

# Image Caption

Shusen Wang

# Flickr8K Image Caption Dataset

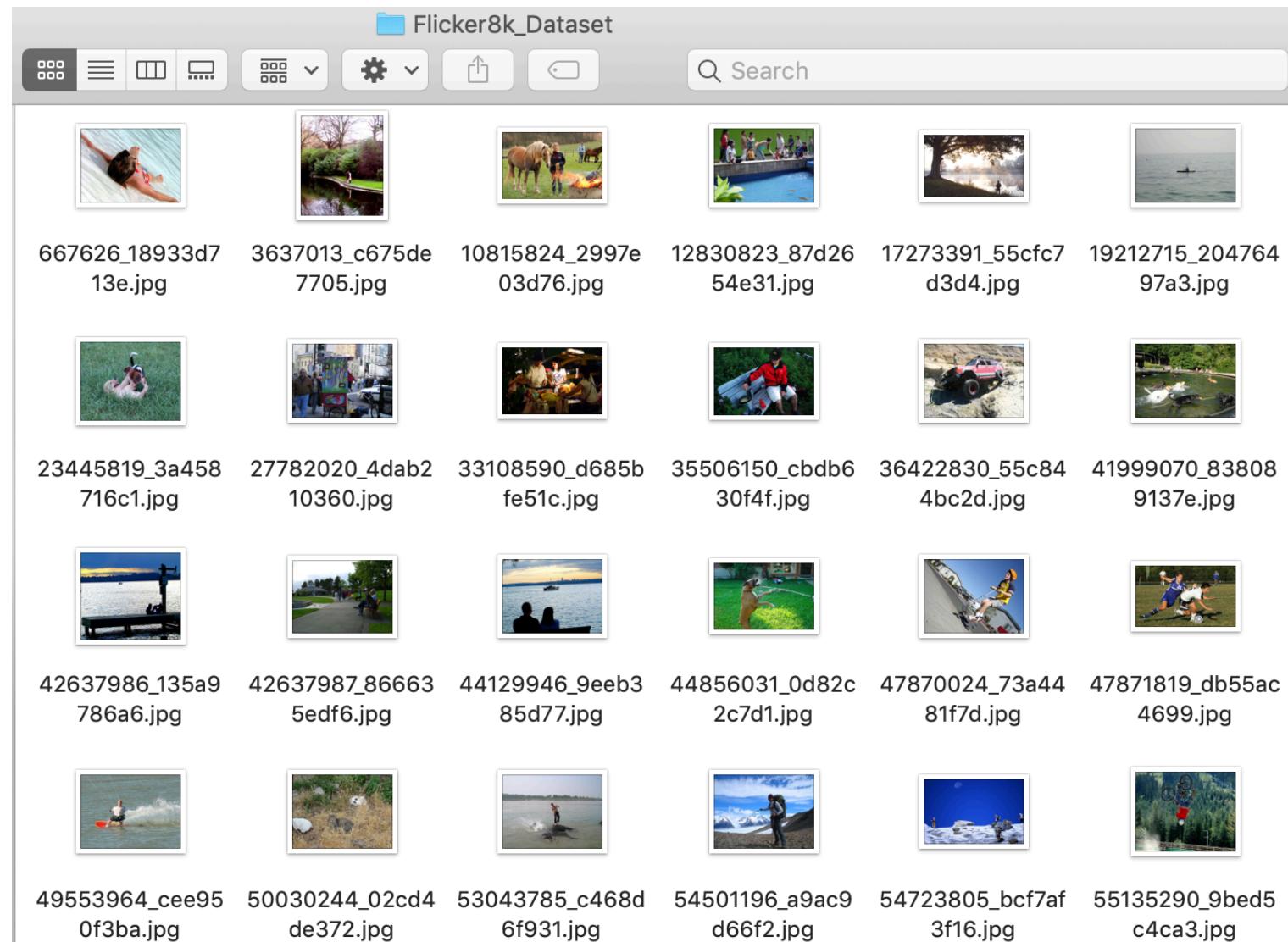
Image ID: 1000268201\_693b08cb0e



## Captions

- A child in a pink dress is climbing up a set of stairs in an entry way
- A girl going into a wooden building
- A little girl climbing into a wooden playhouse
- A little girl climbing the stairs to her playhouse
- A little girl in a pink dress going into a wooden cabin

