remove stop words - Keep words over 3 occurrences - Choose best 500, 1000, 2000, 5000 features - ModApte split - One-vs-all classification	Precision / recall	86.4

Reference: [1] Text Categorization with Support Vector Machines

Research

- Previous research only makes classifier on topics with most frequency
- BERT shows capacity in imbalanced dataset, but for over 100 classes?
- A close look at the dataset

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Hierarchical Classifier

1. Multilabel Classification

2. Few Shot Learner

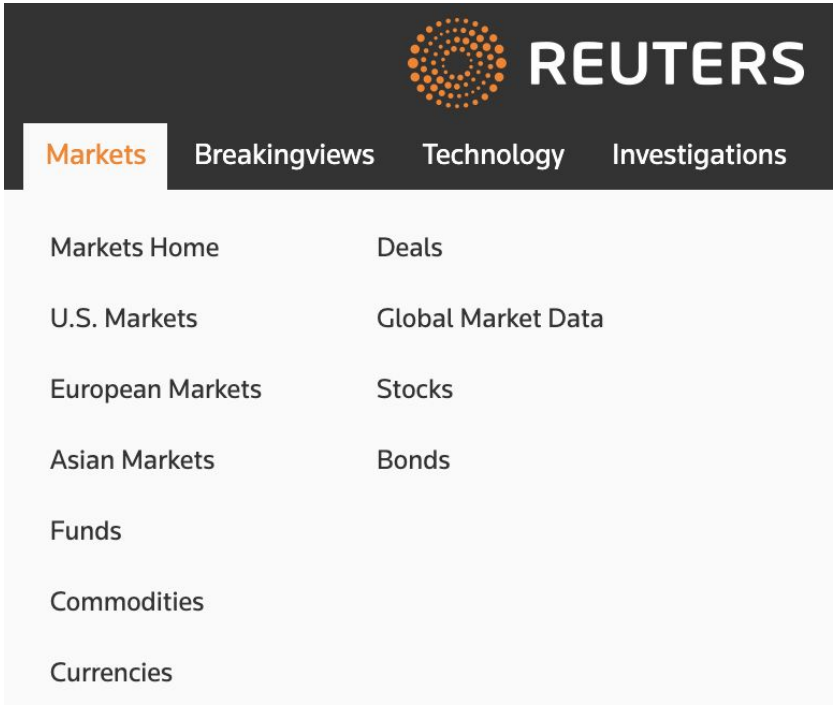
Hierarchical Classifier

```
****Subject Codes (135)

Money/Foreign Exchange (MONEY-FX)
Shipping (SHIP)
Interest Rates (INTEREST)

**Economic Indicator Codes (16)

Balance of Payments (BOP)
Trade (TRADE)
Consumer Price Index (CPI)
Wholesale Price Index (WPI)
Unemployment (JOBS)
Industrial Production Index (IPI)
Capacity Utilisation (CPU)
Gross National/Domestic Product (GNP)
Money Supply (MONEY-SUPPLY)
Reserves (RESERVES)
```



First Layer:

Multilabel Classifier

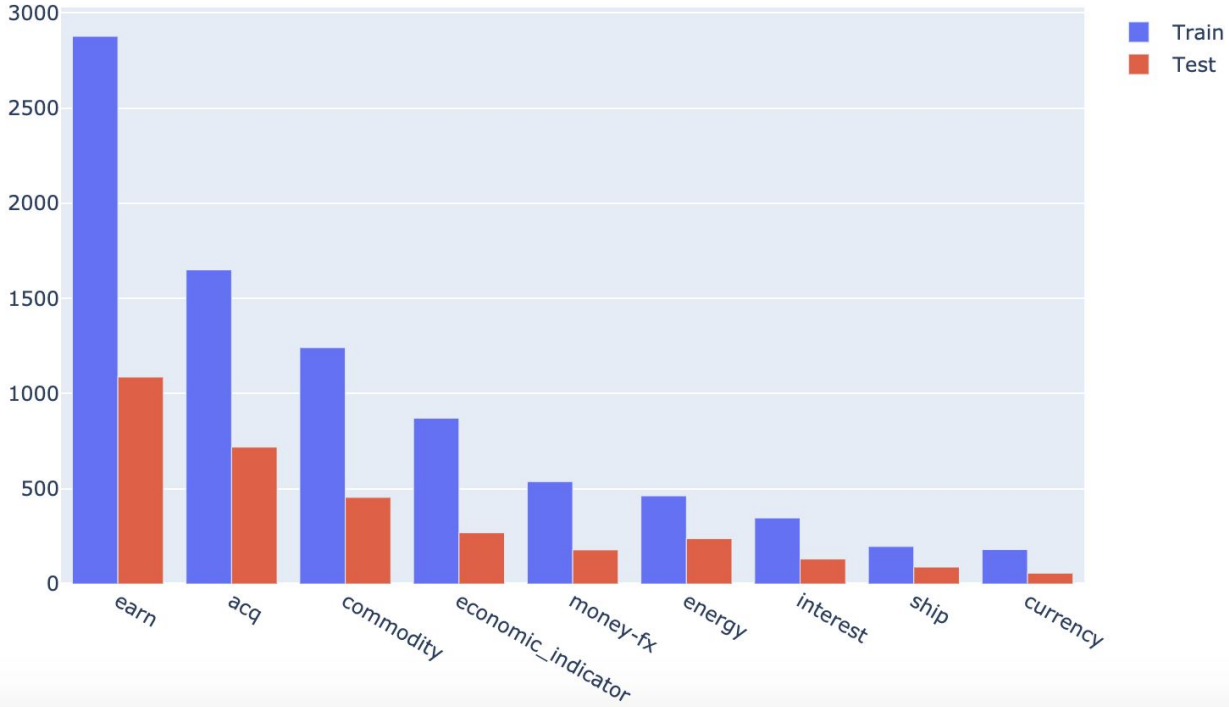
1. Data Preprocessing
2. Model Construction
3. Model Evaluation

