Machine Learning in Real Estate

Khoa Tran
Niantong Dong
Bryan Jaimes
Amanda Justiniano-Pagan

Agenda

- Paper evaluation
- BERT Model
- Data cleaning
- API
- Next sprint Goal

Our fundamental

EXTRACT USEFUL INFORMATION FROM BUILDING PERMITS DATA TO PROFILE A CITY'S BUILDING RETROFIT HISTORY

Wanni Zhang[1], Tianzhen Hong[1], and Xuan Luo[1]
Lawrence Berkeley National Laoratory, Berkeley, CA

About this paper

- Using Machine learning to extract useful information from the building permit datasets over the past several decades.
    - Tokenize written natural language as input to the model
    - Compared CNN and BERT performance
- Individual building analysis such as time intervals between permits for each building.

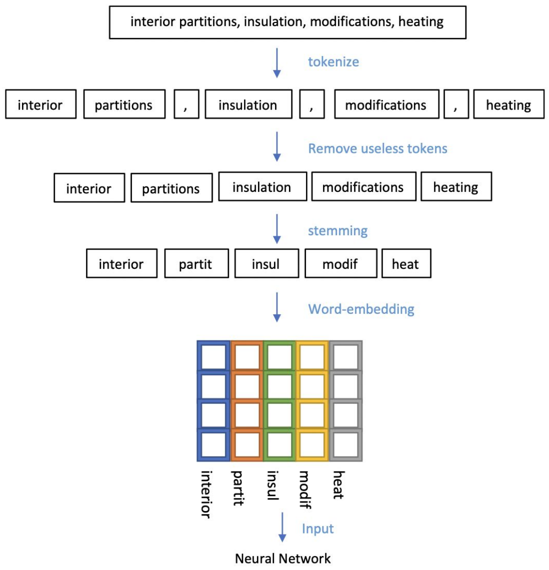


Figure 2 Diagram of data preprocessing and word-embedding process

Why starts with this paper ?

- This paper has a similar purpose with what we are doing now, to train a model to classifier different kind of permits.
- This paper gives us a solid reason to go with BERT model because it is better accuracy compared to CNN.

Table 1 Classification accuracy for the type of work of different models

	BUILDING	**ELECTRICAL**	**MECHANICAL**	**PLUMBING**
val_acc_cnn	0.9130	0.8916	0.8744	0.9214
val_acc_bert	0.9231	0.9045	0.9104	0.9490
default val_acc	0.8493	0.7834	0.7489	0.6274
test_acc_cnn	0.6886	0.7057	0.7229	0.8371
test_acc_bert	0.7875	0.7125	0.7475	0.8406
default test_acc	0.5657	0.6257	0.5371	0.6400

Data Cleanup

The Dataset: Boston Construction Permits

- Publicly available dataset of construction permit data from Boston.

We made several modifications to the data before feeding it into the BERT model:

1. Dropping columns that are not be necessary for training our model (such as permit applicant name, address, monetary fees)
2. Verifying there are no duplicate records in the dataset
3. Drop any records with missing data

X Dataset

Grab X data

In [29]:
```
X = df.drop(columns=['permittypedescr'])
X.head()
```

Out[29]:

	permitnumber	worktype	description	comments
0	A1000569	INTEXT	Interior/Exterior Work	This work is to Amend Permit ALT347244. Elimin...
1	A100071	COB	City of Boston	Change connector link layout from attached enc...
2	A1001012	OTHER	Other	Amend Alt943748 to erect a roof deck as per pl...
3	A1001201	INTEXT	Interior/Exterior Work	Build steel balcony over garden level with sta...
4	A100137	EXTREN	Renovations - Exterior	Landscaping/stonework - amending permit #2801/...

y Dataset

Grab y data (Labels)