# Flickr8K Image Caption Dataset



# Flickr8K Image Caption Dataset

File: Flickr8k.token.txt

```
1000268201_693b08cb0e.jpg#0 A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1 A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2 A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3 A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4 A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0 A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1 A black dog and a tri-colored dog playing with each other on the road .
1001773457_577c3a7d70.jpg#2 A black dog and a white dog with brown spots are staring at each other in the street .
1001773457_577c3a7d70.jpg#3 Two dogs of different breeds looking at each other on the road .
1001773457_577c3a7d70.jpg#4 Two dogs on pavement moving toward each other .
1002674143_1b742ab4b8.jpg#0 A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .
1002674143_1b742ab4b8.jpg#1 A little girl is sitting in front of a large painted rainbow .
1002674143_1b742ab4b8.jpg#2 A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
1002674143_1b742ab4b8.jpg#3 There is a girl with pigtails sitting in front of a rainbow painting .
1002674143_1b742ab4b8.jpg#4 Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0 A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1 A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2 a man sleeping on a bench outside with a white and black dog sitting next to him .
1003163366_44323f5815.jpg#3 A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4 man laying on bench holding leash of dog sitting on ground
1007129816_e794419615.jpg#0 A man in an orange hat starring at something .
1007129816_e794419615.jpg#1 A man wears an orange hat and glasses .
1007129816_e794419615.jpg#2 A man with gauges and glasses is wearing a Blitz hat .
1007129816_e794419615.jpg#3 A man with glasses is wearing a beer can crocheted hat .
1007129816_e794419615.jpg#4 The man with pierced ears is wearing glasses and an orange hat .
1007320043_627395c3d8.jpg#0 A child playing on a rope net .
1007320043_627395c3d8.jpg#1 A little girl climbing on red roping .
1007320043_627395c3d8.jpg#2 A little girl in pink climbs a rope bridge at the park .
1007320043_627395c3d8.jpg#3 A small child grips onto the red ropes at the playground .
1007320043_627395c3d8.jpg#4 The small child climbs on a red ropes on a playground .
1009434119_febe49276a.jpg#0 A black and white dog is running in a grassy garden surrounded by a white fence .
1009434119_febe49276a.jpg#1 A black and white dog is running through the grass .
1009434119_febe49276a.jpg#2 A Boston terrier is running in the grass .
1009434119_febe49276a.jpg#3 A Boston Terrier is running on lush green grass in front of a white fence .
1009434119_febe49276a.jpg#4 A dog runs on the green grass near a wooden fence .
1012212859_01547e3f17.jpg#0 A dog shakes its head near the shore , a red ball next to it .
1012212859_01547e3f17.jpg#1 A white dog shakes on the edge of a beach with an orange ball .
1012212859_01547e3f17.jpg#2 Dog with orange ball at feet , stands on shore shaking off water
1012212859_01547e3f17.jpg#3 White dog playing with a red ball on the shore near the water .
1012212859_01547e3f17.jpg#4 White dog with brown ears standing near water with head turned to one side .
1015118661_980735411b.jpg#0 A boy smiles in front of a stony wall in a city .
1015118661_980735411b.jpg#1 A little boy is standing on the street while a man in overalls is working on a stone wall .
1015118661_980735411b.jpg#2 A young boy runs across the street .
1015118661_980735411b.jpg#3 A young child is walking on a stone paved street with a metal pole and a man behind him .
1015118661_980735411b.jpg#4 Smiling boy in white shirt and blue jeans in front of rock wall with man in overalls behind him .
1015584366_dfcecc3c85a.jpg#0 A black dog leaps over a log .
1015584366_dfcecc3c85a.jpg#1 A grey dog is leaning over a fallen tree .
```

# Feature Extraction Using Pretrained VGG16

# Pretrained VGG16

```
from keras.applications import VGG16
from keras.models import Model

vgg16 = VGG16(weights='imagenet',
               include_top=True,
               input_shape=(224, 224, 3))

vgg16 = Model(inputs=vgg16.inputs,
              outputs=vgg16.layers[-2].output)

vgg16.summary()
```

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
Total params: 134,260,544		
Trainable params: 134,260,544		
Non-trainable params: 0		

# Pretrained VGG16

Layer (type)	Output Shape	Param #
=====		
input_1 (InputLayer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
• • •		
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
=====		

Total params: 134,260,544

Trainable params: 134,260,544

Non-trainable params: 0

# Feature Extraction

Convert  $224 \times 224 \times 3$  images to 4096-dim vectors

```
from keras.preprocessing import image
from keras.applications.vgg16 import preprocess_input

img_path = 'Flicker8k_Dataset/667626_18933d713e.jpg'

img = image.load_img(img_path, target_size=(224, 224))
x = image.img_to_array(img)
x = preprocess_input(x)
x = x.reshape(1, 224, 224, 3)

features = vgg16.predict(x) ——————> 4096-dim vector
```

