Classifying Multimodal Data Using Transformers

Watson Chua, Lu Li, Alvina Goh, Amelia Lee
Data Science and Artificial Intelligence Division
Government Technology Agency of Singapore

Background

- GovTech successfully built a transformers-based classifier which can automatically predict the government agencies to handle feedback on municipal issues from residents, using the following:
  - Text
  - Geolocation
  - Images

- With this experience, the team is conducting a hands-on tutorial to flatten the learning curve for data scientists and machine learning engineers who want to apply machine learning on multimodal data.

Objective

- At the end of the tutorial, you will be able to use PyTorch and Hugging Face Transformers to build:
  - A text classifier using BERT
  - A text-image classifier with dual encoders using BERT and ResNet
  - A text-image classifier with joint encoders using ALBEF

- We will be using a selected subset of the Webvision1.0 dataset for the tutorial.

Tutorial Outline

No.	Item	Duration
1	Sharing on the Municipal Issues Feedback Classifier	25 mins
2	Building a Text Classifier with BERT	30 mins
3	Break	15 mins
4	Building a Dual-Encoder Text-Image Classifier with ResNet and BERT	20 mins
5	Building a Joint-Encoder Text-Image Classifier with ALBEF	25 mins
6	Break	15 mins
7	Wrap-up/ Q&A	15 mins

OFFICIAL (OPEN)

Sharing on the Municipal Issues Feedback Classifier

Municipal Services Office (MSO)

- Works with government agencies and town councils to improve feedback management and service delivery for municipal services

Building and Construction Authority
HOUSING & DEVELOPMENT BOARD
Land Transport Authority
We Keep Your World Moving
National Environment Agency
Safeguard · Nurture · Cherish
NATIONAL PARKS
LET'S MAKE SINGAPORE OUR CITY IN NATURE
People's Association
PUB SINGAPORE'S NATIONAL WATER AGENCY
SLA SINGAPORE LAND AUTHORITY
SINGAPORE POLICE FORCE
SAFEGUARDING EVERY DAY
URBAN REDEVELOPMENT AUTHORITY
ONESERVICE
MUNICIPAL SERVICES OFFICE
ALJUNIED HOUGANG TOWN COUNCIL
Ang Mo Kio Town Council
Bishan-Toa Payoh Town Council
Chua Chu Kang TOWN COUNCIL
EAST COAST TOWN COUNCIL
HOLLAND-BUKIT PANJANG TOWN COUNCIL
Jalan Besar TOWN COUNCIL
Jurong-Clementi Town Council
Marine Parade Town Council
Marsiling-Yew Tee Town Council
NEE SOON TOWN COUNCIL
PASIR RIS-PUNGGOL TOWN COUNCIL
Sembawang Town Council
SENGKANG TOWN COUNCIL
BRINGING THE FUTURE CLOSER
TAMPINES TOWN COUNCIL
TANJONG PAGAR TOWN COUNCIL
WEST COAST TOWN COUNCIL

OFFICIAL (OPEN)

Municipal Issues Feedback

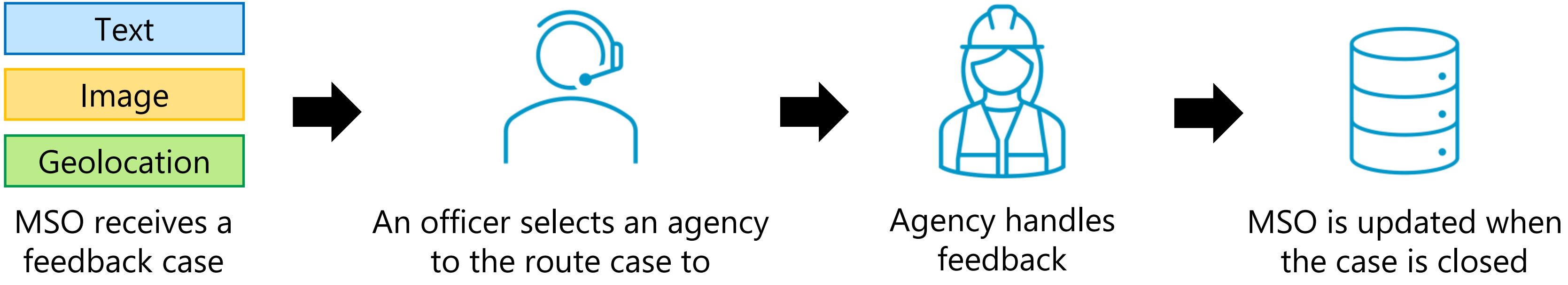
Pigeons at
hawker centres

Overgrown greenery

Fallen lamppost

Feedback is received from 3 channels:

Manual Case Routing Process



Text

Image

Geolocation

MSO receives a feedback case

An officer selects an agency to the route case to

Agency handles feedback

MSO is updated when the case is closed

Case ID	Description	Images	X-Coord	Y-Coord	Agency
1	…	…	…	…	LTA
2	…	…	…	…	NParks
3	…	…	…	…	NEA
…	…	…	…	…	…

417k cases handled each year!

Predicting Agency to Handle Feedback

New Feedback

The recently patched road has a new pothole appearing. Very obvious.

Text

Image

1.3404329
103.6901702

Geolocation

App or Chatbot

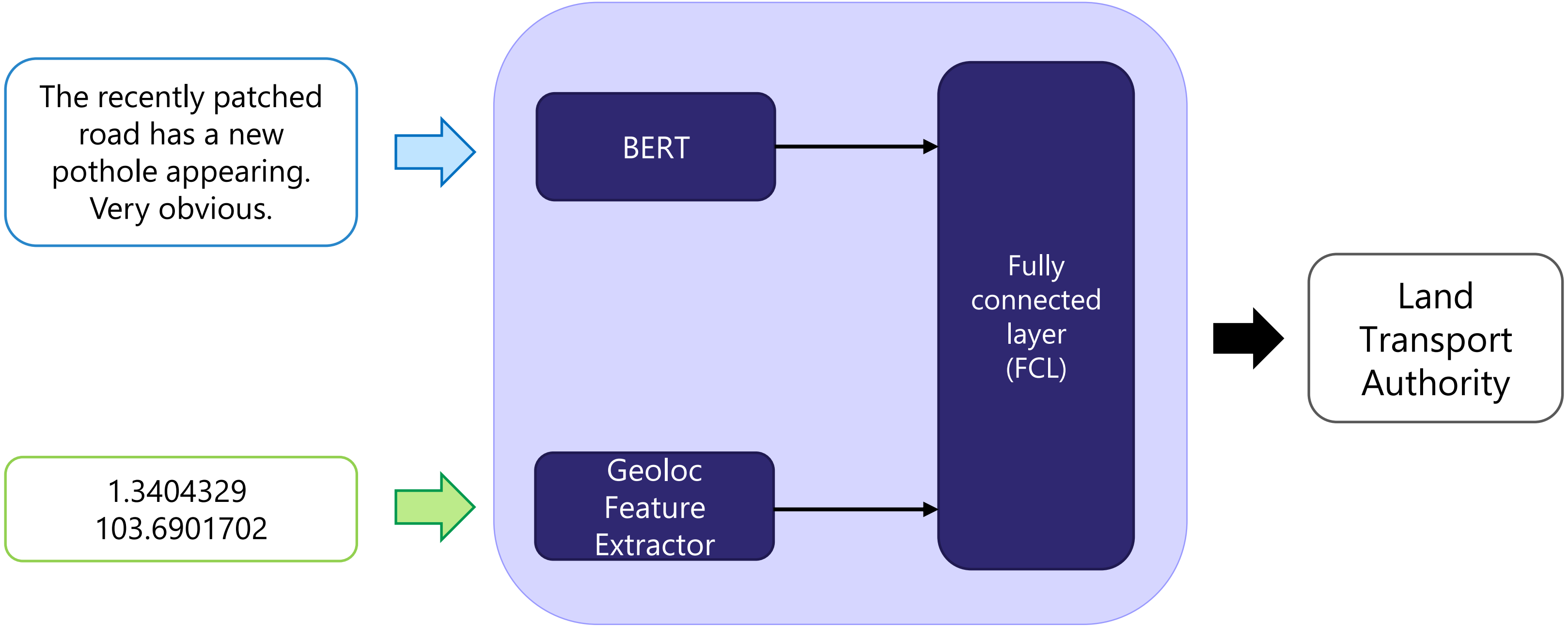
Classifier trained on historical data from CRM