Data Preprocess

- Generate mapping for topics and categories
- Add categories columns
- Generate labels
    - 7% of news have over 2 categories.

```
"money-fx": ["money-fx"],
"ship": ["ship"],
"interest": ["interest"],
"acq": ["acq"],
"earn": ["earn"],
"economic_indicator": ["bop", "trade", "cpi", "wpi",....],
"currency": ["dlr", "austdlr", "hk", "singdlr", ...],
"commodity": ["alum", "barley", "carcass", "castor-meal",... ],
 "energy": ["crude", "heat", "fuel", "gas", "nat-gas",...]}
```

Category Distribution in Training and Testing Dataset



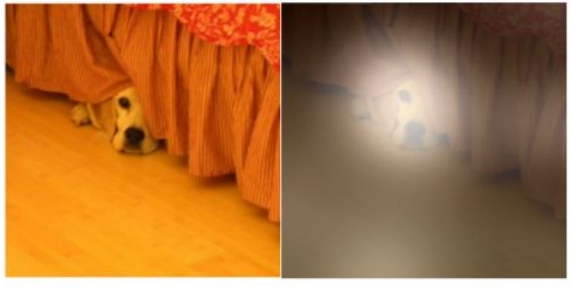
Data Preprocess

topics	texts
[interest, retail, ipi]	u.s. economic data key to debt futures outlook...
[earn]	bank of british columbia 1st qtr jan 31 netope...
[earn]	restaurant associates inc <ra> 4th qtr jan 3sh...

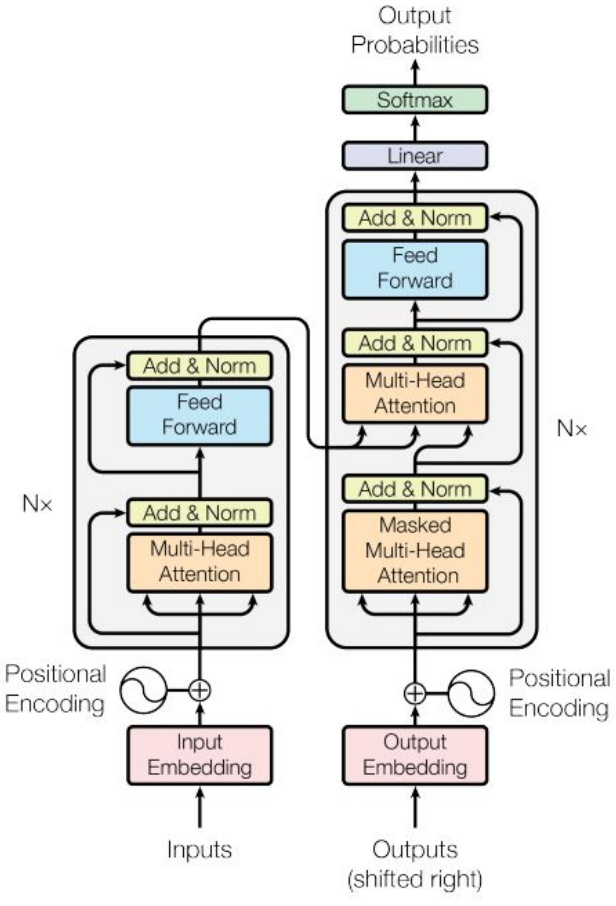
texts	categories	labels
key to debt futures outlook...	[interest, economic_indicator]	[0, 0, 1, 1, 0, 0, 0, 0, 0]
lumbia 1st qtr jan 31 netope...	[earn]	[0, 0, 0, 0, 0, 0, 0, 0, 1]
ates inc <ra> 4th qtr jan 3sh...	[earn]	[0, 0, 0, 0, 0, 0, 0, 0, 1]