# Feature Extraction

Images

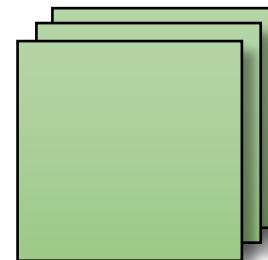
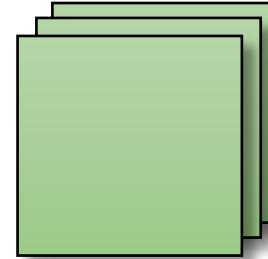


•  
•  
•

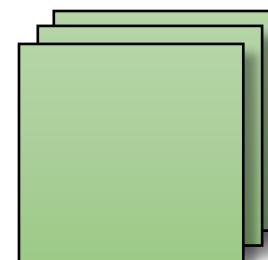


Tensors

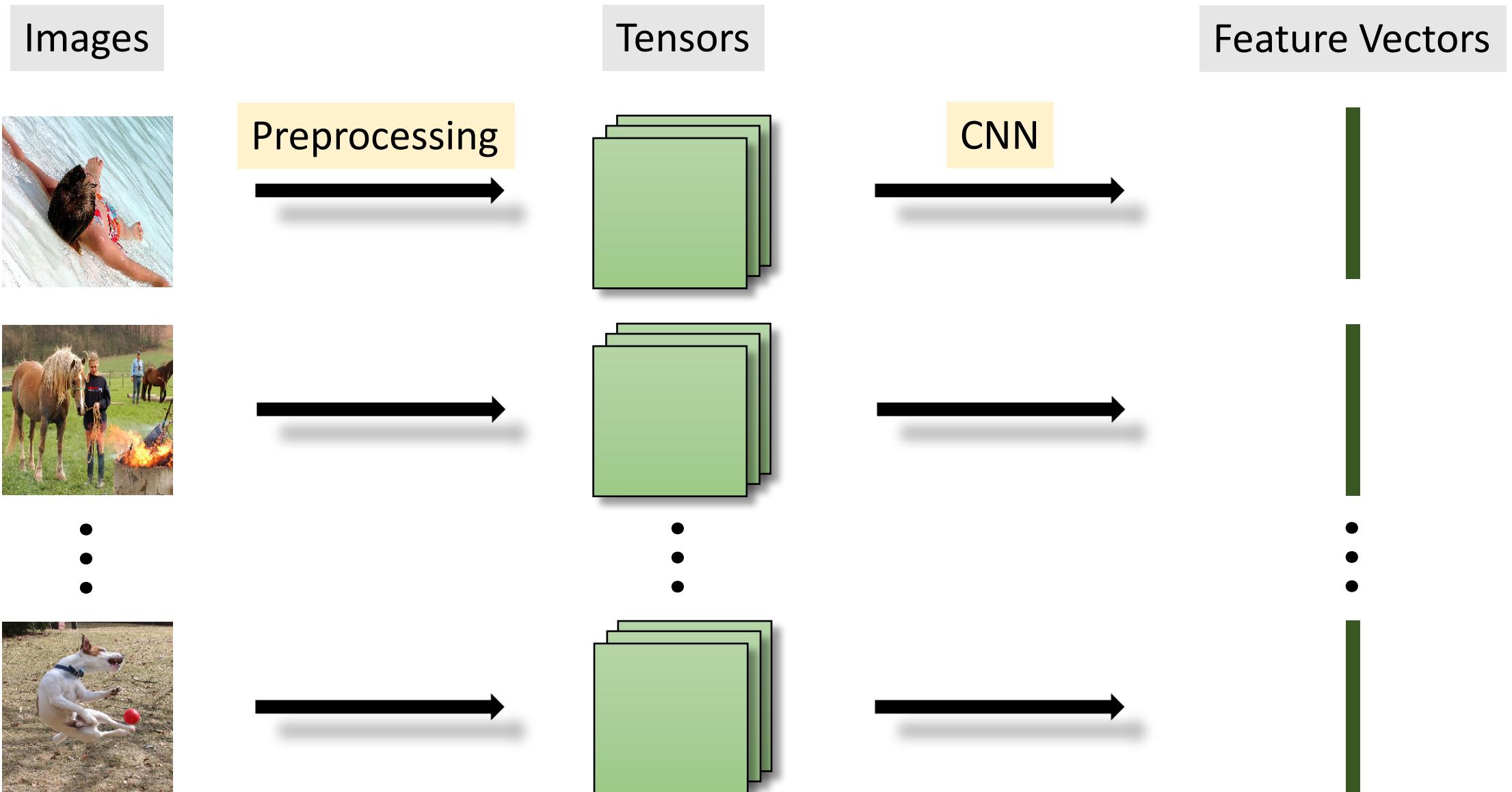
Preprocessing



•  
•  
•



# Feature Extraction



# Text Processing

# Processing Text Data

5 different  
captions for  
the same  
image

Image Path	Caption
1000268201_693b08cb0e.jpg#0	A child in a pink dress is climbing up a set of stairs in an
1000268201_693b08cb0e.jpg#1	A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2	A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg#3	A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg#4	A little girl in a pink dress going into a wooden cabin .
1001773457_577c3a7d70.jpg#0	A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg#1	A black dog and a tri-colored dog playing with each other on
1001773457_577c3a7d70.jpg#2	A black dog and a white dog with brown spots are staring at
1001773457_577c3a7d70.jpg#3	Two dogs of different breeds looking at each other on the ro
1001773457_577c3a7d70.jpg#4	Two dogs on pavement moving toward each other .
1002674143_1b42ab4b8.jpg#0	A little girl covered in paint sits in front of a painted ra
1002674143_1b742ab4b8.jpg#1	A little girl is sitting in front of a large painted rainbow
1002674143_1b742ab4b8.jpg#2	A small girl in the grass plays with fingerpaints in front o
1002674143_1b742ab4b8.jpg#3	There is a girl with pigtails sitting in front of a rainbow
1002674143_1b742ab4b8.jpg#4	Young girl with pigtails painting outside in the grass .
1003163366_44323f5815.jpg#0	A man lays on a bench while his dog sits by him .
1003163366_44323f5815.jpg#1	A man lays on the bench to which a white dog is also tied .
1003163366_44323f5815.jpg#2	a man sleeping on a bench outside with a white and black dog
1003163366_44323f5815.jpg#3	A shirtless man lies on a park bench with his dog .
1003163366_44323f5815.jpg#4	man laying on bench holding leash of dog sitting on ground
