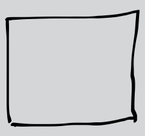
I/P.

→ f (I/P Img) → Coption.

Data → Image
→ Text (captions).

Image



=> Some Shape.

224 × 224 × 3.

Reshaping.

Batch Size – 32/64.

32 × 224 × 224 × 3

32 × 3 × 224 × 224

B × C × H × W.

Normalization.

Zero Mean | Unit Variance.

Text.

Train [— — — — —]

Test [ ]

① In Train Data.

② Iterate over the captions.

③ Tokenize using Spacy Tokenizer.

I am a student. ← Lowercase

=> ["I", "am", "a", "student"]

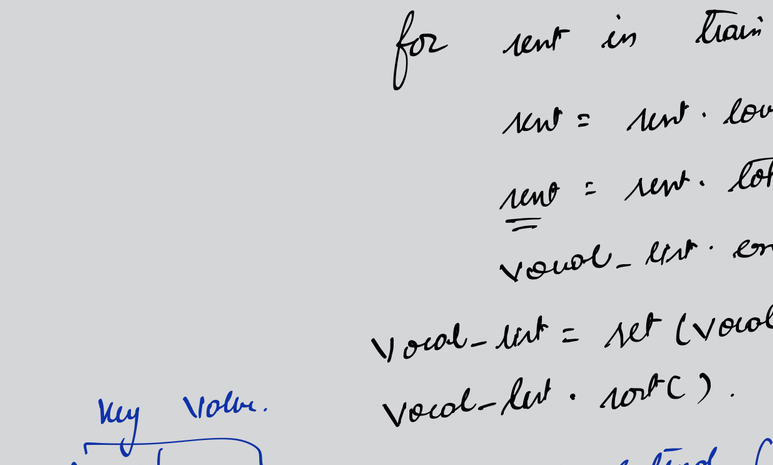
④ Vocab List = [ ] ← Initially Empty.

```
vocab_list = [ ]

for sent in train_sent:

    sent = sent.lower()

    sent = sent.tokenize()   ["i", "am", ...]

    vocab_list.extend(sent).

vocab_list = set(vocab_list).

vocab_list.sort()

vocab_list.extend(["<SOS>", "<EOS>", "<PAD>", "<UNK>"]).

word2int = { }

for i, w in enumerate(vocab_list):

    word2int[w] = i

int2word = { } ← Reverse of word2int.
```

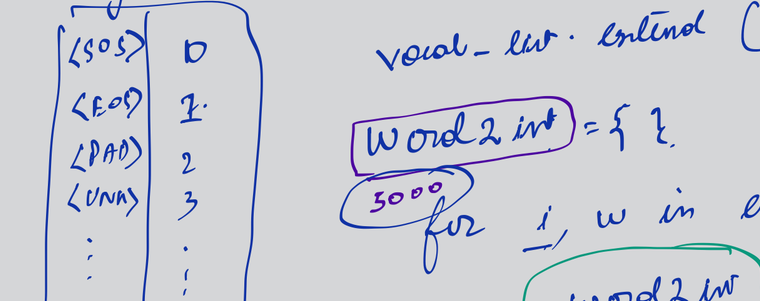
Key	Value
<SOS>	0
<EOS>	1
<PAD>	2
<UNK>	3
⋮	⋮

Batch

32



① max length of Sentence in a batch.
= 15, 20



Target Tensor

Src Tensor

32 × 3 × 224 × 224.

Tgt Tensor

32 × 20.

32 × 45.

Resnet.

LSTM.

Img → Resnet → 512 dim.