Model Construction

- **Pretrained DistilBert** -- light, fast
- A **drop out** layer is added for Regularization
- A **linear layer** are added for Classification

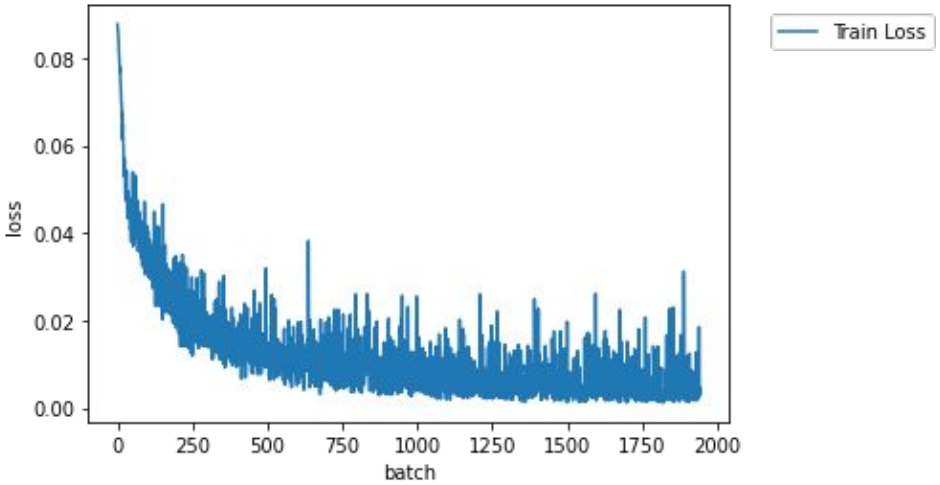


A dog is standing on a hardwood floor.



Multilabel Classification

- Loss function: Binary Cross Entropy
- Optimizer: Adam Optimizer
- Hyperparameter:
    - Max length: 200
    - Learning rate: 1e-05
    - Batch size: 8
    - EPOCH: 2



Model Evaluation

Metrics:

- Hamming Score: 0.94
- Hamming Loss: 0.013
- Accuracy Score = 0.91
- F1 Score (Micro) = 0.94
- F1 Score (Macro) = 0.89

```
>>> from sklearn.metrics import hamming_loss
>>> y_pred = [1, 2, 3, 4]
>>> y_true = [2, 2, 3, 4]
>>> hamming_loss(y_true, y_pred)
0.25
```

Second Layer:

Few Shot Learner

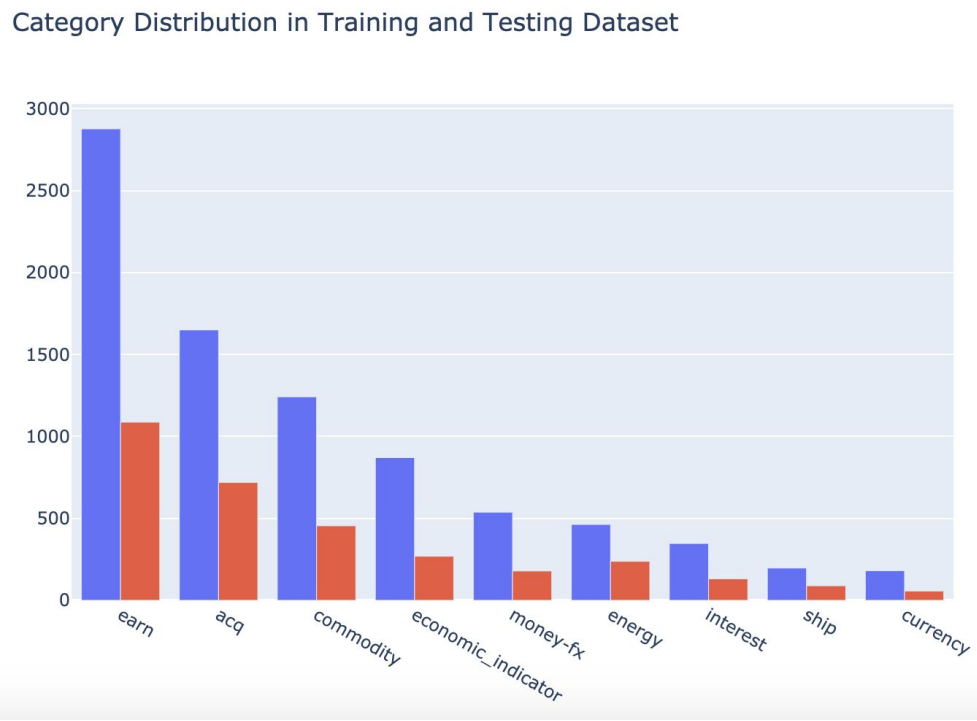
1. Few Shot Learner
2. Data Preprocess
3. Model Construction
4. Model Evaluation