1007129816_e794419615.jpg#0	A man in an orange hat staring at something .
1007129816_e794419615.jpg#1	A man wears an orange hat and glasses .
1007129816_e794419615.jpg#2	A man with gauges and glasses is wearing a Blitz hat .
1007129816_e794419615.jpg#3	A man with glasses is wearing a beer can crocheted hat .
1007129816_e794419615.jpg#4	The man with pierced ears is wearing glasses and an orange h
1007320043_627395c3d8.jpg#0	A child playing on a rope net .
1007320043_627395c3d8.jpg#1	A little girl climbing on red roping .
1007320043_627395c3d8.jpg#2	A little girl in pink climbs a rope bridge at the park .
1007320043_627395c3d8.jpg#3	A small child grips onto the red ropes at the playground .
1007320043_627395c3d8.jpg#4	The small child climbs on a red ropes on a playground .

# Processing Text Data

**Convert the file “Flickr8k.token” to a list of (ID, caption)**

- ('1000268201\_693b08cb0e', 'a child in a pink dress is climbing up a set of stairs in an entry way'),
- ('1000268201\_693b08cb0e', 'a girl going into a wooden building'),
- ('1000268201\_693b08cb0e', 'a little girl climbing into a wooden playhouse'),
- ('1000268201\_693b08cb0e', 'a little girl climbing the stairs to her playhouse'),
- ('1000268201\_693b08cb0e', 'a little girl in a pink dress going into a wooden cabin')
- ('1001773457\_577c3a7d70', 'a black dog and a spotted dog are fighting'),
- ('1001773457\_577c3a7d70', 'a black dog and a tri-colored dog playing with each other on the road'),  
•  
•  
•

# Texts to Sequences

texts[i] :

'**startseq** a child in a pink dress is climbing up  
a set of stairs in an entry way **endseq**'



Tokenization

tokens[i] :

[ '**startseq**', 'a', 'child', 'in', 'a', 'pink',  
'dress', 'is', 'climbing', 'up', 'a', 'set', 'of',  
'stairs', 'in', 'an', 'entry', 'way', '**endseq**' ]



Encoding

seqs[i] :

[ **2**, 1, 43, 4, 1, 90, 172, 7, 119, 51, 1, 394, 12,  
395, 4, 28, 5159, 670, **3** ]

# Processing Text Data

A list of (ID, caption) to (ID, sequence)