Agency Classifier

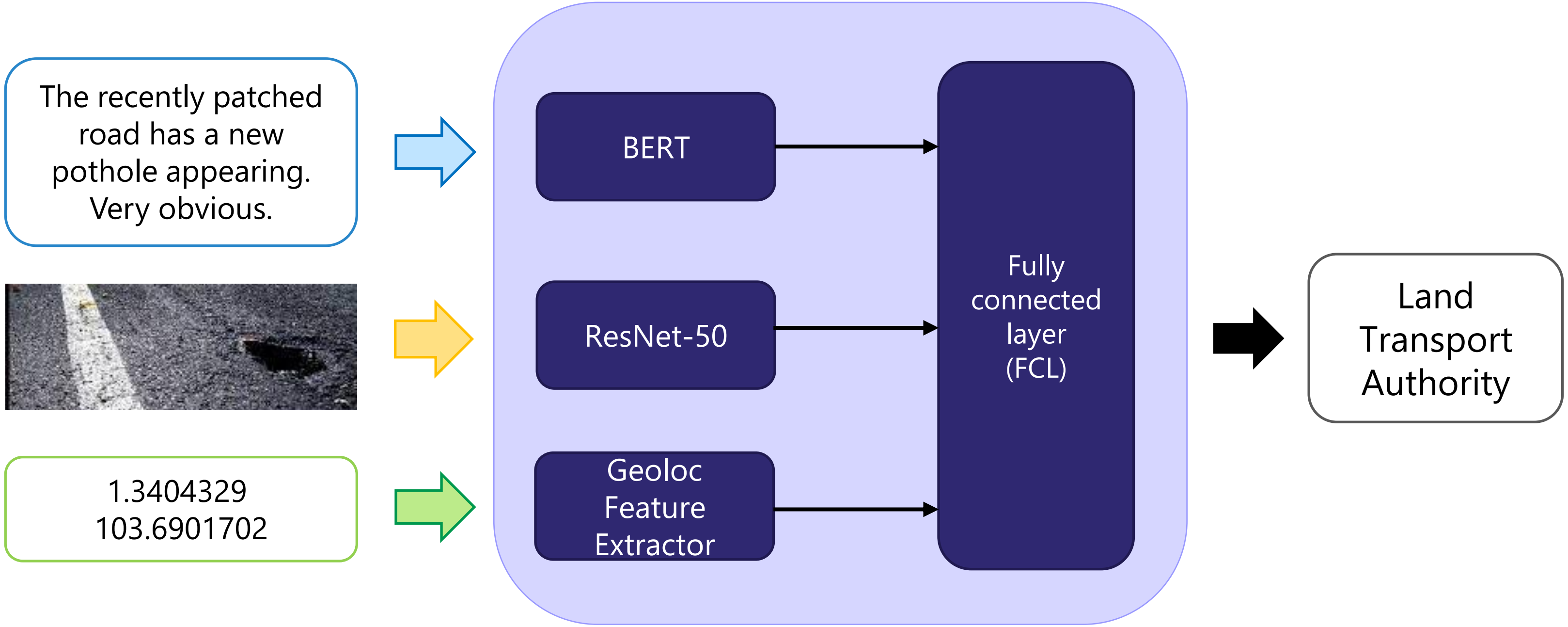
Prediction

Land Transport Authority

Training the Classifier: Text Encoder + Geolocs Architecture


The recently patched road has a new pothole appearing. Very obvious.
1.3404329
103.6901702
BERT
Geoloc Feature Extractor
Fully connected layer (FCL)
Land Transport Authority

Training the Classifier: Dual-Encoder Text-Image + Geolocs Architecture



The recently patched road has a new pothole appearing. Very obvious.

1.3404329
103.6901702

BERT

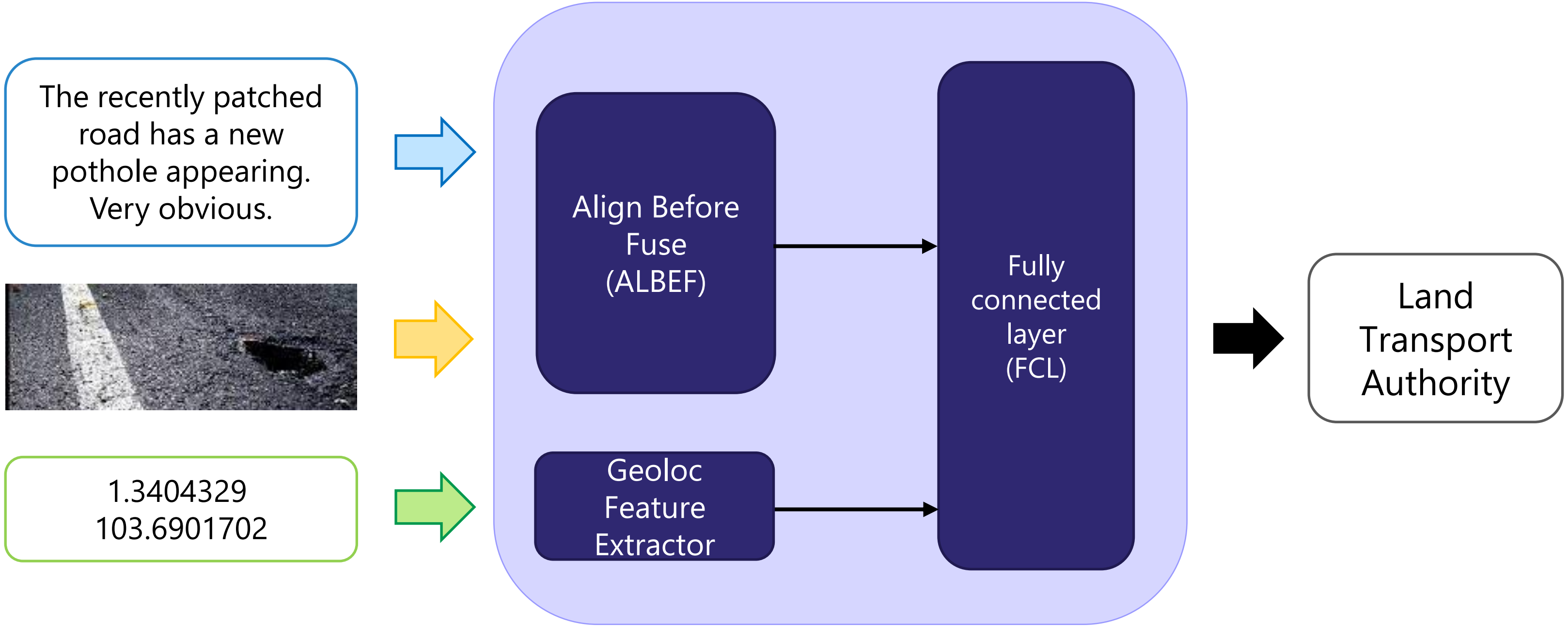
ResNet-50

Geoloc Feature Extractor

Fully connected layer (FCL)

Land Transport Authority

Training the Classifier:
Joint-Encoder Text-Image + Geolocs Architecture



The recently patched road has a new pothole appearing. Very obvious.

