IMAGE CAPTIONING USING DEEP LEARNING

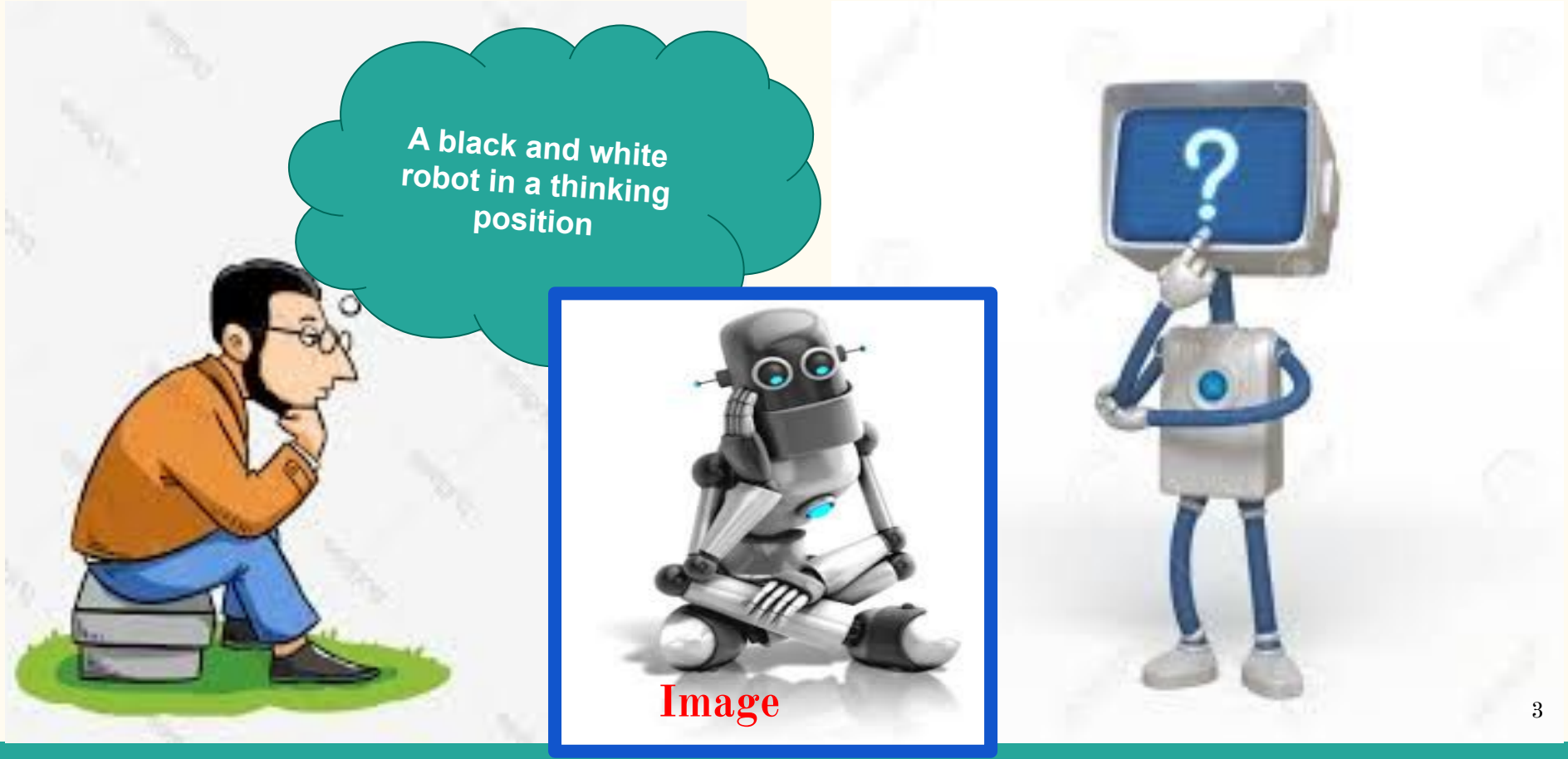
Mahendra Nandi

RKMVERI
BDA-ML COURSE

Basic and Important Topics

1. **What** is Image Captioning ?
2. **How** It Can be Done !
3. **Dataset** For this Project
4. Transfer Learning with **ResNet50**
5. Concept of memory in RNN [i.e, **LSTM**]
6. **Training** Results
7. **BLEU** score
8. **Evaluation** With BLEU Score
9. **Caption** Generated by the Model
10. **Further Work** That Can Be Done

1.What is Image Captioning ?



2. How Human do it !!

Find key things

Dog- object
Brown dog- about object
Running -action
Grass - background
Green grass - about background

Play with the words and Create a meaningful sentence

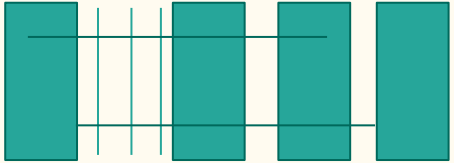
- A brown dog run
- A brown dog run over grass .
- A brown dog with its front paw off the ground on a grassy surface near red and purple flower .
- A dog run across a grassy lawn near some flower
- A yellow dog be play in a grassy area near flower .

How it can be done by a machine

Feature extraction

Features for the whole image

LSTM model



a brown dog is running over green grass.

How it can be done by a machine

CNN

Transfer learning

Long Feature vector
all information are encoded

LSTM model

<end> brown a <start>

a brown dog is running over green grass.

During Training

➢ <start> A dog run across ... lawn <end>

processing

CNN

Vector representation of a word

Long Feature vector

To calculate error

To easy and quick learn

LSTM model

A brown dog is running over green grass.

During Validation

processing

CNN

Long Feature vector

To calculate error

LSTM model

A brown dog is running over green grass.

During Testing

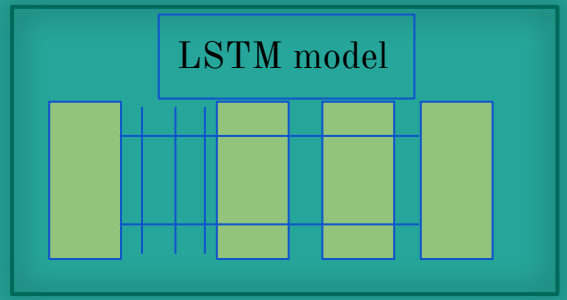
➢ <start> A dog run across ... lawn <end>

processing

CNN

Long Feature vector

LSTM model



A brown dog is running over green grass.

