SDAIA Academy T5C04 Bootcamps: Data Science

Deep Learning Module

Automatic Video Description Generation

Presented by: Nada Rambu

01

Problem and Solution

Project background and objectives

Objective

Generate a textual description of a video content using neural network

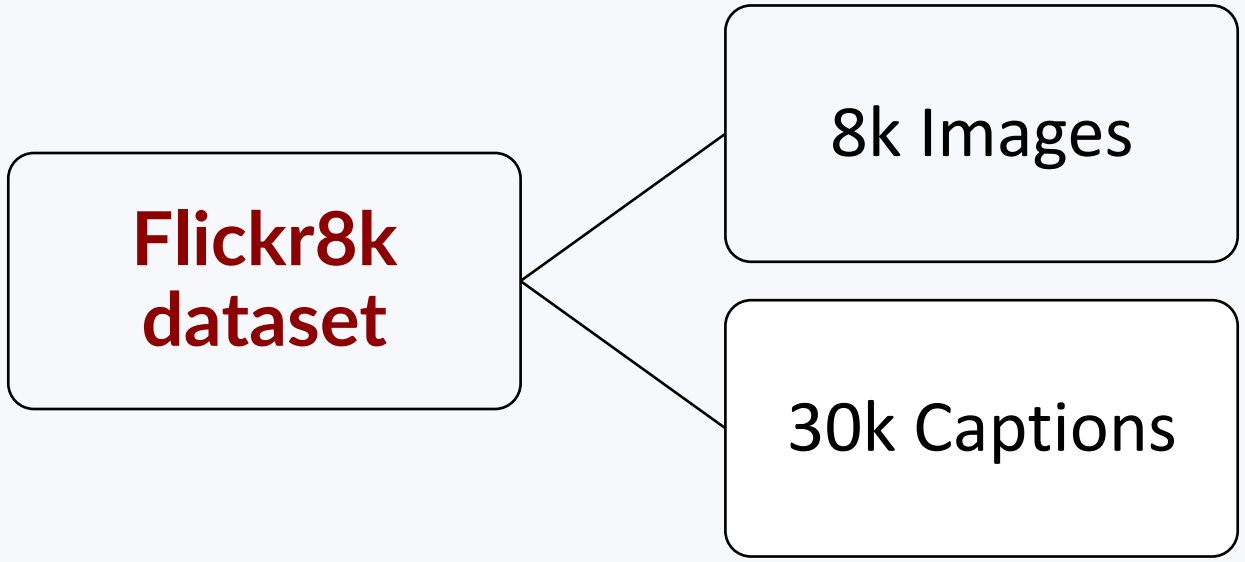

DATA SCIENCE

02

Data Preprocessing

Bringing the data from the its source to be processed

Dataset



Dataset

Dataset

Captions (5)

- A brown and white dog is running through the snow
- A dog is running in the snow
- A dog running through snow
- A white and brown dog is running through a snow covered field
- The white and brown dog is running over the surface of the snow