- ('1000268201\_693b08cb0e', [2, 1, 43, 4, 1, 90, 172, 7, 119, 51, 1, 394, 12, 395, 4, 28, 5159, 670, 3]),
- ('1000268201\_693b08cb0e', [2, 1, 19, 316, 64, 1, 196, 117, 3]),
- ('1000268201\_693b08cb0e', [2, 1, 40, 19, 119, 64, 1, 196, 2437, 3]),
- ('1000268201\_693b08cb0e', [2, 1, 40, 19, 119, 5, 395, 20, 60, 2437, 3]),
- ('1000268201\_693b08cb0e', [2, 1, 40, 19, 4, 1, 90, 172, 316, 64, 1, 196, 2981, 3]),
- ('1001773457\_577c3a7d70', [2, 1, 15, 9, 8, 1, 843, 9, 17, 343, 3]),
- ('1001773457\_577c3a7d70', [2, 1, 15, 9, 8, 1, 1575, 235, 9, 34, 10, 137, 82, 6, 5, 151, 3])
  - 
  - 
  -

# Prepare Training Data

# Training Data

- Inputs: (image feature, sequence)
- Output: next word

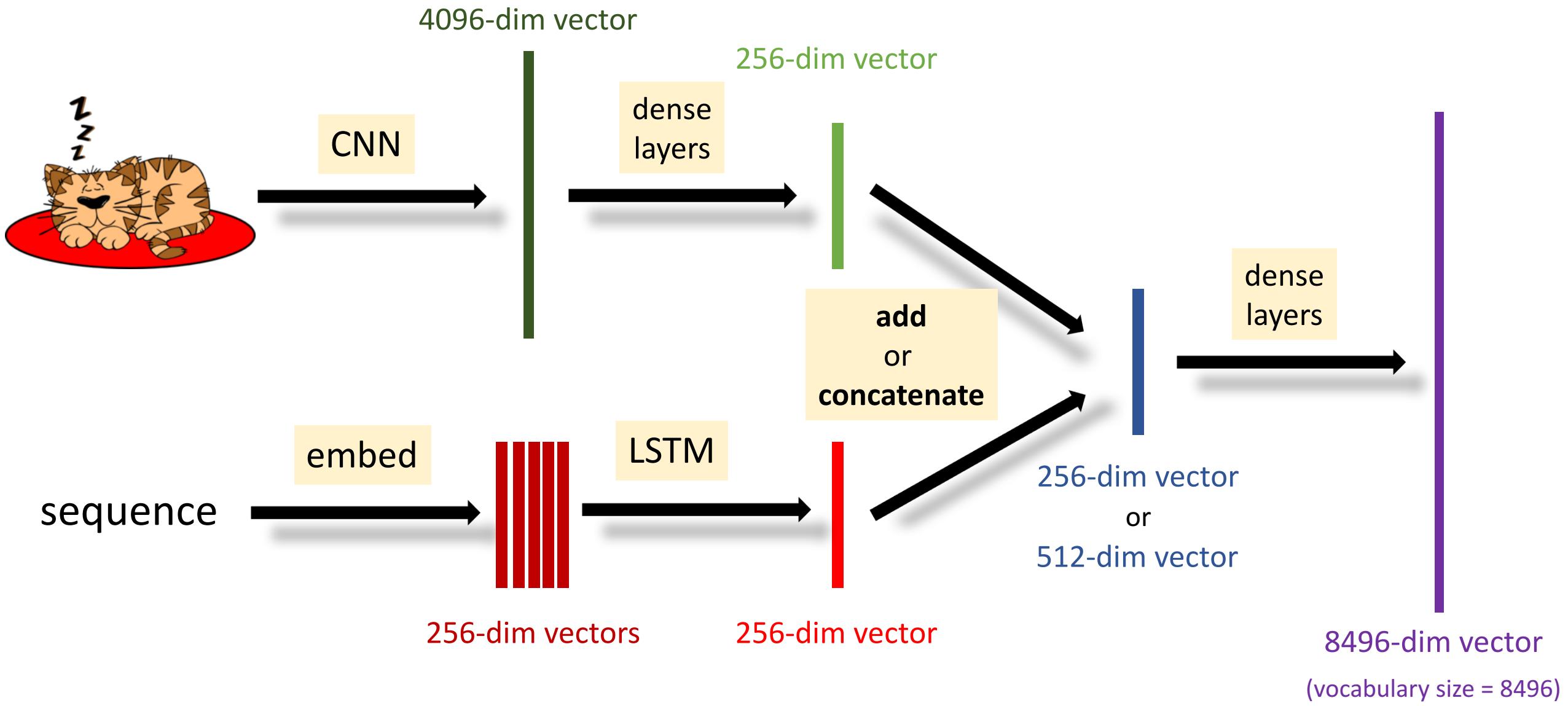
x1	x2	y
image feature	'startseq'	'a'
image feature	'startseq a'	'cat'
image feature	'startseq a cat'	'sat'
image feature	'startseq a cat sat'	'on'
image feature	'startseq a cat sat on'	'a'
image feature	'startseq a cat sat on a'	'mat'
image feature	'startseq a cat sat on a mat'	'endseq'