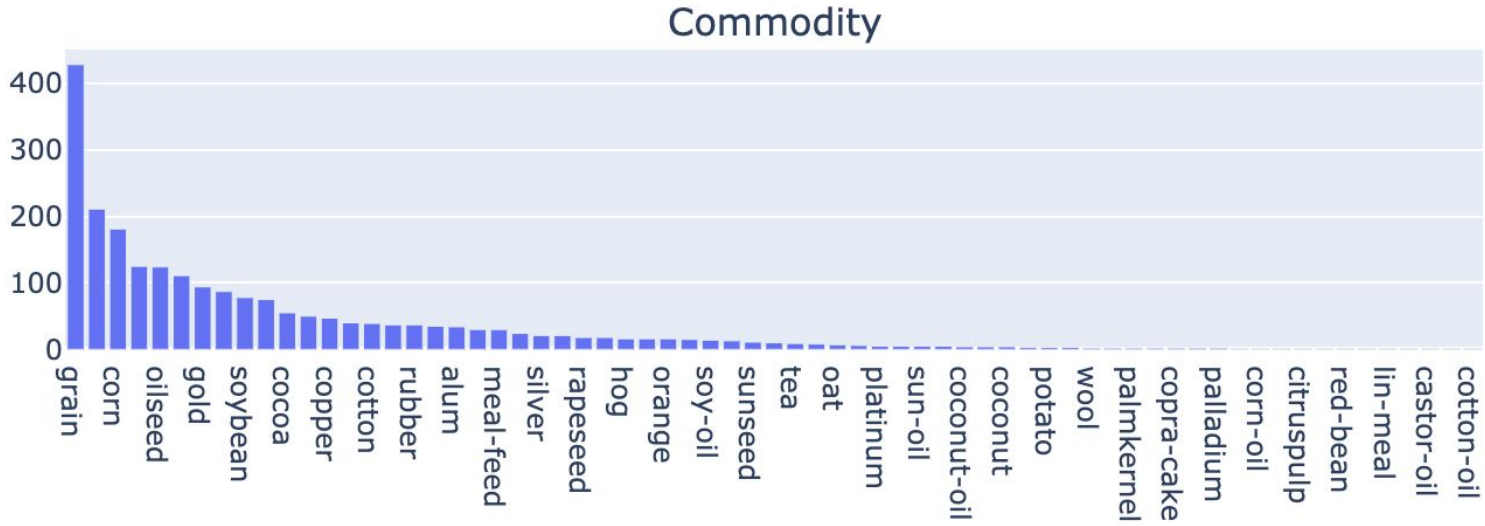
Dataset

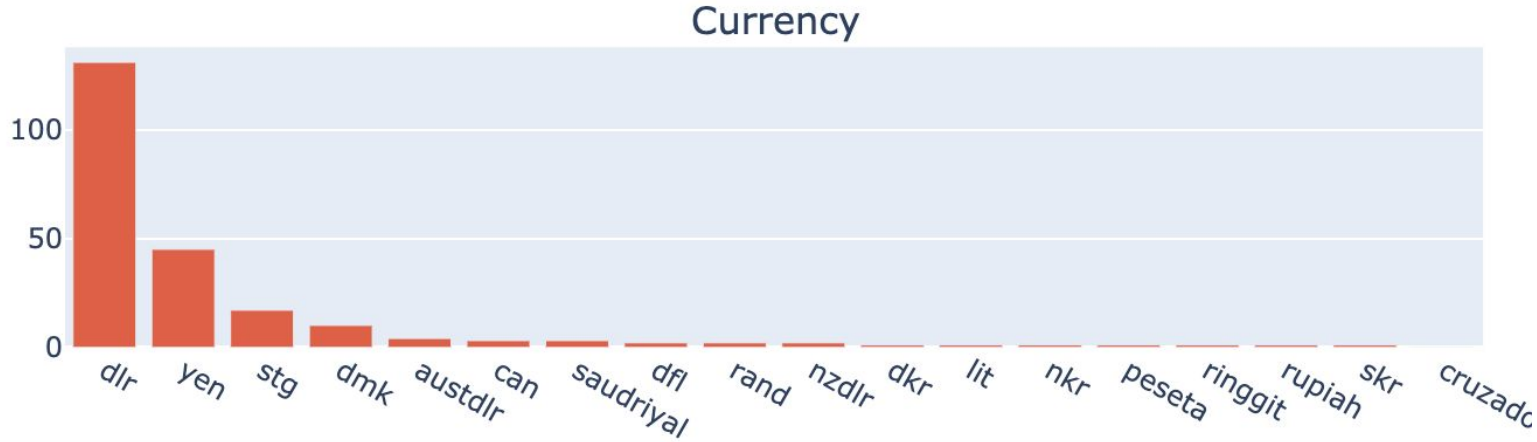
Main classes (5):
money-fx, ship, interest, acq, earn