03

Model Development

Train and test the model to estimate results

Transfer Learning

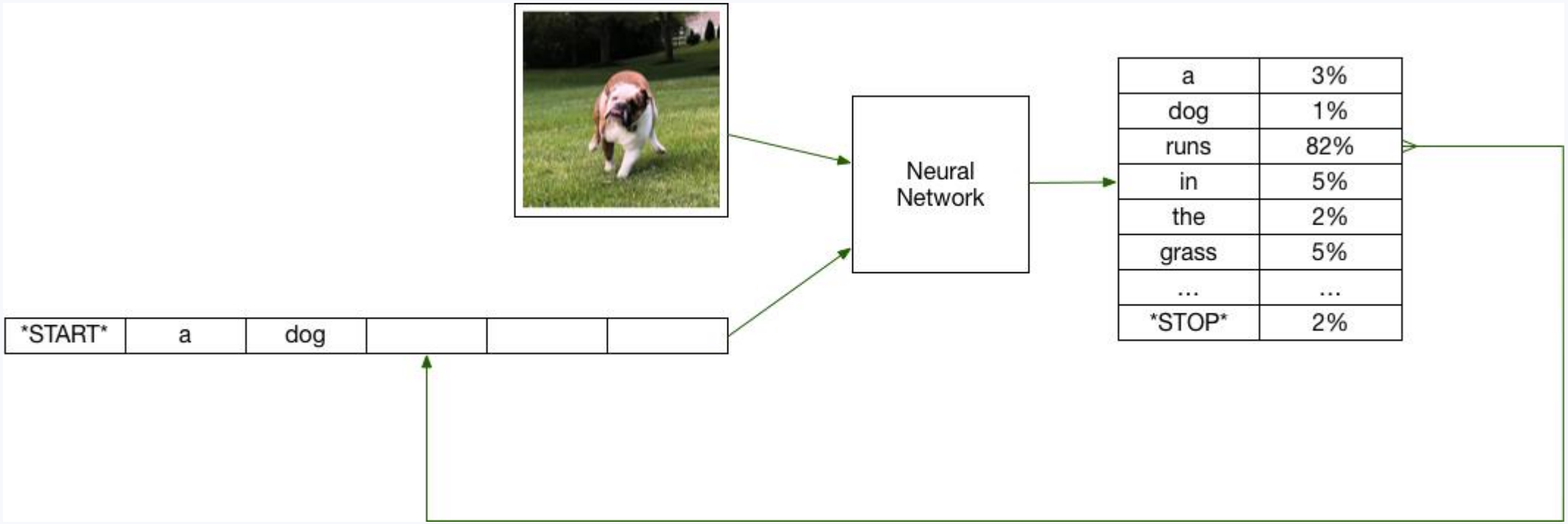
Inception V3

A large CNN model used for image analysis

GloVe (300d)

Word embeddings used for text generation

Model


Neural Network
a 3%
dog 1%
runs 82%
in 5%
the 2%
grass 5%
... ...
STOP 2%
START
a
dog

Model

Baseline model

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 34)]	0	
input_2 (InputLayer)	[(None, 2048)]	0	
embedding (Embedding)	(None, 34, 200)	330400	input_3[0][0]
dropout (Dropout)	(None, 2048)	0	input_2[0][0]
dropout_1 (Dropout)	(None, 34, 200)	0	embedding[0][0]
dense (Dense)	(None, 256)	524544	dropout[0][0]
lstm (LSTM)	(None, 256)	467968	dropout_1[0][0]
add (Add)	(None, 256)	0	dense[0][0] lstm[0][0]
dense_1 (Dense)	(None, 256)	65792	add[0][0]
dense_2 (Dense)	(None, 1652)	424564	dense_1[0][0]

Model

Baseline model

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 34)]	0	
input_2 (InputLayer)	[(None, 2048)]	0	
embedding (Embedding)	(None, 34, 200)	330400	input_3[0][0]
dropout (Dropout)	(None, 2048)	0	input_2[0][0]
dropout_1 (Dropout)	(None, 34, 200)	0	embedding[0][0]
dense (Dense)	(None, 256)	524544	dropout[0][0]
lstm (LSTM)	(None, 256)	467968	dropout_1[0][0]
add (Add)	(None, 256)	0	dense[0][0] lstm[0][0]
dense_1 (Dense)	(None, 256)	65792	add[0][0]
dense_2 (Dense)	(None, 1652)	424564	dense_1[0][0]

Model

Baseline model

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 34)]	0	
input_2 (InputLayer)	[(None, 2048)]	0	
embedding (Embedding)	(None, 34, 200)	330400	input_3[0][0]
dropout (Dropout)	(None, 2048)	0	input_2[0][0]
dropout_1 (Dropout)	(None, 34, 200)	0	embedding[0][0]
dense (Dense)	(None, 256)	524544	dropout[0][0]
lstm (LSTM)	(None, 256)	467968	dropout_1[0][0]
add (Add)	(None, 256)	0	dense[0][0] lstm[0][0]
dense_1 (Dense)	(None, 256)	65792	add[0][0]
dense_2 (Dense)	(None, 1652)	424564	dense_1[0][0]

Model

Baseline model

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 34)]	0	
input_2 (InputLayer)	[(None, 2048)]	0	
embedding (Embedding)	(None, 34, 200)	330400	input_3[0][0]
dropout (Dropout)	(None, 2048)	0	input_2[0][0]
dropout_1 (Dropout)	(None, 34, 200)	0	embedding[0][0]
dense (Dense)	(None, 256)	524544	dropout[0][0]
lstm (LSTM)	(None, 256)	467968	dropout_1[0][0]
add (Add)	(None, 256)	0	dense[0][0] lstm[0][0]
dense_1 (Dense)	(None, 256)	65792	add[0][0]
dense_2 (Dense)	(None, 1652)	424564	dense_1[0][0]