1.3404329
103.6901702

Align Before Fuse (ALBEF)

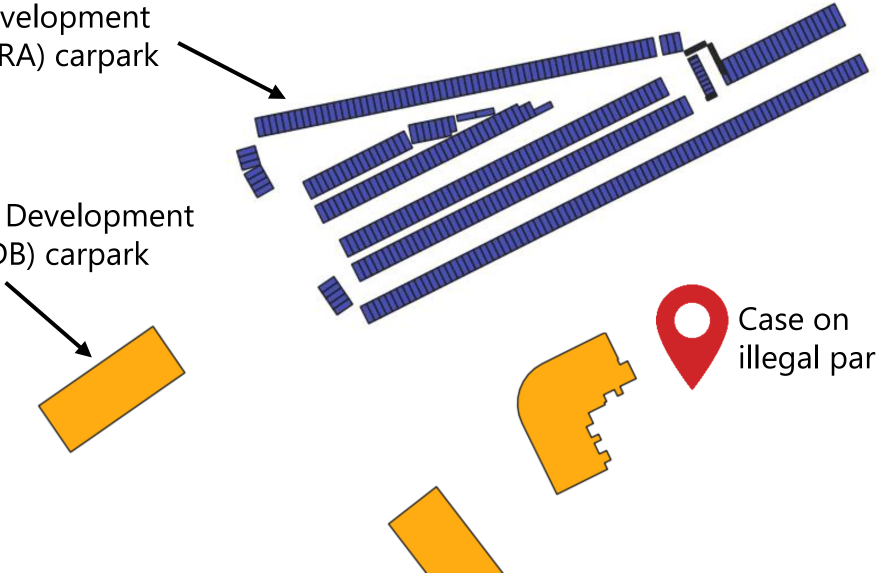
Geoloc Feature Extractor

Fully connected layer (FCL)

Land Transport Authority

Geolocation Features

- Geolocation data provides valuable information in a small subset of cases where multiple agencies handle similar cases in the same area



Urban Redevelopment Authority (URA) carpark

Housing and Development Board (HDB) carpark

Case on illegal parking

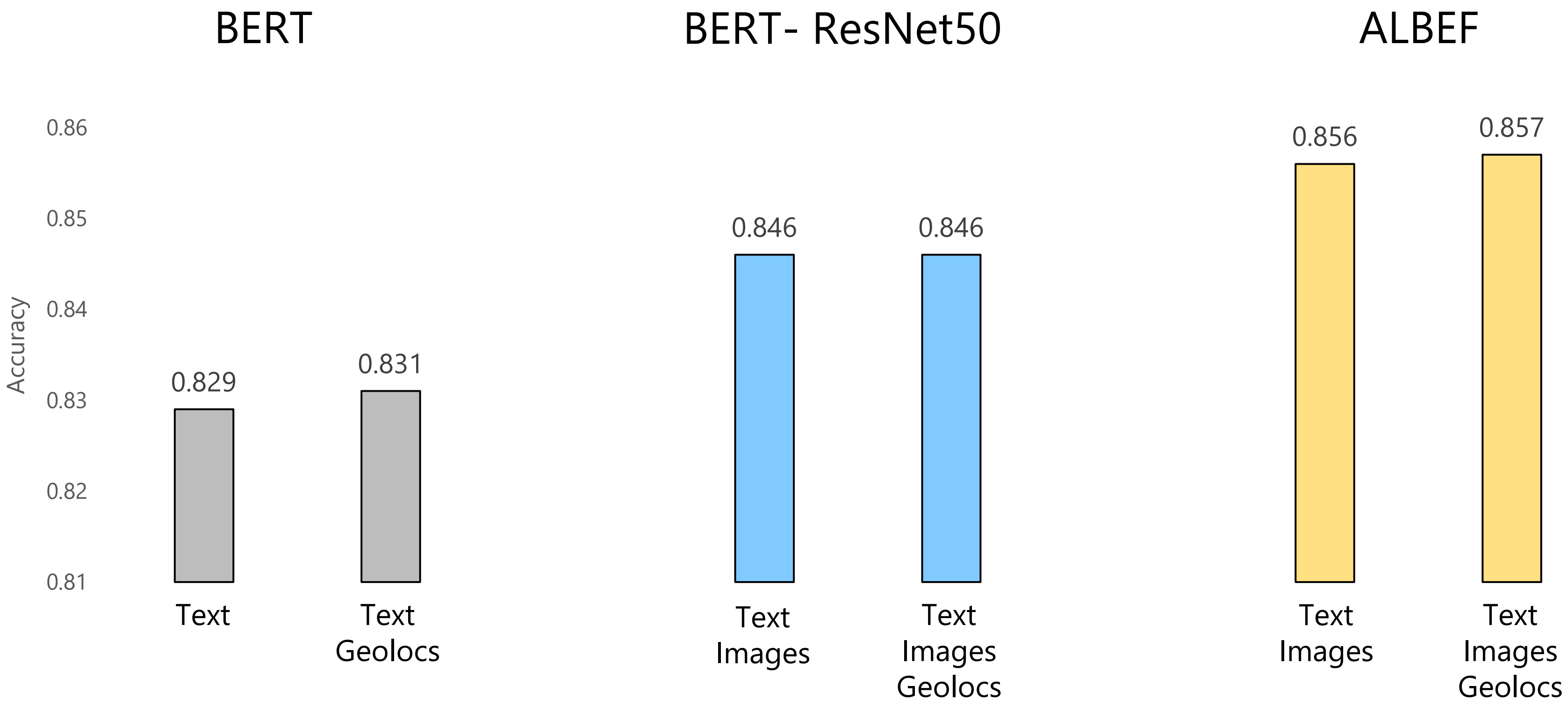
Both agencies own and maintain carparks in the same area.

Since the incident location is nearer to the orange areas, the case is likely under the purview of HDB.

Model Accuracy Comparison

Data used: 30k cases
Num classes: 13
Train-Test Split: 80%-20%

Randomly selected subset from the most recent data with images



Hands-on Session

Objective (Recap)

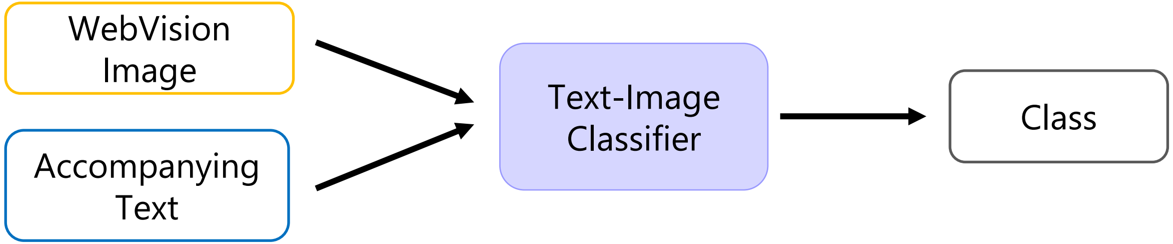
- At the end of the tutorial, you will be able to use PyTorch and Hugging Face Transformers to build:
  - A text classifier using BERT
  - A text-image classifier with dual encoders using BERT and ResNet
  - A text-image classifier with joint encoders using ALBEF

- We exclude the geolocation portion from the tutorial because it is too specific to our use case

Dataset - WebVision

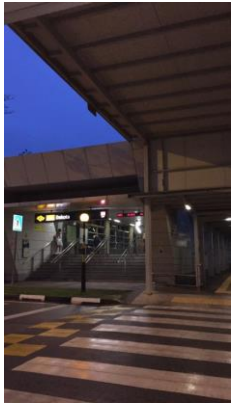
- The municipal issues dataset cannot be shared as it contains sensitive information
- WebVision dataset shares similar characteristics

- Text-image pairs are tagged with a class
- Images are crawled from the Flickr website and Google Images search
- Texts are captions, user tags or descriptions



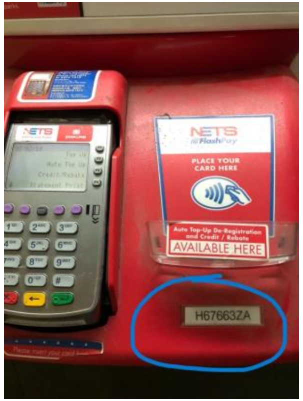
- We selected 1,000 cases from 10 classes for the hands-on session, with a 80%-20% train-test split

Sample Text-Image Pairs (Municipal Issue Dataset)



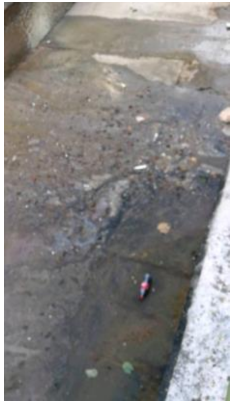
Text: The covered walkway outside Dakota Station Exit B has holes in it and leaks water on your head, please fix it.

Class: LTA



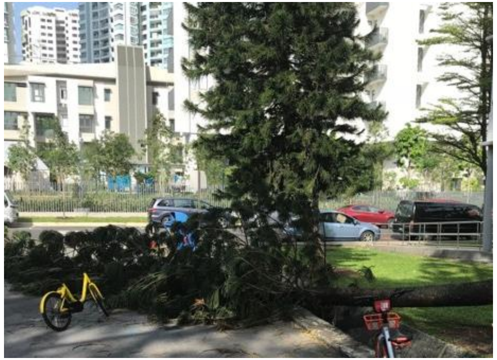
Text: The receipt is jammed up and I could not retrieve a copy of the statement. Can someone clear up the jammed paper?