Train:
The proportion of main classes is 0.58.
The proportion of minor classes is 0.42.

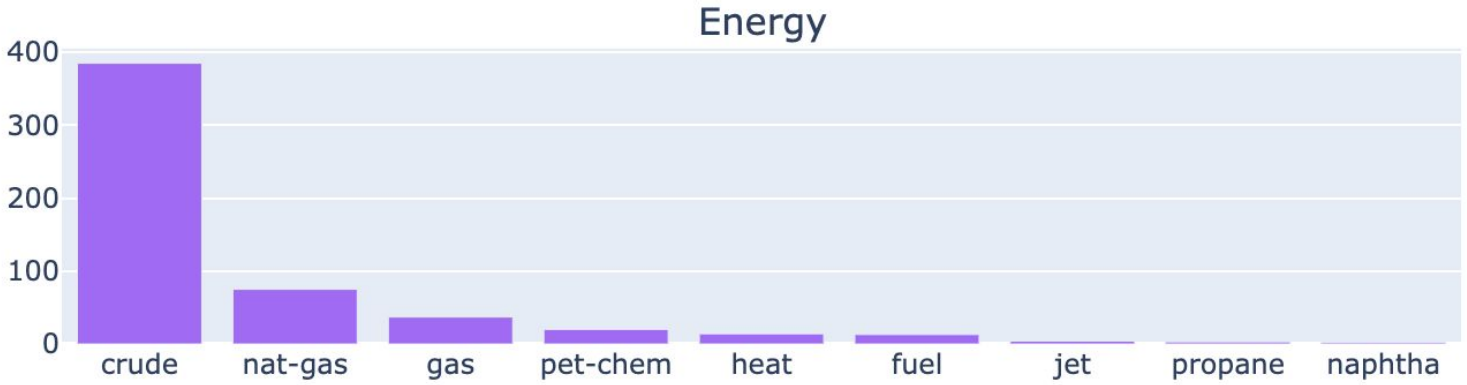
Test:
The proportion of main classes is 0.59.
The proportion of minor classes is 0.41.




Commodity
400
300
200
100
0
grain
corn
oilseed
gold
soybean
cocoa
copper
cotton
rubber
alum
meal-feed
silver
rapeseed
hog
orange
soy-oil
sunseed
tea
oat
platinum
sun-oil
coconut-oil
coconut
potato
wool
palmkernel
copra-cake
palladium
corn-oil
citruspulp
red-bean
lin-meal
castor-oil
cotton-oil