To see model performance

3. OUR DATASET

So, What should our DATASET contain ?

At least , to train the model we need ,
Some **IMAGE** and corresponding **CAPTION(s)**
YES !

Flicker8K Dataset

- Training
  6000 images
- Validation
  1000 images
- Testing
  1000 images

Total 8000 images
In a folder

X 5 = total 40000 captions in a file

This is how it is given

```
1 test_imd_id_df.head(12)
```

	img_id
0	3385593926_d3e9c21170.jpg
1	2677656448_6b7e7702af.jpg
2	311146855_0b65fdb169.jpg
3	1258913059_07c613f7ff.jpg
4	241347760_d44c8d3a01.jpg
5	2654514044_a70a6e2c21.jpg
6	2339106348_2df90aa6a9.jpg
7	256085101_2c2617c5d0.jpg
8	280706862_14c30d734a.jpg
9	3072172967_630e9c69d0.jpg
10	3482062809_3b694322c4.jpg
11	1167669558_87a8a467d6.jpg

time: 9.92 ms (started: 2021-06-23 08:06:30 +00:00)

```
1 all_cap_df.head(12)
```

	img_id	img_caption
0	1305564994_00513f9a5b.jpg#0	A man in street racer armor be examine the tir...
1	1305564994_00513f9a5b.jpg#1	Two racer drive a white bike down a road .
2	1305564994_00513f9a5b.jpg#2	Two motorist be ride along on their vehicle th...
3	1305564994_00513f9a5b.jpg#3	Two person be in a small race car drive by a g...
4	1305564994_00513f9a5b.jpg#4	Two person in race uniform in a street car .
5	1351764581_4d4fb1b40f.jpg#0	A firefighter extinguish a fire under the hood...
6	1351764581_4d4fb1b40f.jpg#1	a fireman spray water into the hood of small w...
7	1351764581_4d4fb1b40f.jpg#2	A fireman spray inside the open hood of small ...
8	1351764581_4d4fb1b40f.jpg#3	A fireman use a firehose on a car engine that ...
9	1351764581_4d4fb1b40f.jpg#4	Firefighter use water to extinguish a car that...
10	1358089136_976e3d2e30.jpg#0	A boy sand surf down a hill
11	1358089136_976e3d2e30.jpg#1	A man be attempt to surf down a hill make of s...

time: 15.6 ms (started: 2021-06-23 08:06:30 +00:00)

Some Visualizations

Two dog play in the snow .

Two brown dog wrestle in the snow .

Two brown dog playful fight in the snow .

Dog play on the snow .

Dog be in the snow in front of a fence .

A small brown and white dog be in a pool .

Small dog be paddle through the water in a pool .

A dog swim in a pool near a person .

A dog in a swim pool swim toward sombody we cannot see .

a brown and white dog swim towards some in a pool

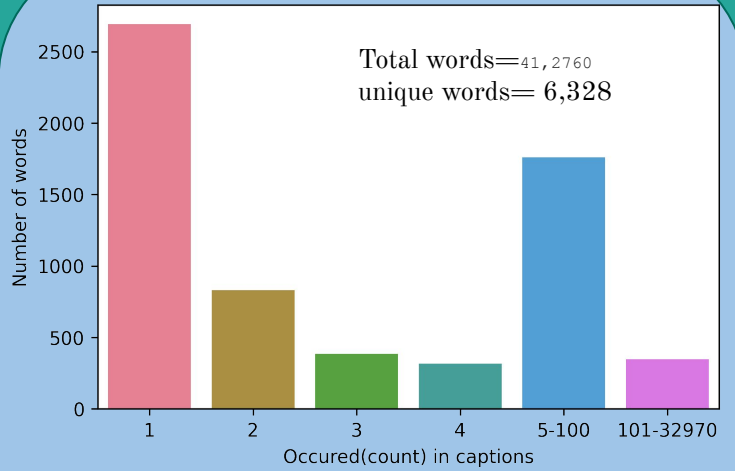
Two person be dance with drum on the right and a crowd behind them .

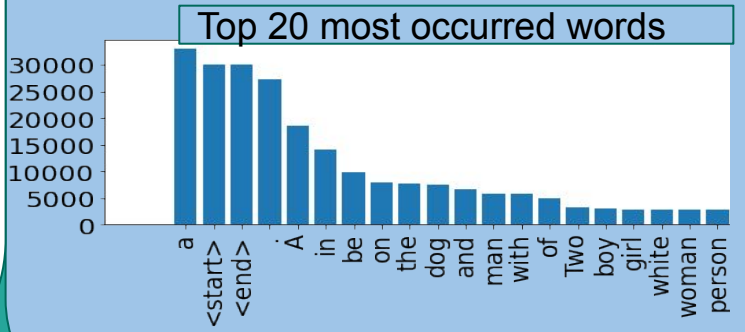
one performer wear a feathered headdress dance with another performer in street

A man and a woman wear decorative costume and dance in a crowd of onlooker .