Class: HDB



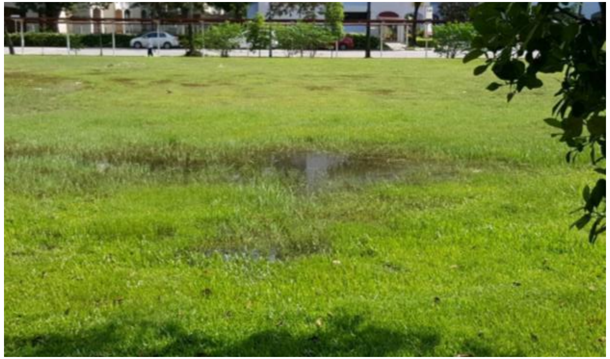
Text: Long time never clean with dead bird and oil stain.

Class: PUB



Text: Fallen tree next to rubbish collection center.

Class: NParks



Text: The ponding has been around for few months. It could be a potential mosquito breeding ground.

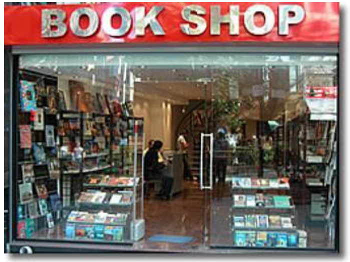
Class: NEA

Sample Text-Image Pairs (WebVision Dataset)

Text: Looks like a ferret!
Class: Polecat

Text: Sea Waves over Sand Beach
Class: Breakwater

Text: Female wearing a respirator posing indoors
Class: Gasmask



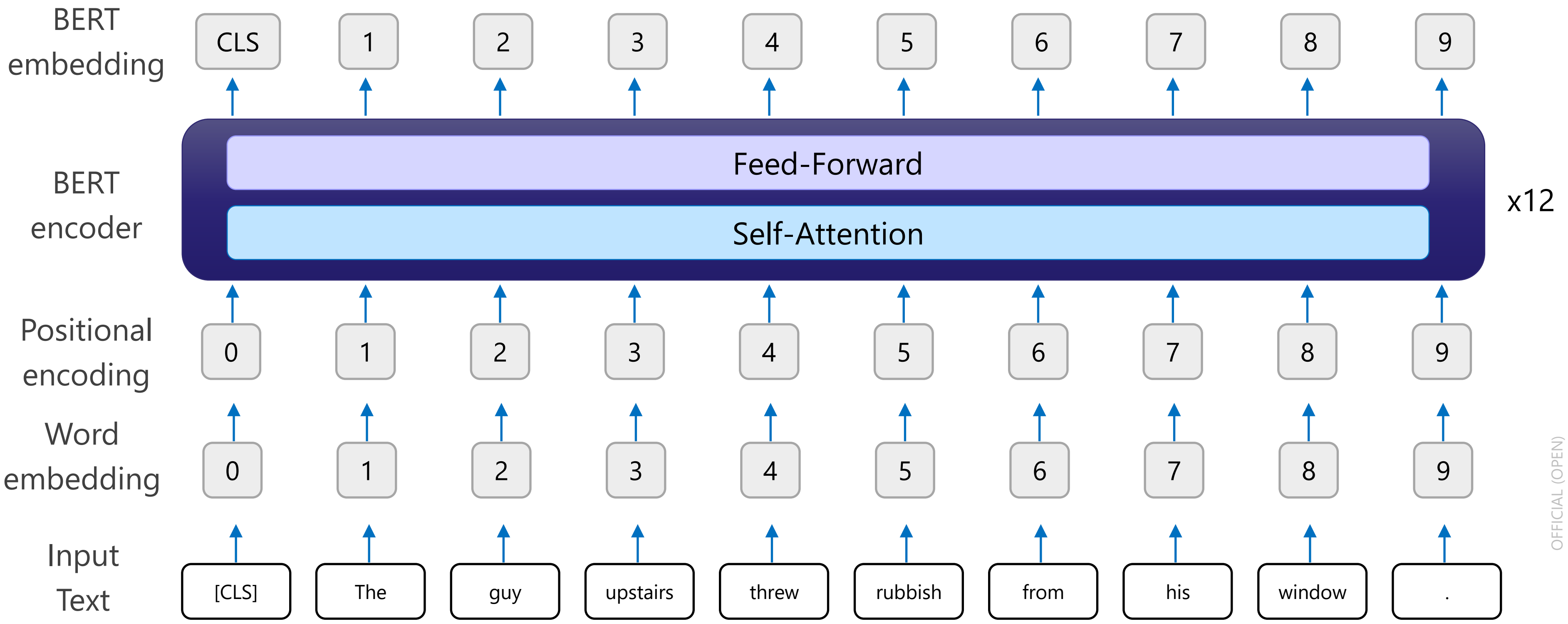
Text: Book shop Istanbul, Turkey
Class: Bookshop

Text: Fishing Photos, Blue Marin
Class: Gar

Text: Sweet Candy Store Flyer Template 2
Class: Confectionary

Task 1: Build a Text Classifier with BERT

BERT Architecture



BERT embedding: CLS 1 2 3 4 5 6 7 8 9

BERT encoder

Feed-Forward

Self-Attention

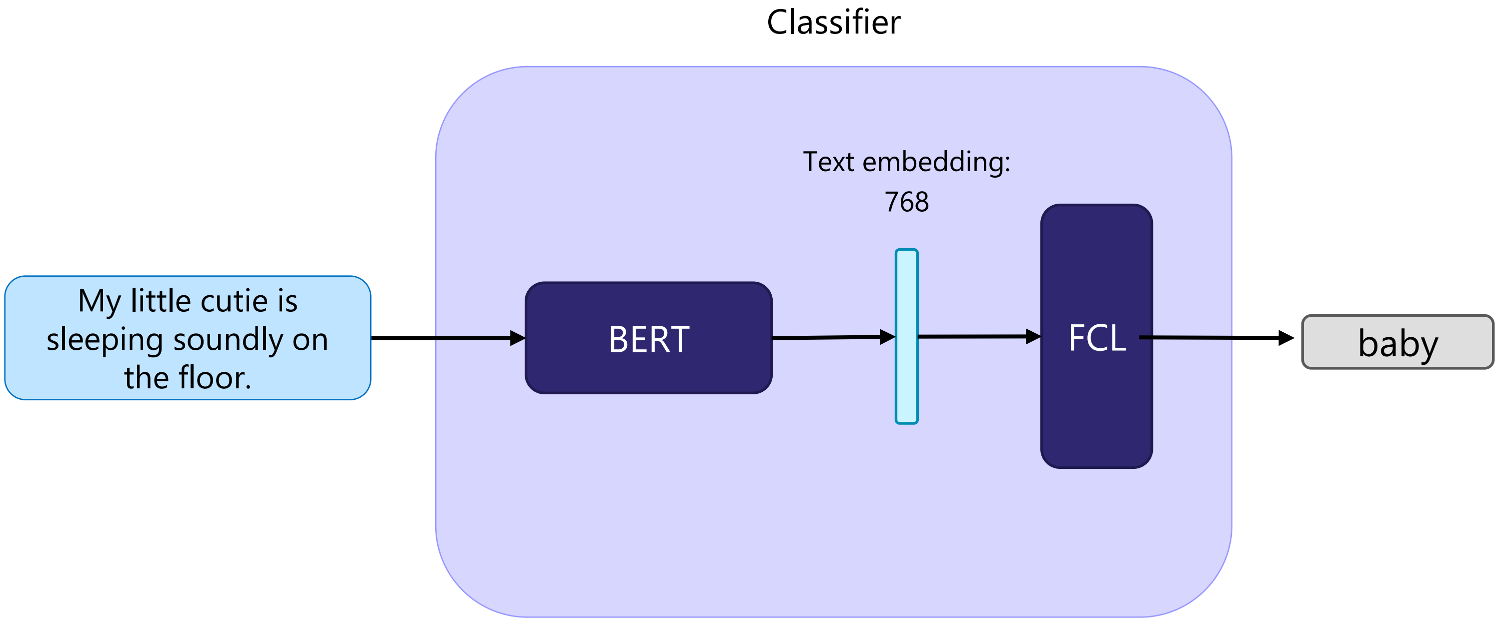
x12

Positional encoding: 0 1 2 3 4 5 6 7 8 9

Word embedding: 0 1 2 3 4 5 6 7 8 9

Input Text: [CLS] The guy upstairs threw rubbish from his window .