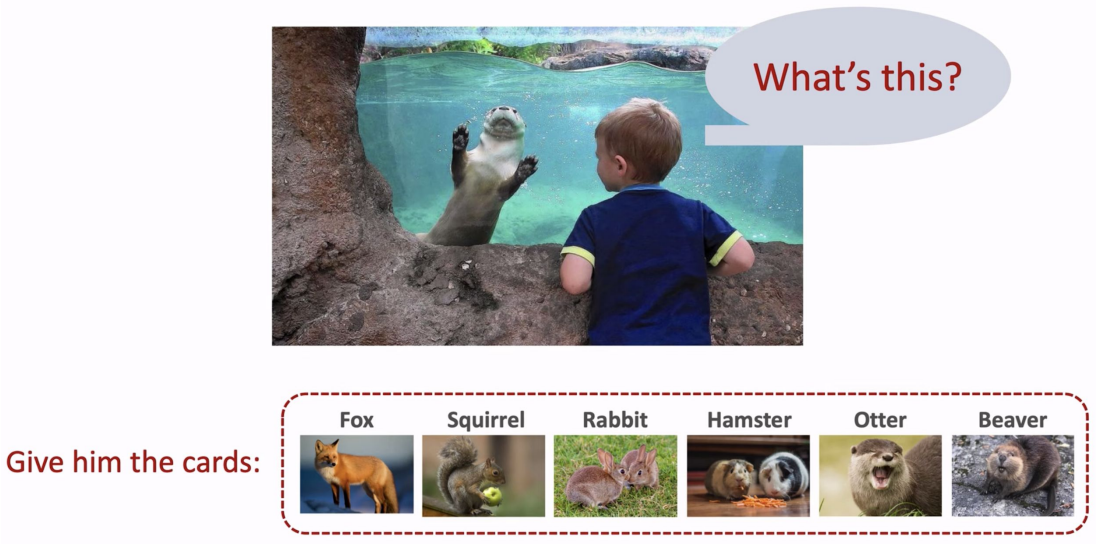

Currency
100
50
0
dlr
yen
stg
dmk
austdlr
can
saudriyal
dfl
rand
nzdlr
dkr
lit
nkr
peseta
ringgit
rupiah
skr
cruzado


Economic Indicator
300
200
100
0
trade
money-supply
gnp
bop
cpi
reserves
jobs
ipi
retail
wpi
housing
lei
income
instal-debt
inventories
cpu


Energy
400
300
200
100
0
crude
nat-gas
gas
pet-chem
heat
fuel
jet
propane
naphtha

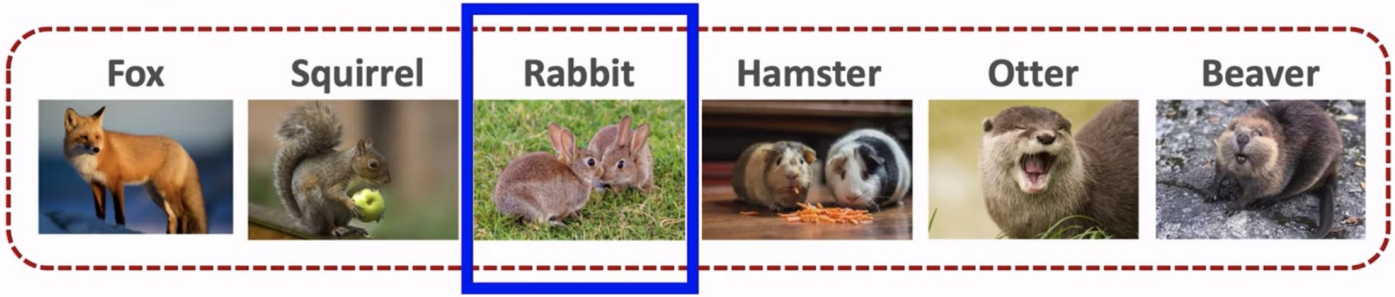
Few Shot Learner

- Classify new data having seen only a few training examples
- Learn to learn

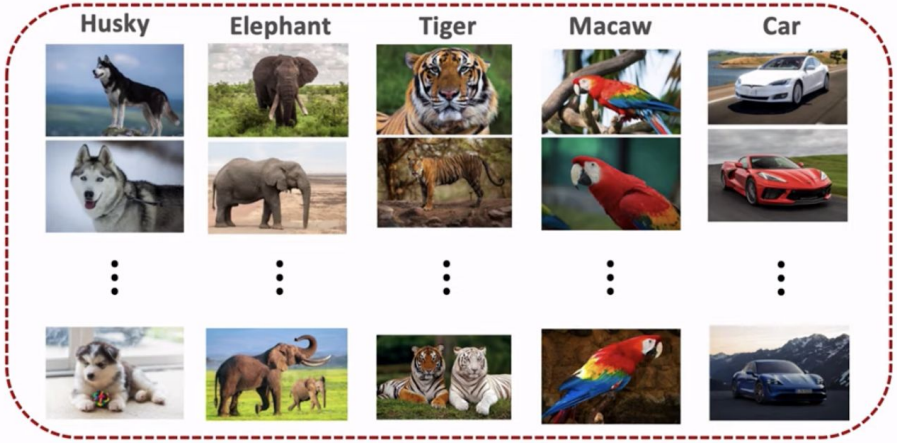


[Youtube: Few shot learner]

Support Set:



Training Set



Query Sample

Few Shot Learner

- Learning a **similarity** function

- Running two identical CNN on two different inputs and then comparing them

Few Shot Learner

Step 1: Learn a similarity function from large-scale training sample

Step 2: Apply the similarity function for prediction

- Compare the query with every sample in the support set
- Find the sample with the highest similarity score

Loss function - triplet loss

$\mathbf{x}^{+}$
(positive)