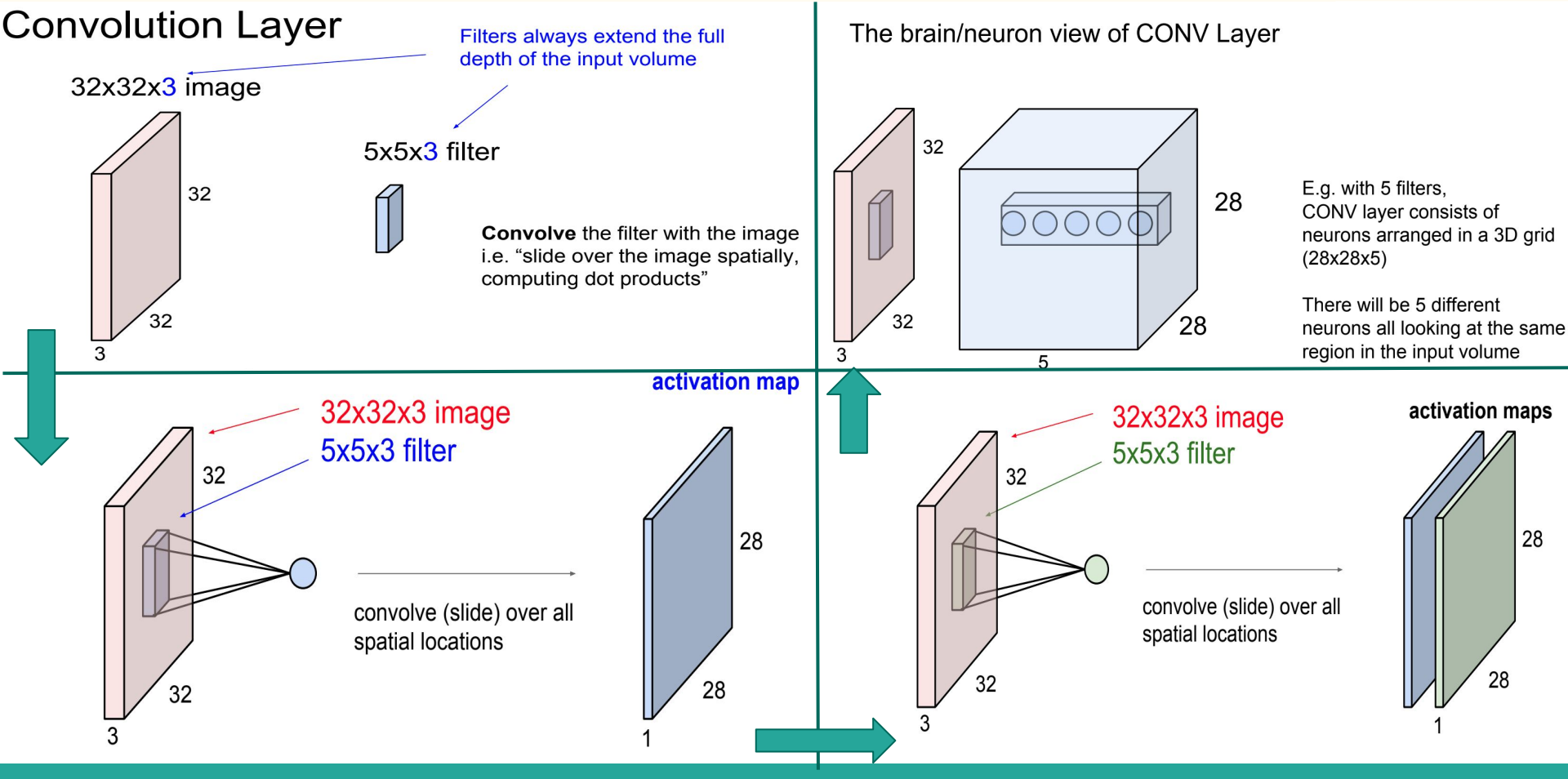
A man and a woman with feather on her head dance .

A man and a woman in festive costume dance .





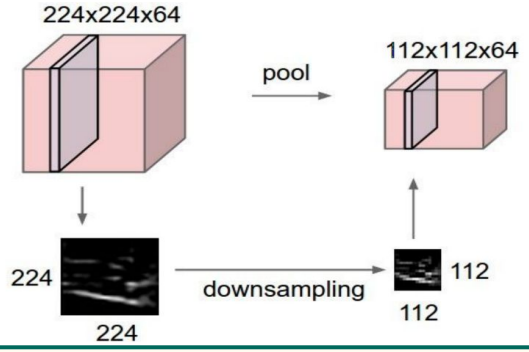
4. Transfer Learning with ResNet50

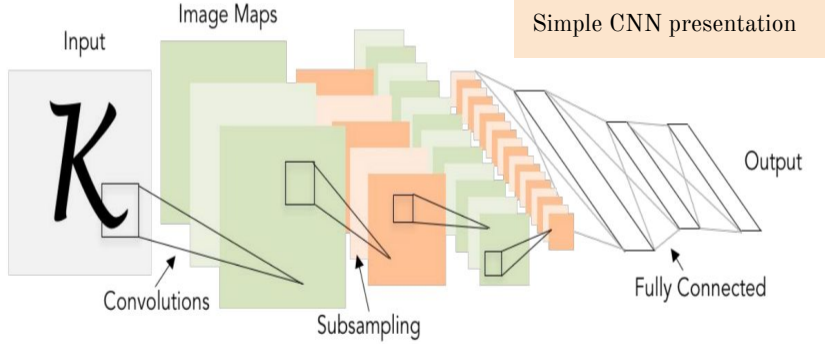


ResNet architecture

Pooling layer

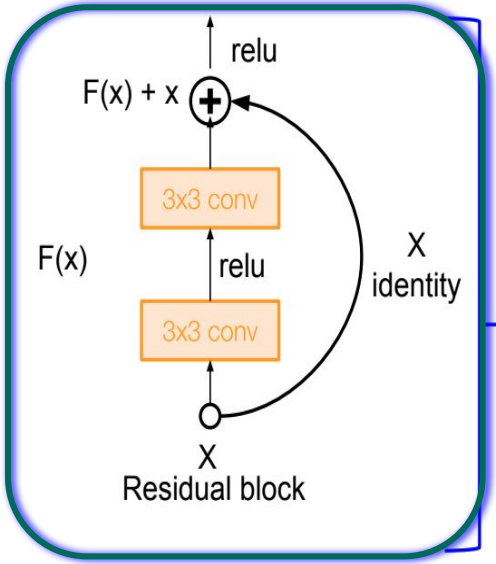
- makes the representations smaller and more manageable
- operates over each activation map independently:

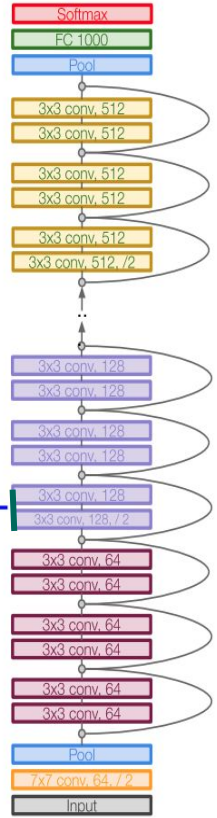




Simple CNN presentation

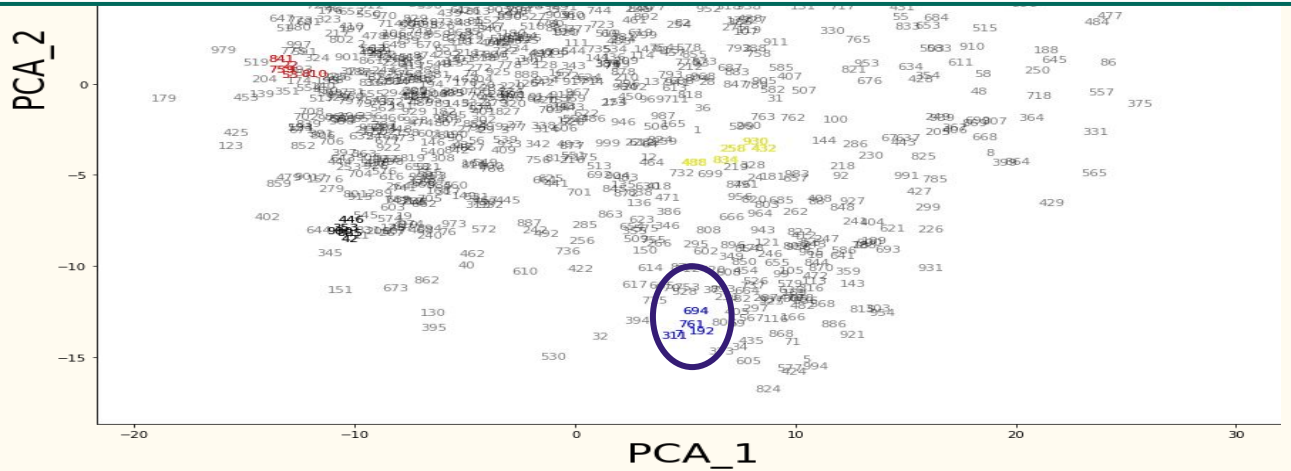
At the end of the pre-trained ResNet50 after global max-pooling , we get a vector of length 2048 Which is our required image feature vector. So we exclude the very last layer.





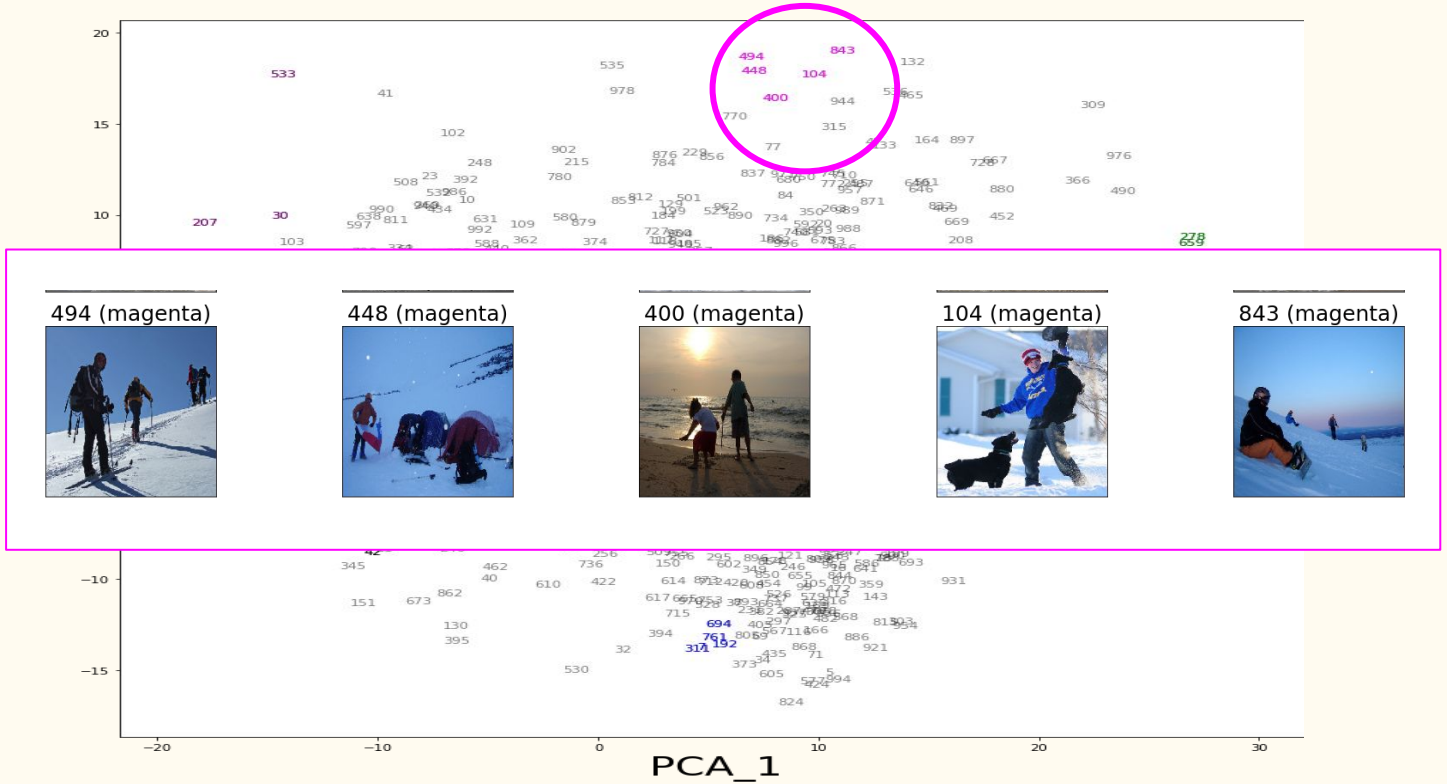
How nice the Feature vector are to collect information from the images?