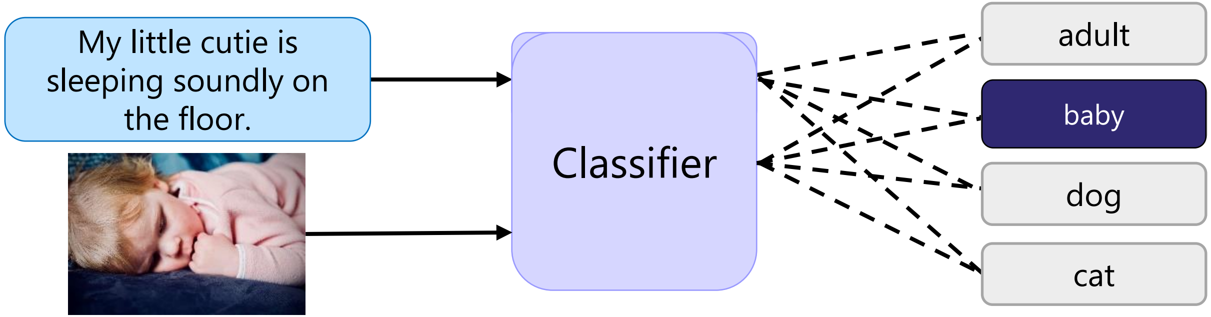
Text Encoder Architecture



Task 2: Build a Dual-Encoder Text-Image Classifier with BERT and ResNet

Combining Text and Image

My little cutie is sleeping soundly on the floor.



Dual-Encoder Text-Image Architecture

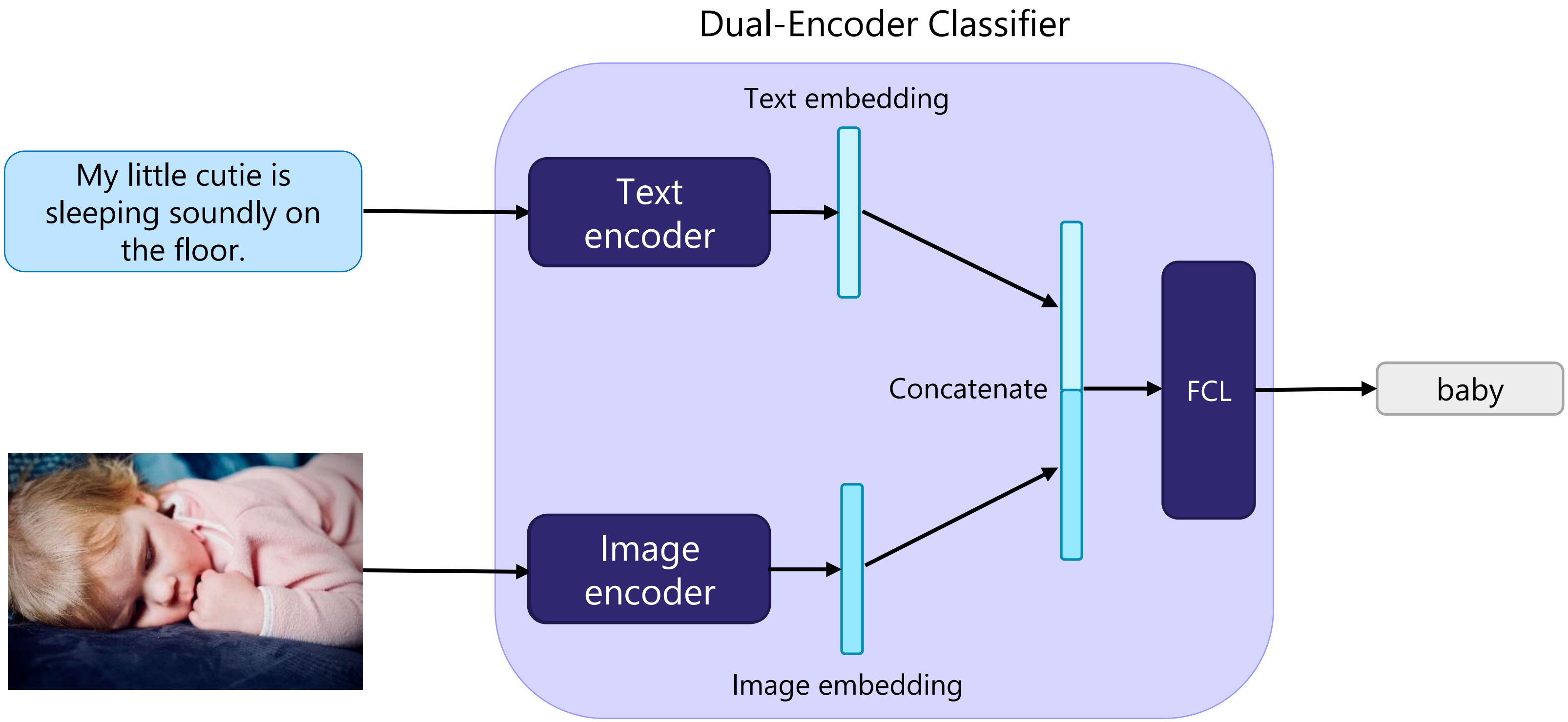


Image Encoder - ResNet

- Very deep network:
  - It can represent very complex functions
  - As the network increases in depth, classical CNNs do not perform well

- Solution: residual learning framework

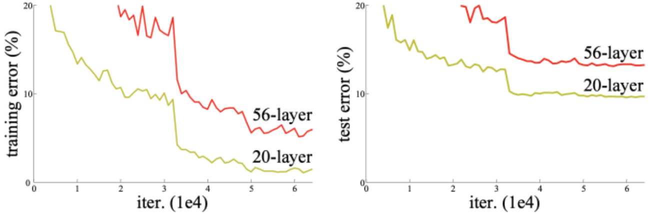
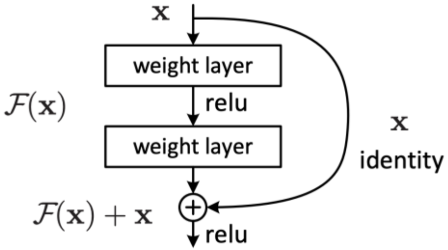