Model

Optimized model

Layer (type)	Output Shape	Param #	Connected to
input_43 (InputLayer)	[(None, 34)]	0	[]
embedding_19 (Embedding)	(None, 34, 300)	495600	['input_43[0][0]']
dropout_91 (Dropout)	(None, 34, 300)	0	['embedding_19[0][0]']
conv1d_35 (Conv1D)	(None, 30, 32)	48032	['dropout_91[0][0]']
max_pooling1d_35 (MaxPooling1D)	(None, 15, 32)	0	['conv1d_35[0][0]']
dropout_92 (Dropout)	(None, 15, 32)	0	['max_pooling1d_35[0][0]']
input_42 (InputLayer)	[(None, 2048)]	0	[]
bidirectional_18 (Bidirectional)	(None, 15, 256)	164864	['dropout_92[0][0]']
dropout_90 (Dropout)	(None, 2048)	0	['input_42[0][0]']
dropout_93 (Dropout)	(None, 15, 256)	0	['bidirectional_18[0][0]']
dense_43 (Dense)	(None, 256)	524544	['dropout_90[0][0]']
lstm_33 (LSTM)	(None, 256)	525312	['dropout_93[0][0]']
add_12 (Add)	(None, 256)	0	['dense_43[0][0]', 'lstm_33[0][0]']
dense_44 (Dense)	(None, 256)	65792	['add_12[0][0]']
dense_45 (Dense)	(None, 1652)	424564	['dense_44[0][0]']

Model

Optimized model

Layer (type)	Output Shape	Param #	Connected to
input_43 (InputLayer)	[(None, 34)]	0	[]
embedding_19 (Embedding)	(None, 34, 300)	495600	['input_43[0][0]']
dropout_91 (Dropout)	(None, 34, 300)	0	['embedding_19[0][0]']
conv1d_35 (Conv1D)	(None, 30, 32)	48032	['dropout_91[0][0]']
max_pooling1d_35 (MaxPooling1D)	(None, 15, 32)	0	['conv1d_35[0][0]']
dropout_92 (Dropout)	(None, 15, 32)	0	['max_pooling1d_35[0][0]']
input_42 (InputLayer)	[(None, 2048)]	0	[]
bidirectional_18 (Bidirectional)	(None, 15, 256)	164864	['dropout_92[0][0]']
dropout_90 (Dropout)	(None, 2048)	0	['input_42[0][0]']
dropout_93 (Dropout)	(None, 15, 256)	0	['bidirectional_18[0][0]']
dense_43 (Dense)	(None, 256)	524544	['dropout_90[0][0]']
lstm_33 (LSTM)	(None, 256)	525312	['dropout_93[0][0]']
add_12 (Add)	(None, 256)	0	['dense_43[0][0]', 'lstm_33[0][0]']
dense_44 (Dense)	(None, 256)	65792	['add_12[0][0]']
dense_45 (Dense)	(None, 1652)	424564	['dense_44[0][0]']

Model

Optimized model

Layer (type)	Output Shape	Param #	Connected to
input_43 (InputLayer)	[(None, 34)]	0	[]
embedding_19 (Embedding)	(None, 34, 300)	495600	['input_43[0][0]']
dropout_91 (Dropout)	(None, 34, 300)	0	['embedding_19[0][0]']
conv1d_35 (Conv1D)	(None, 30, 32)	48032	['dropout_91[0][0]']
max_pooling1d_35 (MaxPooling1D)	(None, 15, 32)	0	['conv1d_35[0][0]']
dropout_92 (Dropout)	(None, 15, 32)	0	['max_pooling1d_35[0][0]']
input_42 (InputLayer)	[(None, 2048)]	0	[]
bidirectional_18 (Bidirectional)	(None, 15, 256)	164864	['dropout_92[0][0]']
dropout_90 (Dropout)	(None, 2048)	0	['input_42[0][0]']
dropout_93 (Dropout)	(None, 15, 256)	0	['bidirectional_18[0][0]']
dense_43 (Dense)	(None, 256)	524544	['dropout_90[0][0]']
lstm_33 (LSTM)	(None, 256)	525312	['dropout_93[0][0]']
add_12 (Add)	(None, 256)	0	['dense_43[0][0]', 'lstm_33[0][0]']
dense_44 (Dense)	(None, 256)	65792	['add_12[0][0]']
dense_45 (Dense)	(None, 1652)	424564	['dense_44[0][0]']

Model

Optimized model

Layer (type)	Output Shape	Param #	Connected to
input_43 (InputLayer)	[(None, 34)]	0	[]
embedding_19 (Embedding)	(None, 34, 300)	495600	['input_43[0][0]']
dropout_91 (Dropout)	(None, 34, 300)	0	['embedding_19[0][0]']
conv1d_35 (Conv1D)	(None, 30, 32)	48032	['dropout_91[0][0]']
max_pooling1d_35 (MaxPooling1D)	(None, 15, 32)	0	['conv1d_35[0][0]']
dropout_92 (Dropout)	(None, 15, 32)	0	['max_pooling1d_35[0][0]']
input_42 (InputLayer)	[(None, 2048)]	0	[]
bidirectional_18 (Bidirectional)	(None, 15, 256)	164864	['dropout_92[0][0]']
dropout_90 (Dropout)	(None, 2048)	0	['input_42[0][0]']
dropout_93 (Dropout)	(None, 15, 256)	0	['bidirectional_18[0][0]']
dense_43 (Dense)	(None, 256)	524544	['dropout_90[0][0]']
lstm_33 (LSTM)	(None, 256)	525312	['dropout_93[0][0]']
add_12 (Add)	(None, 256)	0	['dense_43[0][0]', 'lstm_33[0][0]']
dense_44 (Dense)	(None, 256)	65792	['add_12[0][0]']
dense_45 (Dense)	(None, 1652)	424564	['dense_44[0][0]']