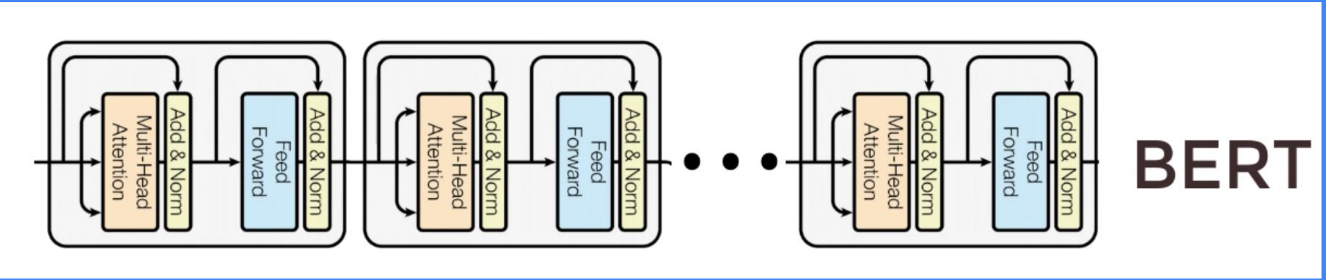
In [31]:
```
y = df.permittypedescr
y.head()
```

Out[31]:
```
0    Amendment to a Long Form
1    Amendment to a Long Form
2    Amendment to a Long Form
3    Amendment to a Long Form
4    Amendment to a Long Form
Name: permittypedescr, dtype: object
```

Permit type

Fine-tuning BERT

BERT



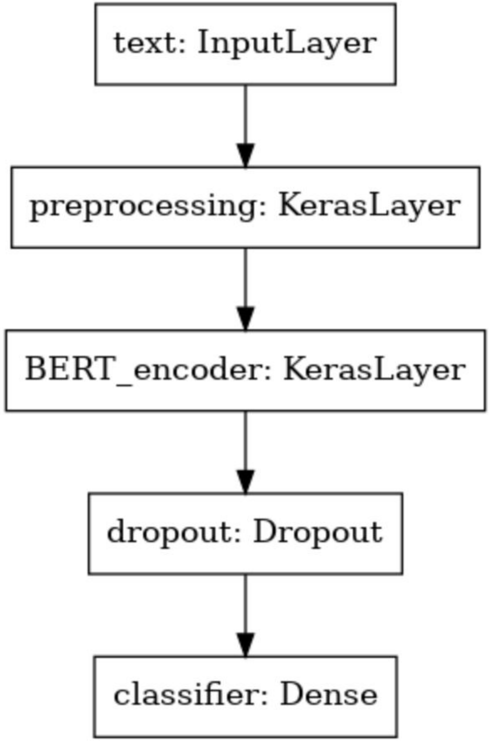
- Bidirectional Encoder Representation from Transformers
- Developed by Google in 2018 to translate languages. Now used as general natural language processing technique
- Faster to train and deeply understands language compared to previous state of the art model (LSTMs)
- Two steps to use BERT:
    - Pre training for language and context (on Wikipedia data)
    - Fine-tuning for our specific use case (text classification)

Fine-tuning BERT

- Used small BERT version (fewer and smaller transformer blocks)
- Used Boston housing permit dataset. Columns used for training included permit number, work type, description, comments -> 14 total output categories.
- Data split 80/20 for training and testing, respective. Datapoints capped at 10k for each category.
- Hyperparameters:
  - Loss function: sparse cross entropy
  - **Optimizer: adaptive moments (AdamW)**
  - **Initial learning rate : 0.00003**
  - **Learning rate: linear decay with first 10% as warm up**
  - Training epochs: 5