# Training Data

- Inputs: (**image feature**, **sequence**)
- Output: **next word**

- **#training images:** 6K
- **#captions:** 30K
- **# (x<sub>1</sub>, x<sub>2</sub>, y) triplets:** 354K

# Deep Learning Model



# Network Structure

```
from keras.layers import Input, Dropout, Dense, Embedding, LSTM, Flatten, Add

img_input = Input(shape=(4096,), name='img_input')
img_dropout = Dropout(0.5, name='img_dropout')(img_input)
img_dense = Dense(256, activation='relu', name='img_dense')(img_dropout)

seq_input = Input(shape=(max_len,), name='seq_input')
seq_embed = Embedding(vocabulary, 256, name='seq_embed')(seq_input)
seq_lstm = LSTM(256, dropout=0.2, name='seq_lstm')(seq_embed)

pred_add = Add(name='pred_add')([img_dense, seq_lstm])
pred_dropout1 = Dropout(0.5, name='pred_dropout1')(pred_add)
pred_dense1 = Dense(256, activation='relu', name='pred_dense1')(pred_dropout1)
pred_dropout2 = Dropout(0.5, name='pred_dropout2')(pred_dense1)
outputs = Dense(vocabulary, activation='softmax', name='pred_dense2')(pred_dropout2)
```

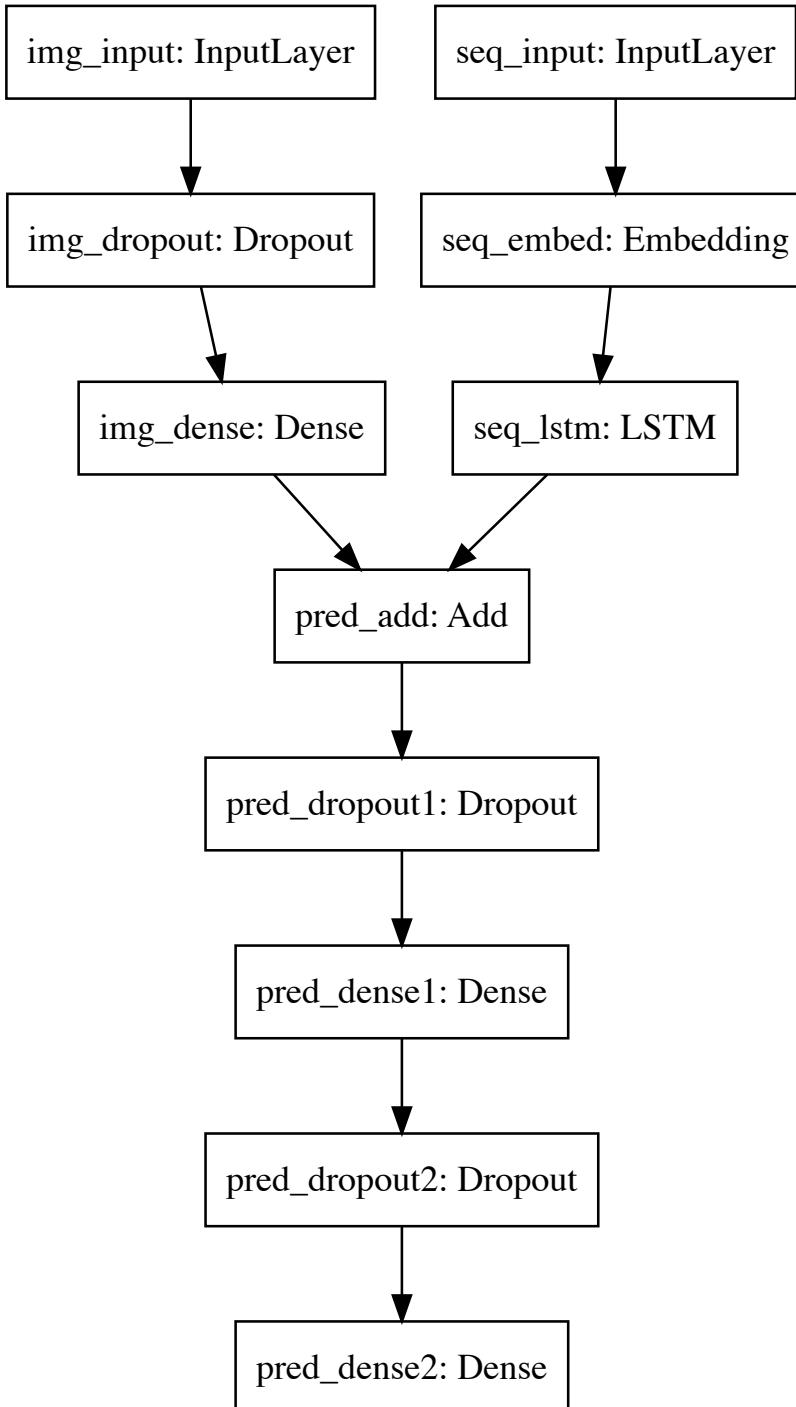
# Network Structure

```
from keras.models import Model  
model = Model(inputs=[img_input, seq_input], outputs=outputs)  
  
model.summary()
```