Here i use PCA to plot 1000 feature vector of dimension 2048 into 2 dimension



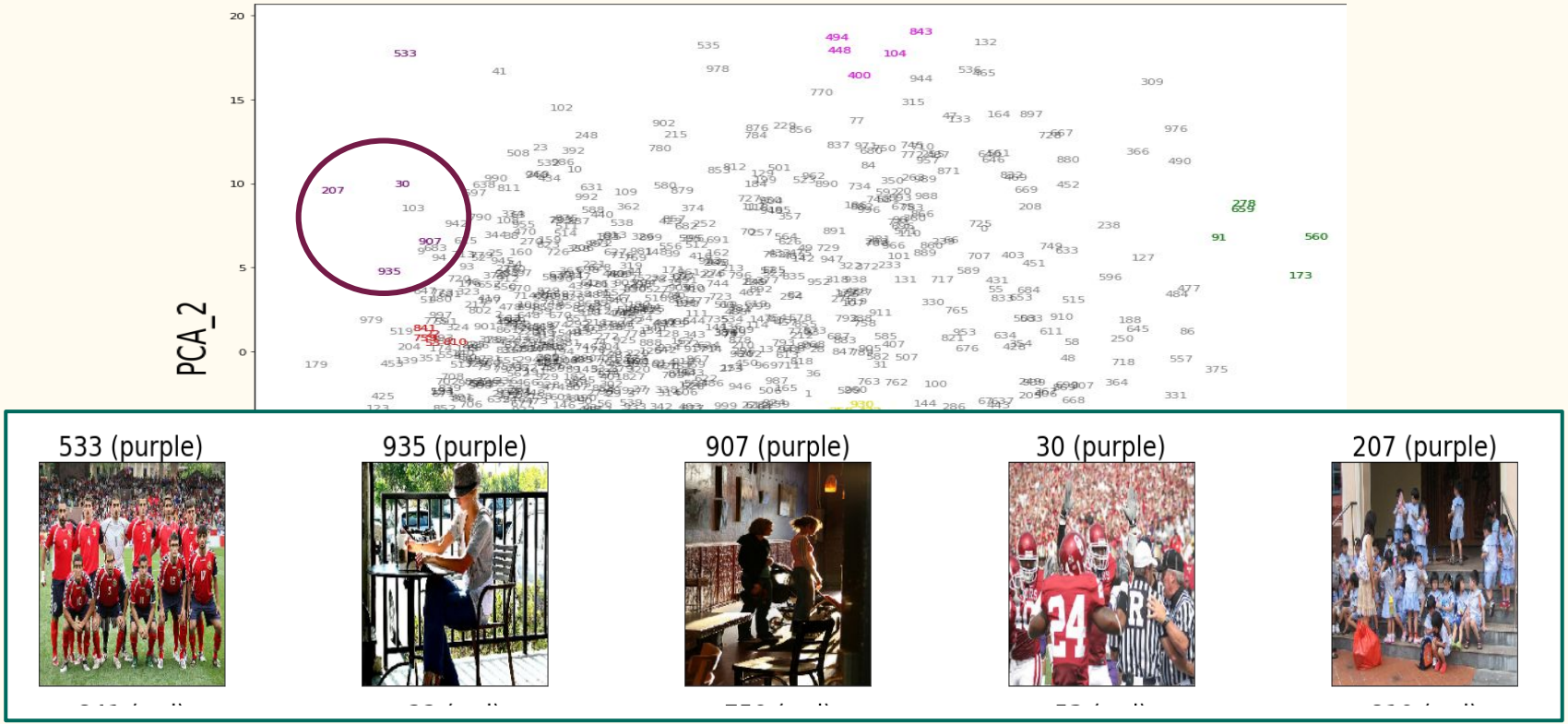
How nice the Feature vector are to collect information from the images?

Here i use PCA to plot 1000 feature vector of dimension 2048 into 2 dimension



How nice the Feature vector are to collect information from the images?

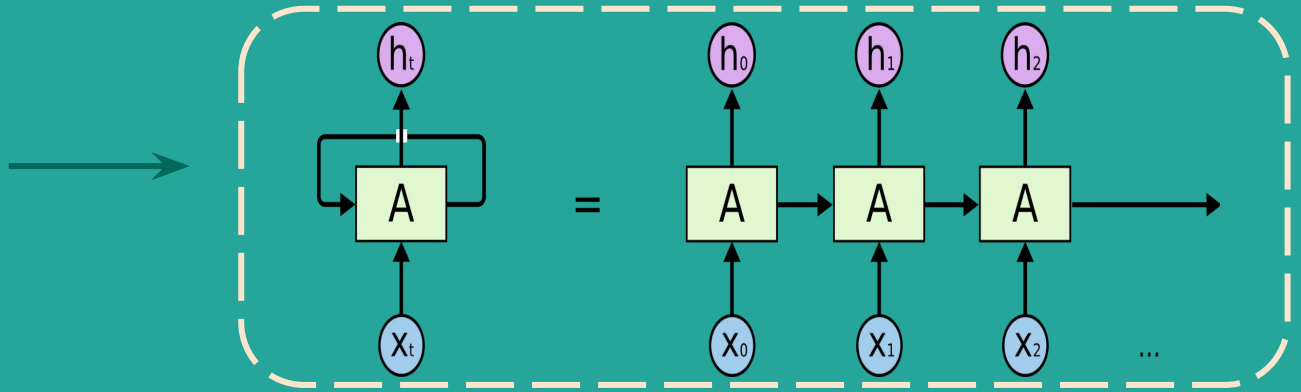
Here I use PCA to plot 1000 feature vector of dimension 2048 into 2 dimension