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

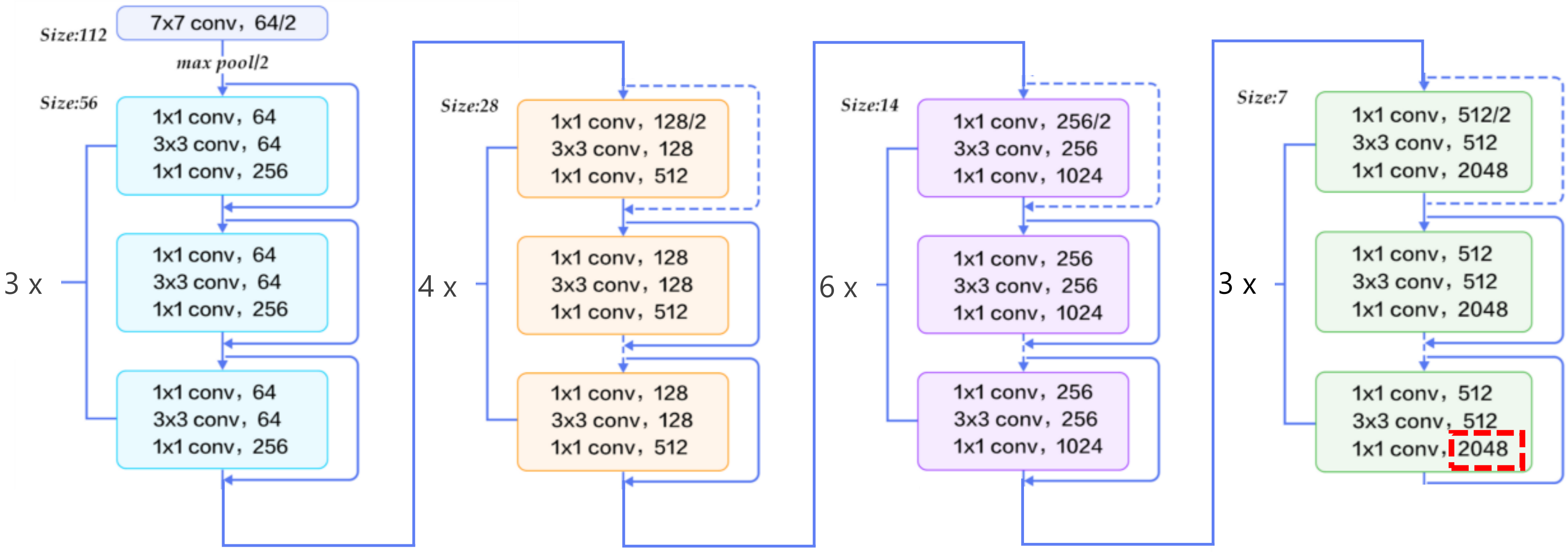


x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

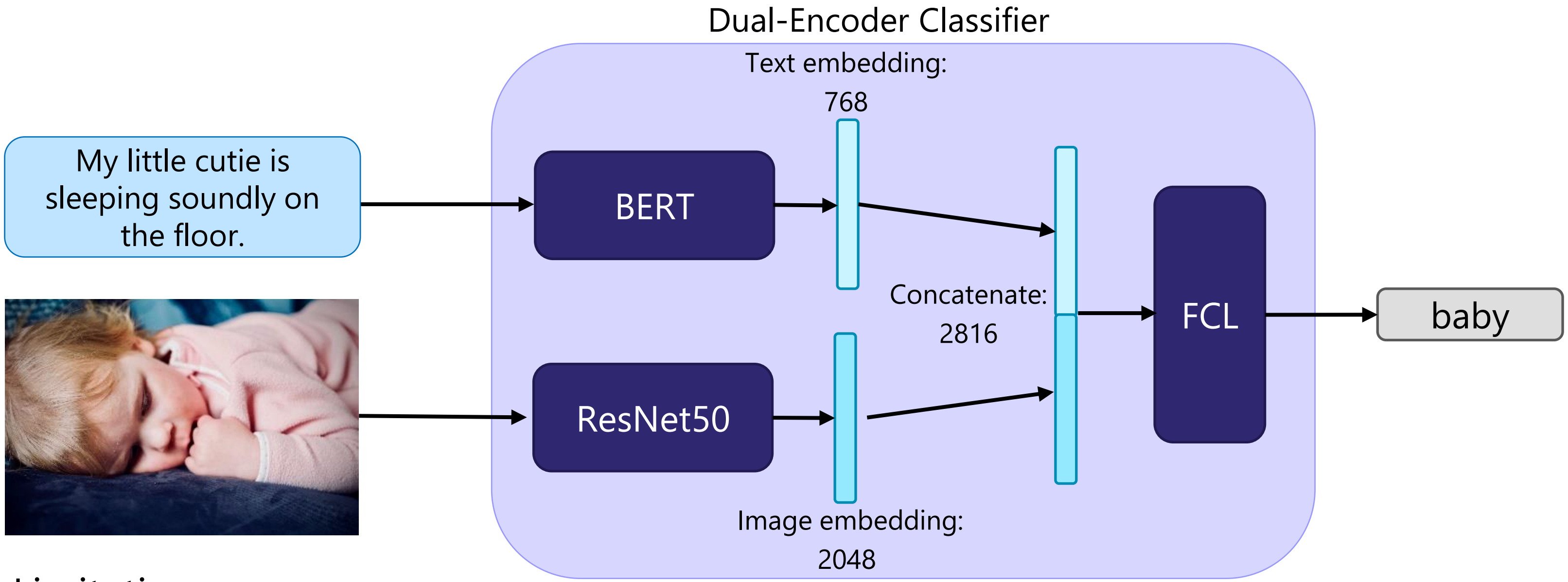
Figure 2. Residual learning: a building block.

"Deep residual learning for image recognition." He, et al. *CVPR*, 2016.

Image Encoder - ResNet50



Dual-Encoder Text-Image Architecture



My little cutie is sleeping soundly on the floor.

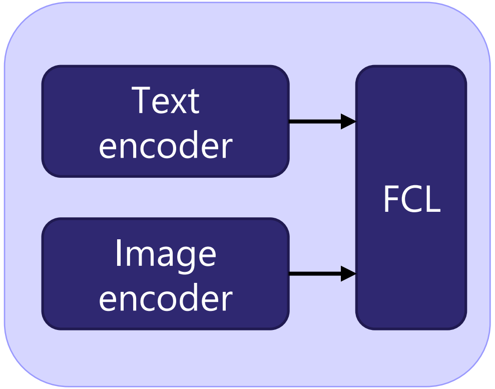
Limitations:

- Architecture only supports linear interactions between image and text
- Pretrained text and image encoders don't have knowledge about text-image pairs because they were trained separately on unimodal datasets

Task 3: Build a Joint-Encoder Text-Image Classifier with ALBEF

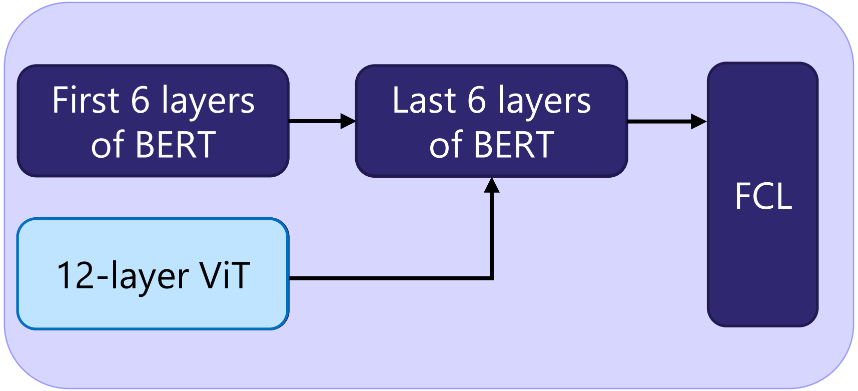
Dual-Encoder vs Joint-Encoder Architecture

Dual-Encoder Classifier



BERT + ResNet50

Joint-Encoder Classifier



ALign the image and text representations BEfore Fusing (ALBEF)

"Align before fuse: Vision and language representation learning with momentum distillation." Li, et al. *NeurIPS*, 2021.

Image Encoder: Vision Transformer

- Vision Transformer (ViT) is the encoder network of transformer
- State-of-the-art for image classification