Layer (type)	Output Shape	Param #	Connected to
<hr/>			
img_input (InputLayer)	(None, 4096)	0	
seq_input (InputLayer)	(None, 39)	0	
img_dropout (Dropout)	(None, 4096)	0	img_input[0][0]
seq_embed (Embedding)	(None, 39, 256)	2174976	seq_input[0][0]
img_dense (Dense)	(None, 256)	1048832	img_dropout[0][0]
seq_lstm (LSTM)	(None, 256)	525312	seq_embed[0][0]
pred_add (Add)	(None, 256)	0	img_dense[0][0] seq_lstm[0][0]
pred_dropout1 (Dropout)	(None, 256)	0	pred_add[0][0]
pred_dense1 (Dense)	(None, 256)	65792	pred_dropout1[0][0]
pred_dropout2 (Dropout)	(None, 256)	0	pred_dense1[0][0]
pred_dense2 (Dense)	(None, 8496)	2183472	pred_dropout2[0][0]
<hr/>			
Total params: 5,998,384			
Trainable params: 5,998,384			
Non-trainable params: 0			

# Network Structure

```
from keras.models import Model  
model = Model(inputs=[img_input, seq_input], outputs=outputs)  
  
model.summary()  
=====  
Layer (type)          Output Shape       Param #  Connected to  
=====  
img_input (InputLayer) (None, 4096)        0  
seq_input (InputLayer) (None, 39)          0  
img_dropout (Dropout)  (None, 4096)        0          img_input[0][0]  
seq_embed (Embedding) (None, 39, 256)      2174976   seq_input[0][0]  
img_dense (Dense)     (None, 256)         1048832   img_dropout[0][0]  
seq_lstm (LSTM)       (None, 256)         525312    seq_embed[0][0]  
pred_add (Add)        (None, 256)         0          img_dense[0][0]  
                           seq_lstm[0][0]  
pred_dropout1 (Dropout) (None, 256)        0          pred_add[0][0]  
pred_dense1 (Dense)   (None, 256)         65792     pred_dropout1[0][0]  
pred_dropout2 (Dropout) (None, 256)        0          pred_dense1[0][0]  
pred_dense2 (Dense)   (None, 8496)        2183472   pred_dropout2[0][0]  
=====  
Total params: 5,998,384  
Trainable params: 5,998,384  
Non-trainable params: 0
```



# Train the Model

```
model.compile(loss='categorical_crossentropy',  
              optimizer='rmsprop')
```

images' features      sequence      target word

The diagram illustrates the mapping between the parameters in the `model.fit()` call and the code above it. A green curly arrow points from "images' features" to the `x1_train` and `x2_train` parameters. A red vertical arrow points from "sequence" to the `y_train` parameter. A red wavy arrow points from "target word" to the `batch_size`, `epochs`, and `validation_split` parameters.

```
model.fit([x1_train, x2_train], y_train,  
          batch_size=32, epochs=20, validation_split=0.2)
```

# Generate Caption

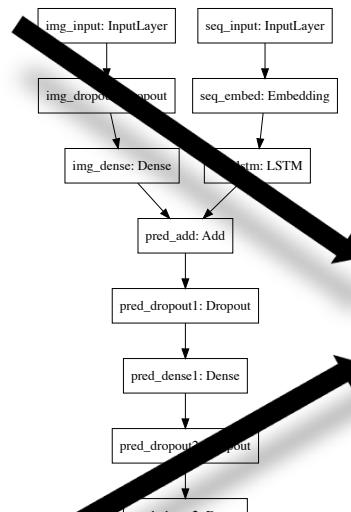
# Make Predictions

Input Image



'startseq'