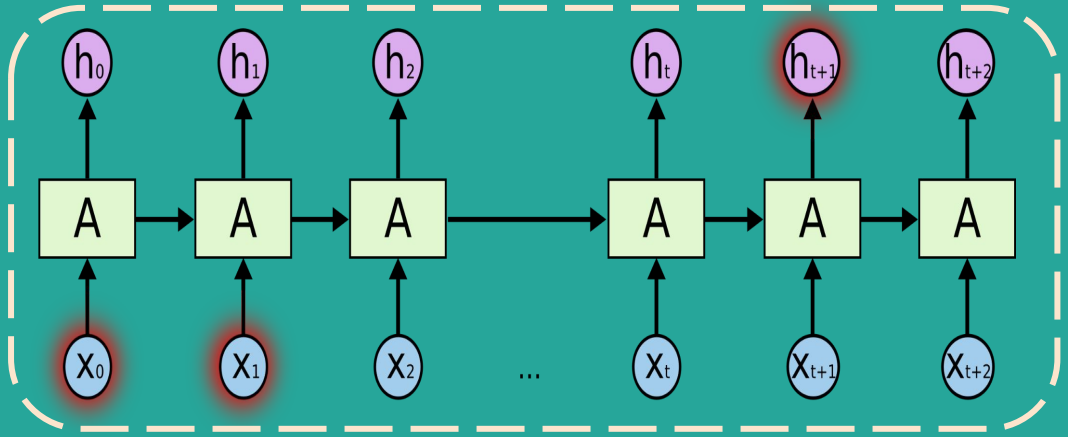


5. LSTM model

A little bit about RNN

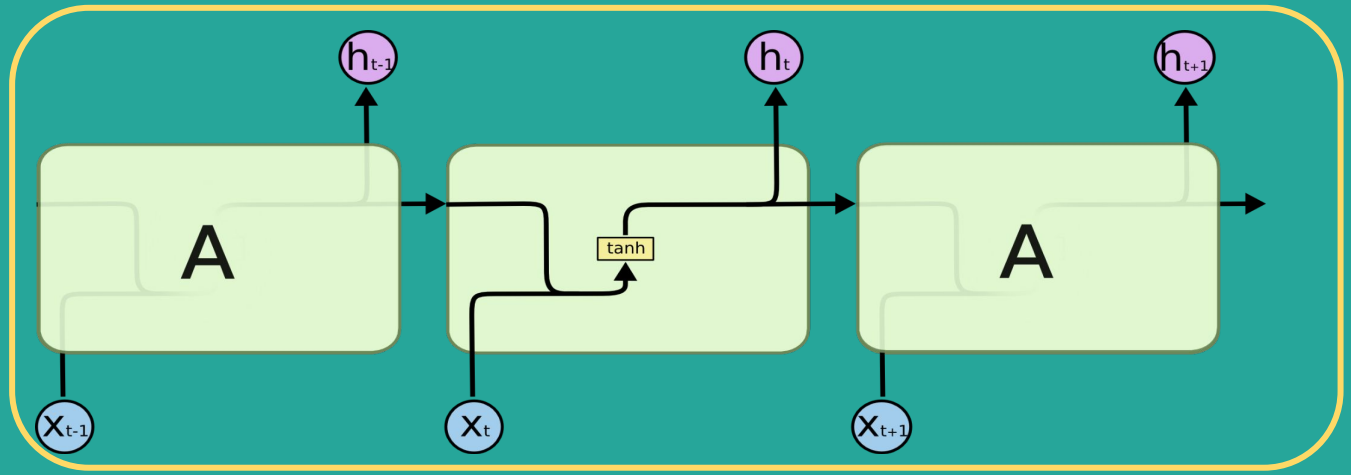
The simple representation of an RNN is such:





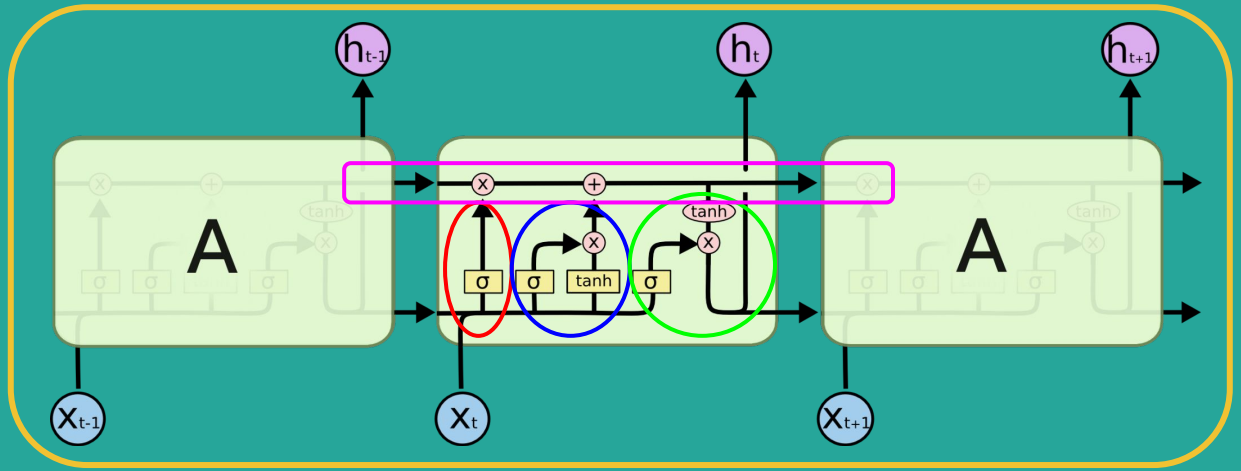
For long term dependencies it fails to keep contains or memory through out

LSTM framework



In RNN we pass the information from the previous input ones but here only one non-linear layer is present.

But in LSTM we have more nonlinear layer . There clearly four gates can be found out of which forgate gate and memory are the interesting and important ones.



How LSTM is being used !

I

IMAGE

CNN

Feature vector

h

LSTM

LSTM

LSTM

LSTM

h_0

h_1

h_2

h_3

h_{k-1}

A

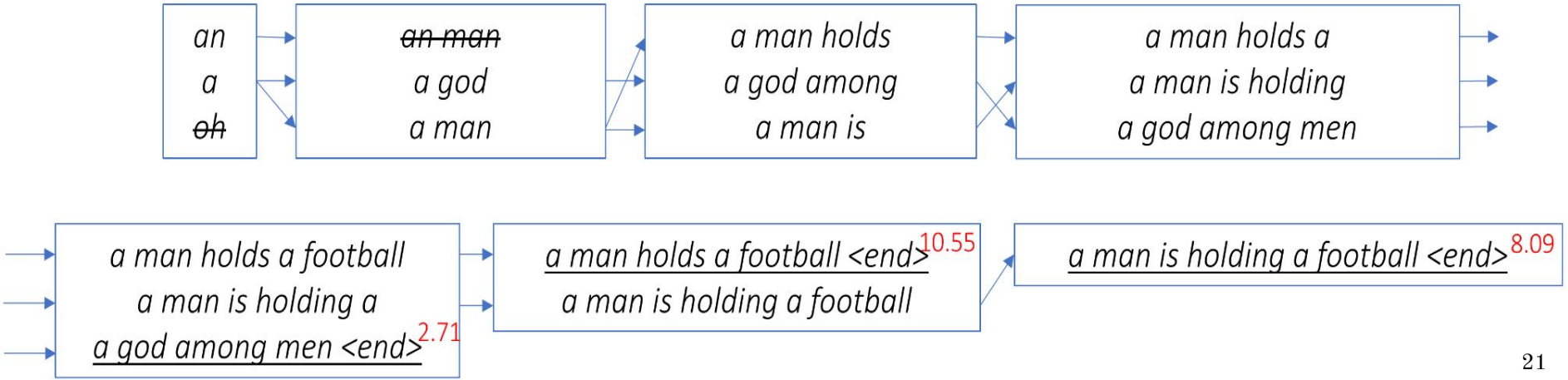
bird

flying

water

Beam Search with k = 3

Choose top 3 sequences at each decode step.
Some sequences fail early.
Choose the sequence with the highest score after all 3 chains complete.