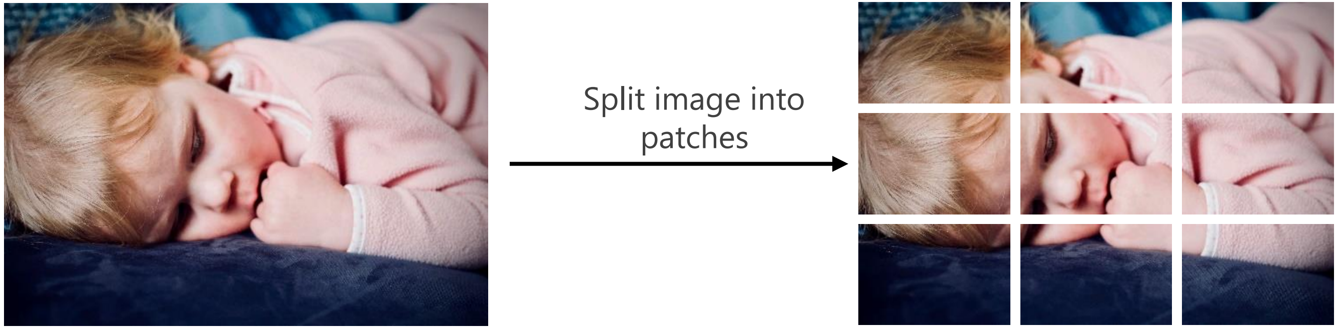
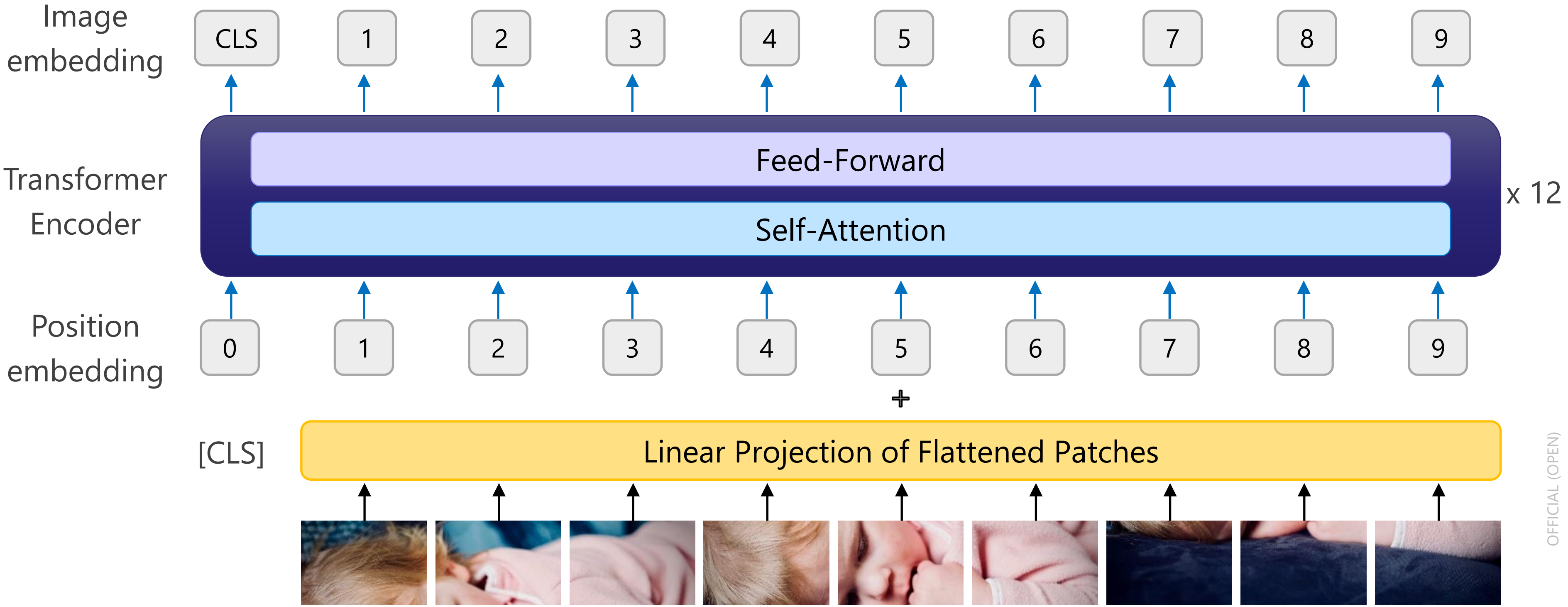
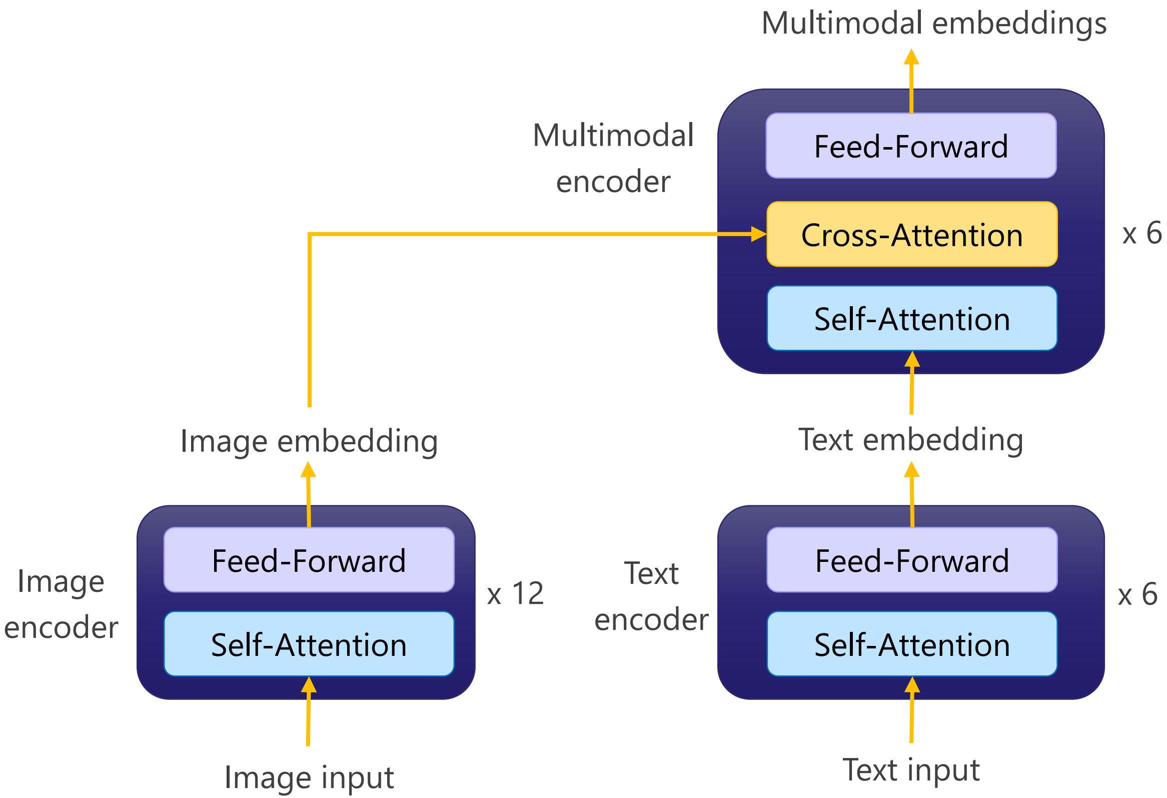


Image Encoder: Vision Transformer



"An image is worth 16x16 words: Transformers for image recognition at scale." Dosovitskiy, et al. *arXiv:2010.11929*, 2020.

Joint-Encoder Text-Image: ALBEF



(Recap) limitations of dual-encoder:

- Architecture only supports linear interactions between image and text

- Pretrained text and image encoders don't have knowledge about text-image pairs because they were trained separately on unimodal datasets

ALBEF Pretraining - Data

Text-image pairs

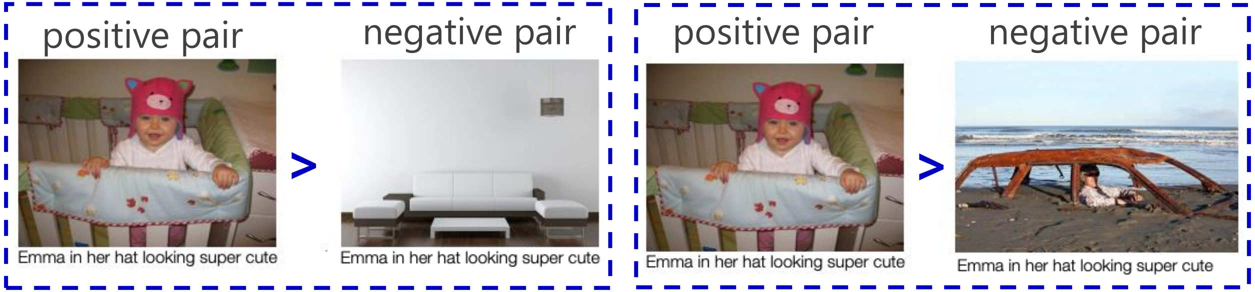
Man sits in a rusted car buried in the sand on Waitarere beach

Little girl and her dog in northern Thailand. They both seemed interested in what we were doing

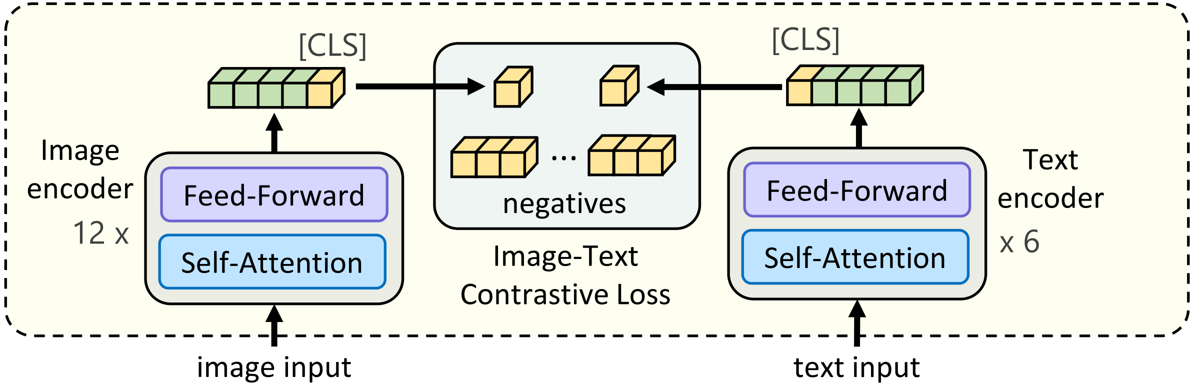
Interior design of modern white and brown living room furniture against white wall with a lamp hanging.