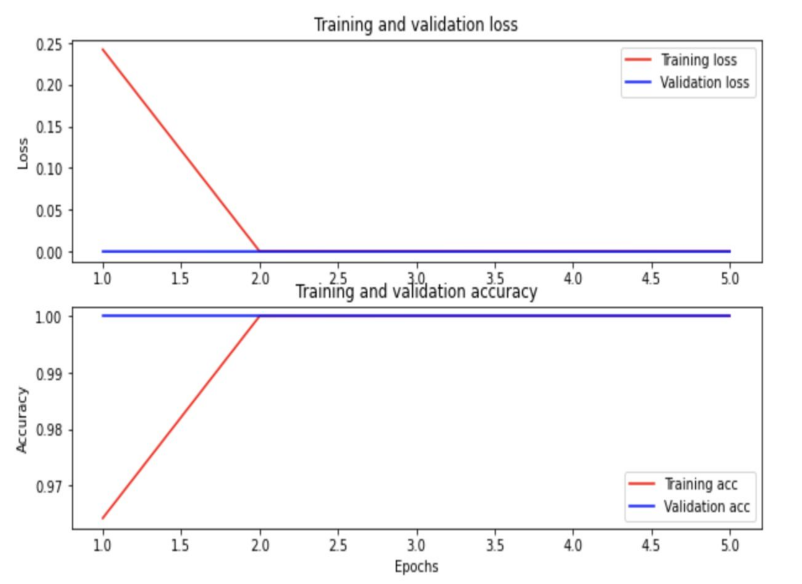
Fine-tuning BERT

1. Input layer to take in our specific inputs
2. BERT preprocessor to process inputs
3. BERT pretrained model
4. Dropout layer
5. Dense output layer as classifier (14 categories)



Results and next steps

- Achieved 100% top 5 categorical accuracy
- Model recognized IDs from dataset, likely biased toward Boston dataset
- Next steps:
  - Fine tune model hyperparameters
  - Retrain on more generalized data from different cities

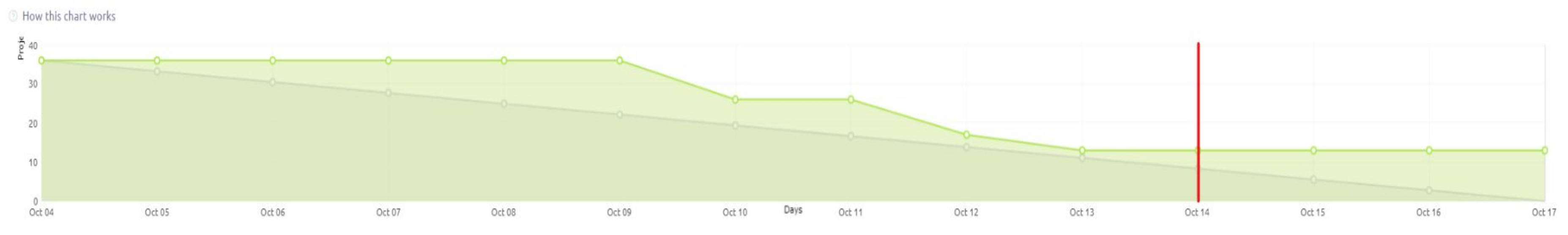


Sample prediction

```
examples = [
    'A1000569,INTEXT,Interior/Exterior Work,This work is to Amend Permit ALT347244. Eliminate construction of two party wall o
penings. Install new wheelchair lift.;;;;; E- Plans',  # this is the same sentence tried earlier
    'G421944,GAS,Gas,Replace leaking gas water heater.',
    'PL917318,PLUMBING,Plumbing,Installation of 2 Bathrooms and 1 Kitchen',
    'ELV765406,LVOLT,Low Voltage,installation of low voltage wireless burglar alarm system',
    'E1128739,ELECTRICAL,Electrical,Apt 54 on 5th floorWire bedroom/living room/bathroom and kitchen Combination of old work a
nd new work Install new sub panel in unit'
]
```

```
Results:
               Amendment_to_a_Long_Form
               Gas_Permit
               Plumbing_Permit
               Electrical_Low_Voltage
               Electrical_Permit
```

Sprint 2: Burndown Chart



<u>Sprint ends Oct 17th</u>

36 points scoped for the sprint

13 points left

Sprint 3 Goals

2021_10_2

18 Oct 2021-31 Oct 2021

0 closed
36 total

#27	Create DevOps/CI Framework in Github	16
#30	Tune BERT model parameter	4
#28	Formalize API Implementation	8
#29	Formalize and deploy ML microservice	8

Second point

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua

Incididunt ut labore et dolore

Consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua

XX%

Use this slide to show a major stat. It can help enforce the presentation's main message or argument.

Final point

A one-line description of it

This is the most important takeaway that everyone has to remember.

Thanks!

Contact us:

Your Company
123 Your Street
Your City, ST 12345

no_reply@example.com
www.example.com