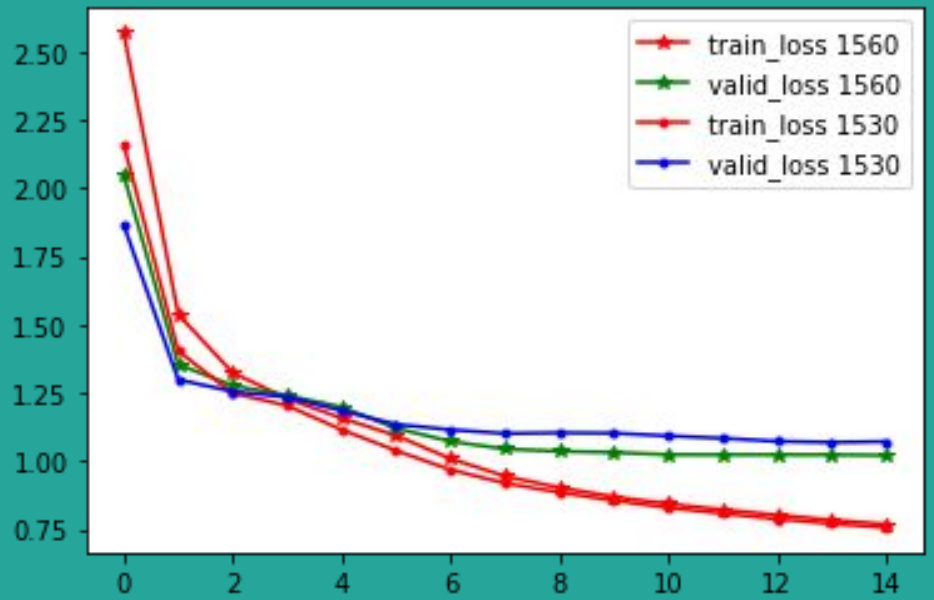


	Xi		Yi
i	Image feature vector	Partial Caption	Target word
1	Image_1	[9, 0, 0 …., 0]	10
2	Image_1	[9, 10, 0, 0 …., 0]	1
3	Image_1	[9, 10, 1, 0, 0 …., 0]	2
4	Image_1	[9, 10, 1, 2, 0, 0 …., 0]	8
5	Image_1	[9, 10, 1, 2, 8, 0, 0 …., 0]	6
6	Image_1	[9, 10, 1, 2, 8, 6, 0, 0 …., 0]	4
7	Image_1	[9, 10, 1, 2, 8, 6, 4, 0, 0 …., 0]	3
8	Image_2	[9, 0, 0 …., 0]	10
9	Image_2	[9, 10, 0, 0 …., 0]	12
10	Image_2	[9, 10, 12, 0, 0 …., 0]	2
11	Image_2	[9, 10, 12, 2, 0, 0 …., 0]	5
12	Image_2	[9, 10, 12, 2, 5, 0, 0 …., 0]	11
13	Image_2	[9, 10, 12, 2, 5, 11, 0, 0 …., 0]	6
14	Image_2	[9, 10, 12, 2, 5, 11, 6, 0, 0 …., 0]	7
15	Image_2	[9, 10, 12, 2, 5, 11, 6, 7, 0, 0 …., 0]	3

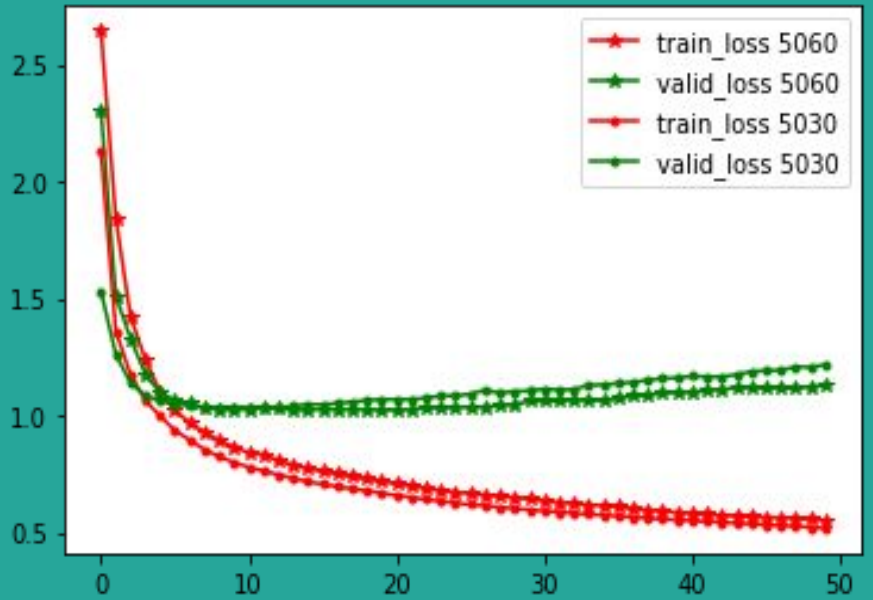
Appending zeros to each sequence to make them all of same length 34

Training results

For 15 epochs with different bath size



For 15 epochs with different bath size



Since we're generating a sequence of words, we use `CrossEntropyLoss`. You only need to submit the raw scores from the final layer in the Decoder, and the loss function will perform the softmax and log operations.

Evaluation with BLEU score

What is bleu score:

BLEU is language independent ,Easy to understand , It is easy to compute.
It lies between [0,1]. Higher the score better the quality of caption

$$\text{modified ngram precision} = \frac{\text{max number of times ngram occurs in referance}}{\text{total number of ngrams in hypothesis}}$$

Some example : To get an idea about the good score I have some example. You can see BLEU does not get the exact meaning rather it plays with n-grams only.