$\xrightarrow{\mathbf{f}}$

$\mathbf{f}(\mathbf{x}^{+})$

$\mathbf{x}^{a}$
(anchor)

$\xrightarrow{\mathbf{f}}$

$\mathbf{f}(\mathbf{x}^{a})$

$\mathbf{x}^{-}$
(negative)

$\xrightarrow{\mathbf{f}}$

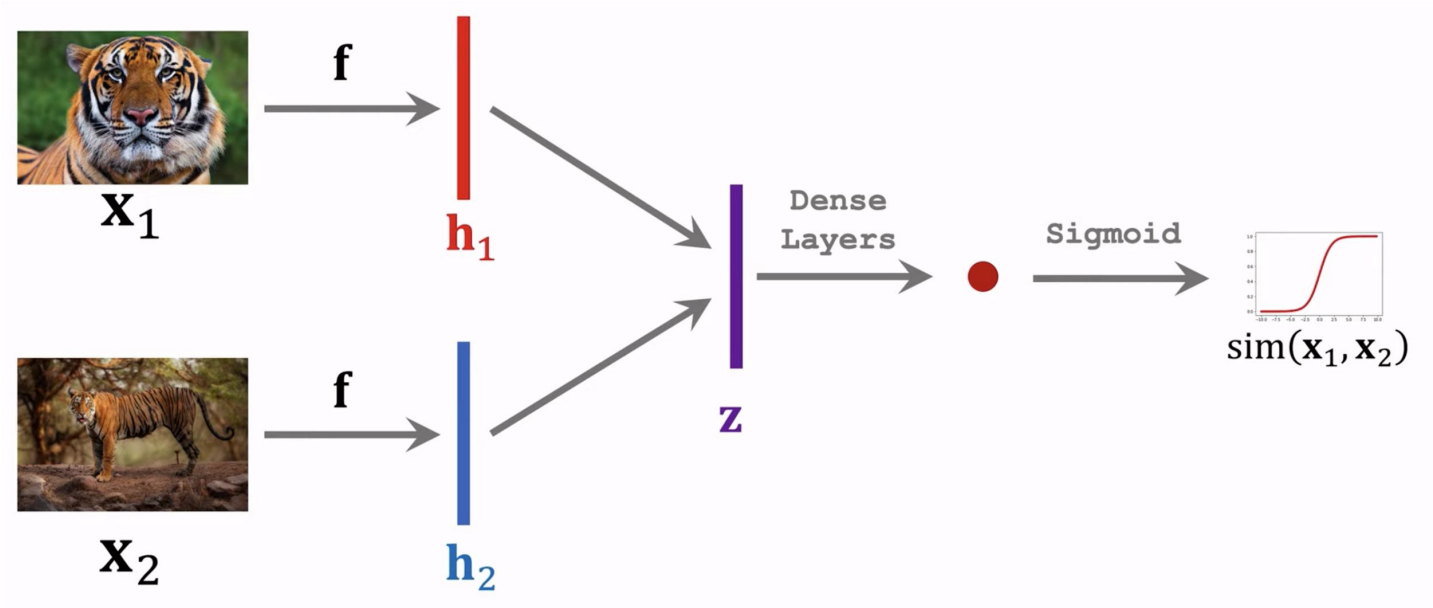
$\mathbf{f}(\mathbf{x}^{-})$

$$d^{+} = \left|\left|\mathbf{f}(\mathbf{x}^{+}) - \mathbf{f}(\mathbf{x}^{a})\right|\right|_2^2$$

$$d^{-} = \left|\left|\mathbf{f}(\mathbf{x}^{a}) - \mathbf{f}(\mathbf{x}^{-})\right|\right|_2^2$$

Loss function - siamese network

Loss function - siamese network

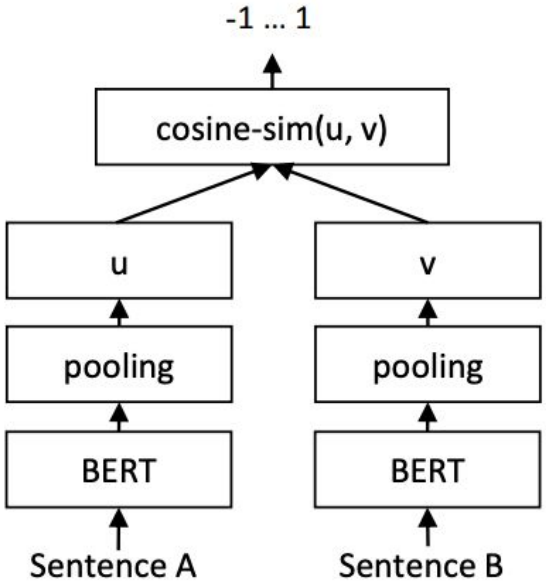


Data Preprocess

- For each category
  - Training set: a pair of data contains anchor, positive and negative
  - Support set: N way 1 shot
  - Test set: all data within the same category