Input Sequence



sampling

' a '

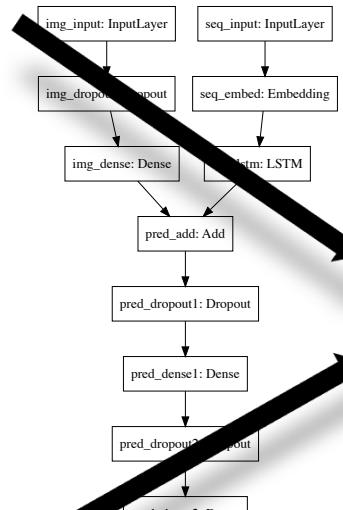
# Make Predictions

Input Image



'startseq a'

Input Sequence



sampling

'cat'

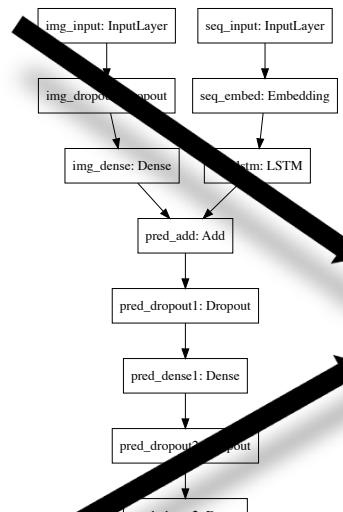
# Make Predictions

Input Image



'startseq a cat'

Input Sequence



sampling

'sat'

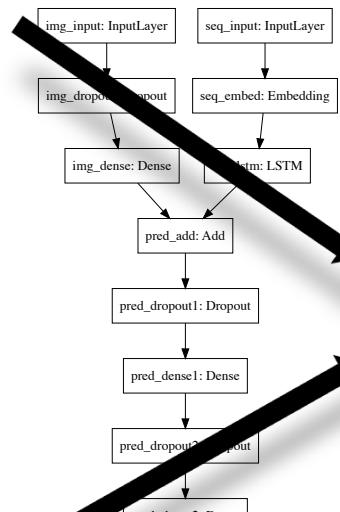
# Make Predictions

Input Image



'startseq a cat sat'

Input Sequence



sampling

' on '

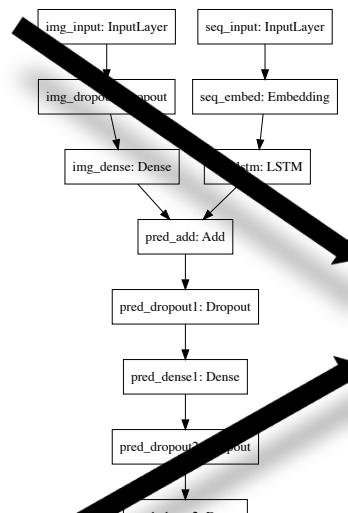
# Make Predictions

Input Image



'startseq a cat sat on'

Input Sequence



sampling

'the'

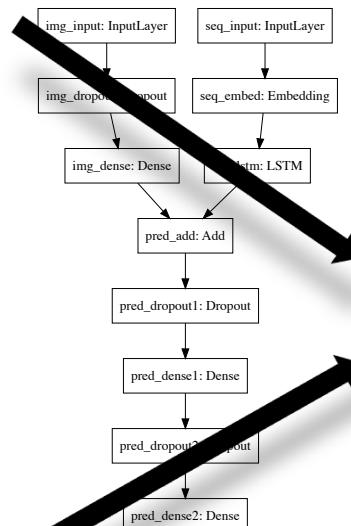
# Make Predictions

Input Image



'startseq a cat sat on the'

Input Sequence



sampling

'mat'

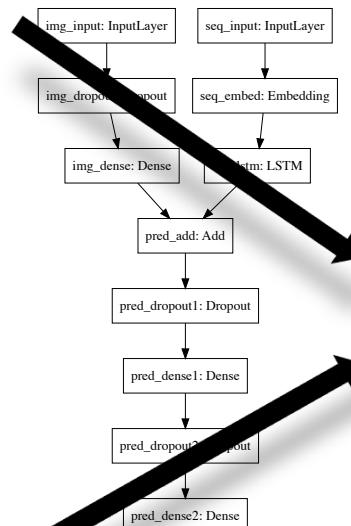
# Make Predictions

Input Image



'startseq a cat sat on the mat'

Input Sequence



sampling

'endseq'

# Details

Refer to the hand-on tutorial:

- <https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/>