Emma in her hat looking super cute

ALBEF Pretraining - Objectives

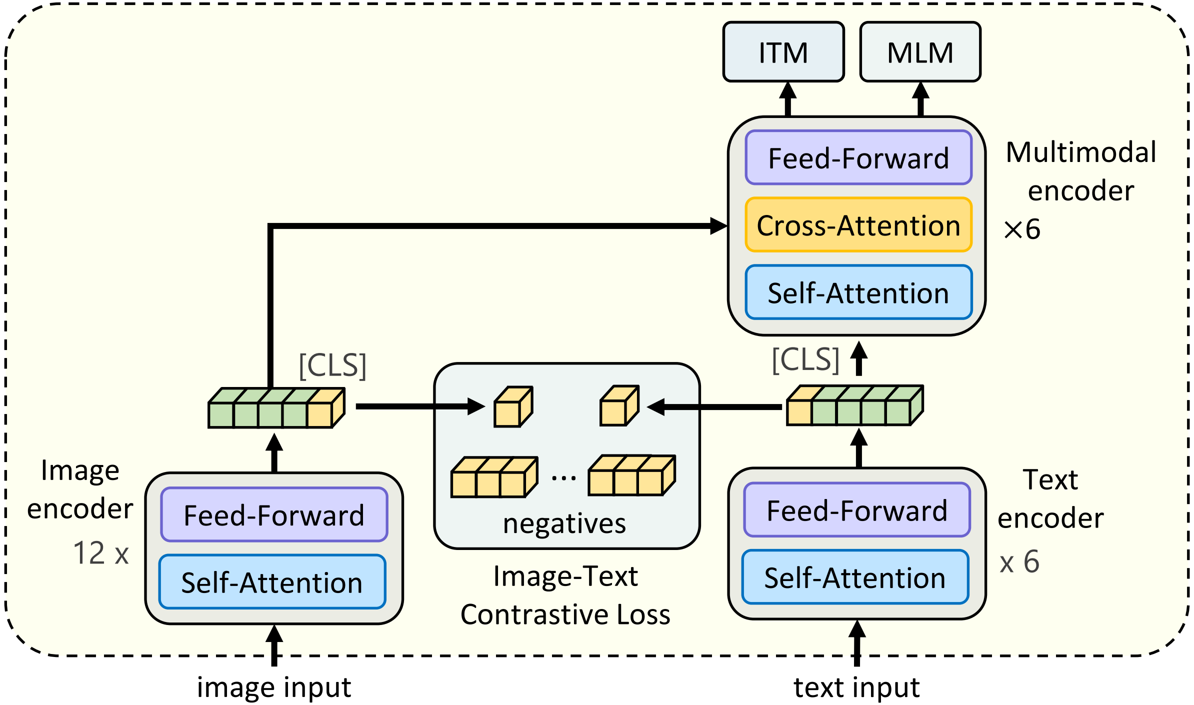

positive pair
negative pair
>
Emma in her hat looking super cute
Emma in her hat looking super cute
positive pair
negative pair
>
Emma in her hat looking super cute
Emma in her hat looking super cute

- Image-Text Contrastive Loss: align the unimodal features



"Align before fuse: Vision and language representation learning with momentum distillation." Li, et al. NeurIPS, 2021.

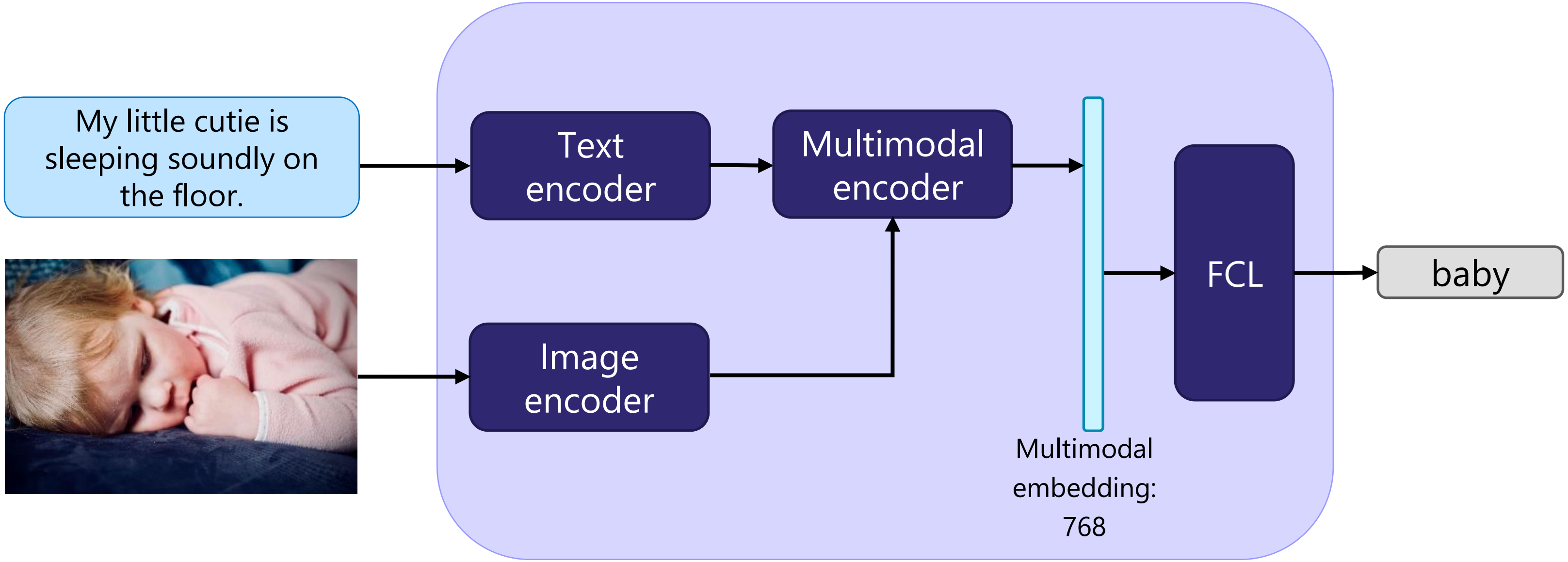
ALBEF Pretraining - Objectives



- Image-Text Contrastive Loss: align the unimodal features
- Image-Text Matching (ITM): binary classification of positives and negatives
- Mask Language Modeling (MLM): predict words based on image and contextual text

Joint-Encoder Text-Image Architecture

Joint-Encoder Classifier


My little cutie is sleeping soundly on the floor.
Text encoder
Multimodal encoder
Image encoder
FCL
baby
Multimodal embedding: 768

Discussion

Best Model Architecture?

- We have shown you **how** to train models using the different architectures on a toy dataset

- Proper evaluation on your own dataset is still required to pick the best model What is best for our dataset may not be best for yours

- Beyond model accuracy/per-class F1 scores, you should also consider:
  - Training time
  - Inference time
  - Resources required: GPU/RAM

Next Steps

- Hyperparameter tuning was not covered in this tutorial.
  You can do it on your own using these steps:

  - Split the dataset into train, validation and test sets
  - **Train set:** run *x (e.g. 50)* trials with different combinations of l*earning rate, learning rate scheduler, batch size, weight decay, training epochs*
  - **Validation set:** select the best model for each model architecture
  - **Test set:** evaluate the performance of the best hyperparameter tuned models for each model architecture, and select the best overall model

  - You can use tools like Optuna or Ray Tune
  - A good guide to hyperparameter tuning can be found here