Model Construction

- **Siamese Neural Network**
- Loss function: Triplet loss
- Optimizer: Adam Optimizer
- Hyperparameter:
  - Max length: 200
  - Learning rate: 1e-05
  - Batch size: 8
  - EPOCH: 2



[Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks]

Model Evaluation

Category	N way one shot (N)	Accuracy
commodity	65	35%
currency	18	43%

Further Study

1. Few shot fine-tuning

2. Top modeling

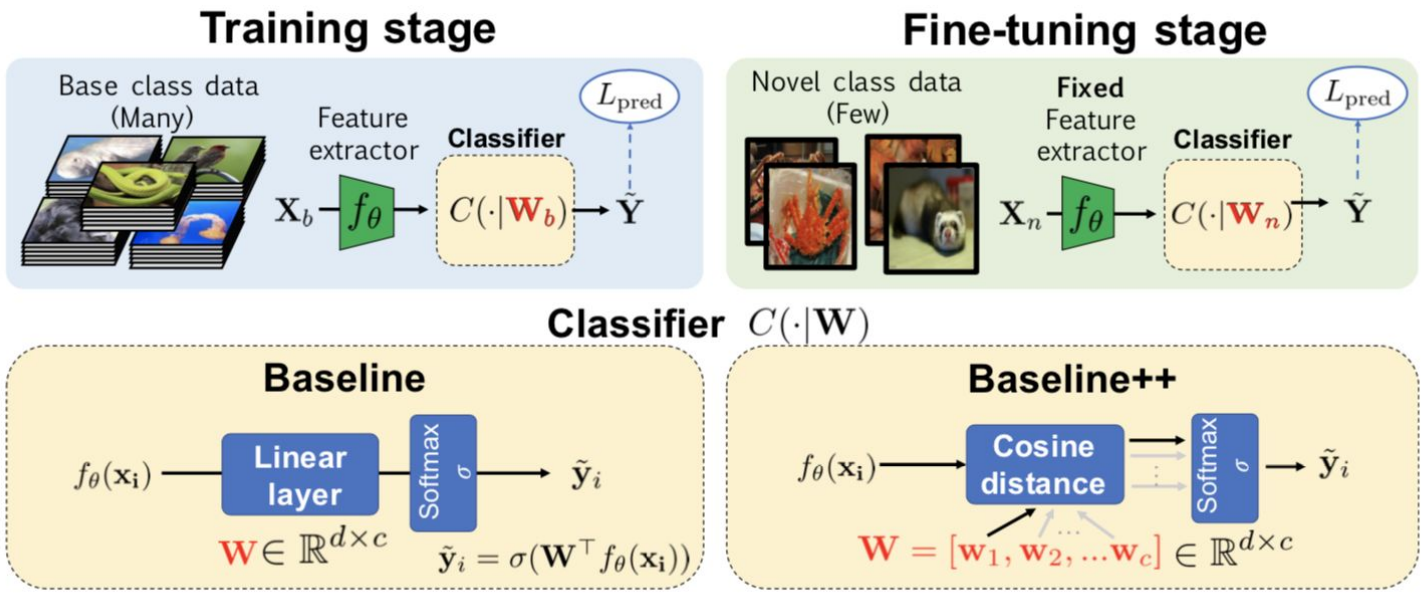


Figure 1: **Baseline and Baseline++ few-shot classification methods.** Both the baseline and baseline++ method train a feature extractor f_θ and classifier $C(.|\mathbf{W}_b)$ with base class data in the training stage In the fine-tuning stage, we fix the network parameters θ in the feature extractor f_θ and train a new classifier $C(.|\mathbf{W}_n)$ with the given labeled examples in novel classes. The baseline++ method differs from the baseline model in the use of cosine distances between the input feature and the weight vector for each class that aims to reduce intra-class variations.

A Closer Look at Few-shot Classification

Timeline and github

Time	Implement	Research
12/21/2020 - 12/24/2020	Parse File	Weak Supervision
12/25/2020 - 12/28/2020	TF-IDF Embedding	Few Shot Learner
12/29/2020 - 1/1/2021	Multilabel Classification	Hugging Face
1/2/2021 -1/6/2021	Few Shot Learner	Few Shot Learner
1/7/2021 - 1/11/2020	Refine Model and Presentation	

GitHub: https://github.com/yinhao0424/reuters/blob/master/README.md

Thank You