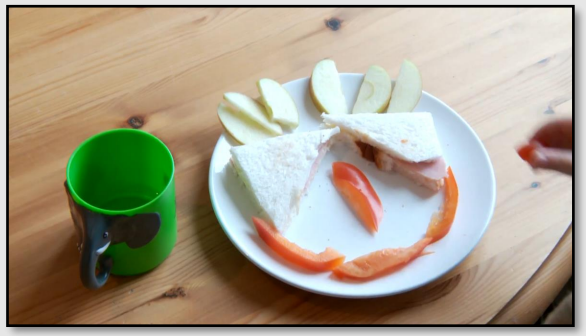
Model

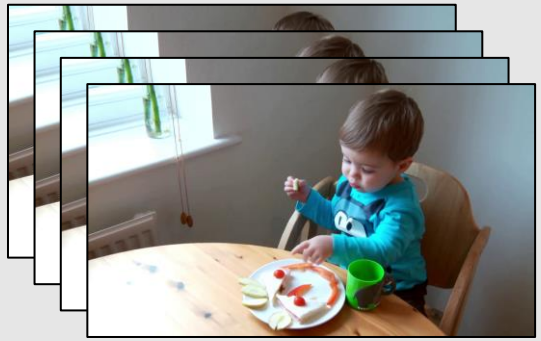
two dogs are playing in the grass

young girl in pink shirt is playing bubbles

man in red jacket and blue pants is standing on snowy mountain

Text Generation





Description 1

Description 2

Description 1
+
Description 2

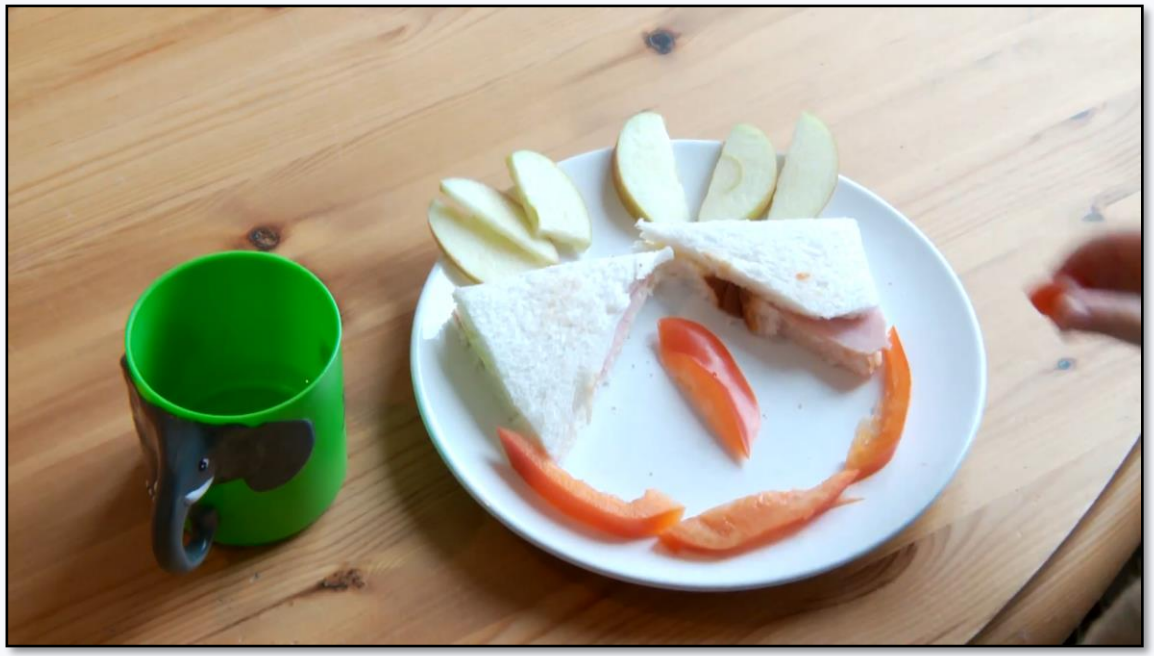
1.
Get a video

2.
Split into frames

3.
Generate a description for each scene

4.
Merge descriptions together

Text Generation



dog is eating in the air

baby in blue shirt is sitting on bed

Thank You!