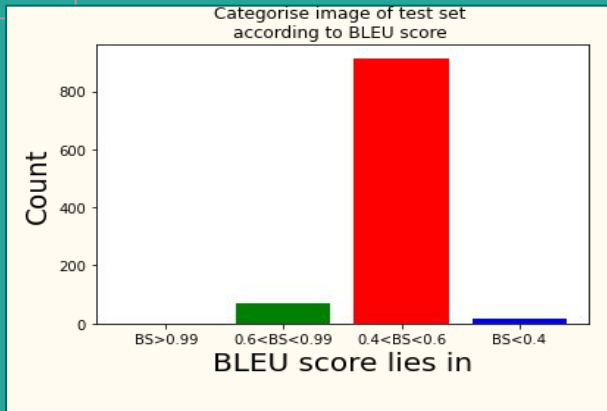
```
reference  = "I like dog  "
hypothesis1 = "I like dog"    1.0
hypothesis2 = "i like dog i like dog"    0.47
hypothesis3 = "dog is liked by me"    0.67
hypothesis4 = "dog is liked"    0.76
```

Comparison of BLEU score

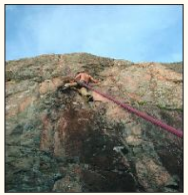
Epoch	Batch size	Method	BLEU score
15	60	Greedy	0.462
,,	,,	Beam k=1	0.475
,,	,,	Beam k=2	0.488
,,	,,	Beam k=3	0.487
,,	,,	Beam k=4	0.485
,,	,,		Average: 0.479

Epoch: 15 , batch size= 30 , average bleu score= 0.482

Epoch: 50, batch size= 60 , average bleu score= 0.486



Some Captions with high BLEU score



true: A young white man be climb a mountain with a rope as a guide .
true: A shirtless man climb up a steep mountain .
true: A man be climb the side of a mountain .
true: A man climb a mountain .
true: A man climb a mountain .

pred: A man be crossing a mountainside .
BLEU score: 0.619

true: A parasailer be skip across the water .
true: A person in a parachute slide across the water .
true: A person hang from a parachute make a splash as their body hit the water .
true: A parasailer splash along the surface of a lake .
true: a man crash into the water with his parachute .

pred: A biker be play in a creek .
BLEU score: 0.615

true: Boy in green sweater with white stripe on sleeve stand on wood plank floor with body of water in the background .
true: A young child stand on a wood dock next to the water .
true: A young boy hold onto a blue handle on a pier .
true: A little boy be smile for the camera on a playground
true: A child in a black shirt stand on a wooden dock near a picnic table .

pred: A boy push a boy in a boat .
BLEU score: 0.65

true: Bmx biker Jump off of ramp
true: Bike rider jump obstacle .
true: A man ride a yellow bike over a ramp while others watch .
true: A man on a bike execute a jump as part of a competition while a crowd watch .
true: A man be do trick on a bicycle on ramp in front of a crowd .

pred: A man be jump off a boat on a boat .
BLEU score: 0.639

Some Captions with low BLEU score

true: People stand at fence watch motor vehicle in field
true: A line of spectator at a race .
true: A line of person stare at vehicle on the dirt track
true: A group of person on the sideline of an Atv race .
true: A crowd watch a dirt bike race

pred: A group of person be walk along a flight of school with a flock of a flock of a flock of a flock of ten a black pylon .
BLEU score: 0.398

true: Several person ride up an escalator .
true: People travel up an elevator together .
true: People ride up an escalator .
true: person go up an escalator .
true: People be stand on an escalator move up .

pred: A man in a white shirt be sit on a white bench outside of a building .
BLEU score: 0.346

true: Two kid play on a street
true: Two child have expression of happiness as they reach for something out of sight .
true: Two baby in sweater play with a toy .
true: One child with a colorful toy and another child reach up .
true: One child reach up for something while another stand beside him .

pred: A child in a colorful fleece and a gray and a man in a striped and a blue and a blue star .
BLEU score: 0.385

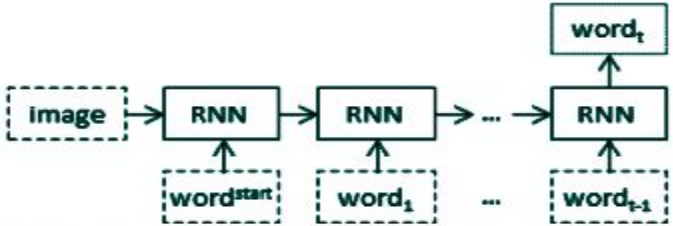
true: Two young child and an old man play with sparkler .
true: Two young boy play with sparkler .
true: Two small child be twirl sparkle rope .
true: Little kid play with sparkler at night .
true: child play with large hoop .

pred: A person dress in a red dress be fire in the grass .
BLEU score: 0.372

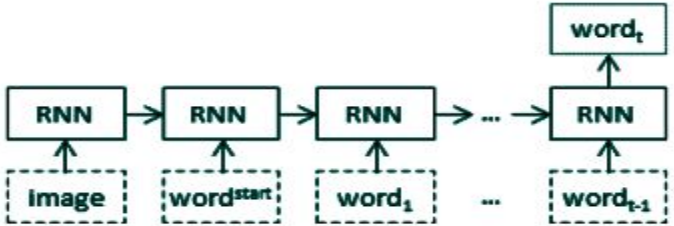
true: A girl be sit at a counter between the bucket of flower and cardboard box .
true: A woman be stand at a counter that be hold bucket of flower .
true: A woman dress in black stand at a green counter with barrel of flower place on it .
true: A person look to the right stand between plant in bucket and a banana box .
true: An oriental florist arrange flower

pred: A man stand on a wall with a large large large large large large large large large large large large large large large large large large large display .
BLEU score: 0.379

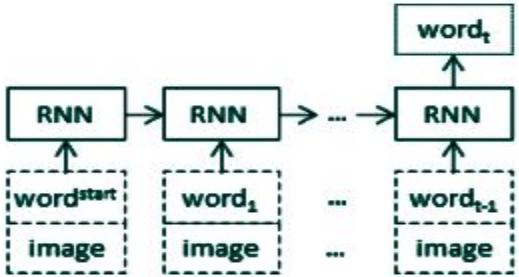
10. Further work to do



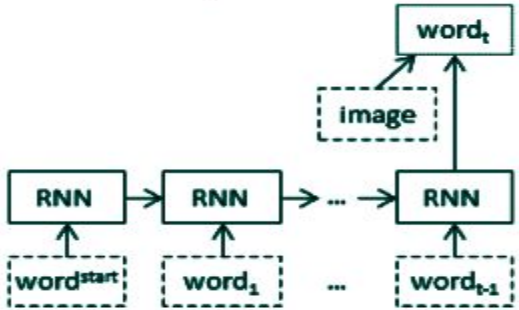
(a) Init-inject: The image vector is used as an initial hidden state vector for the RNN.



(b) Pre-inject: The image vector is used as a first word in the prefix.

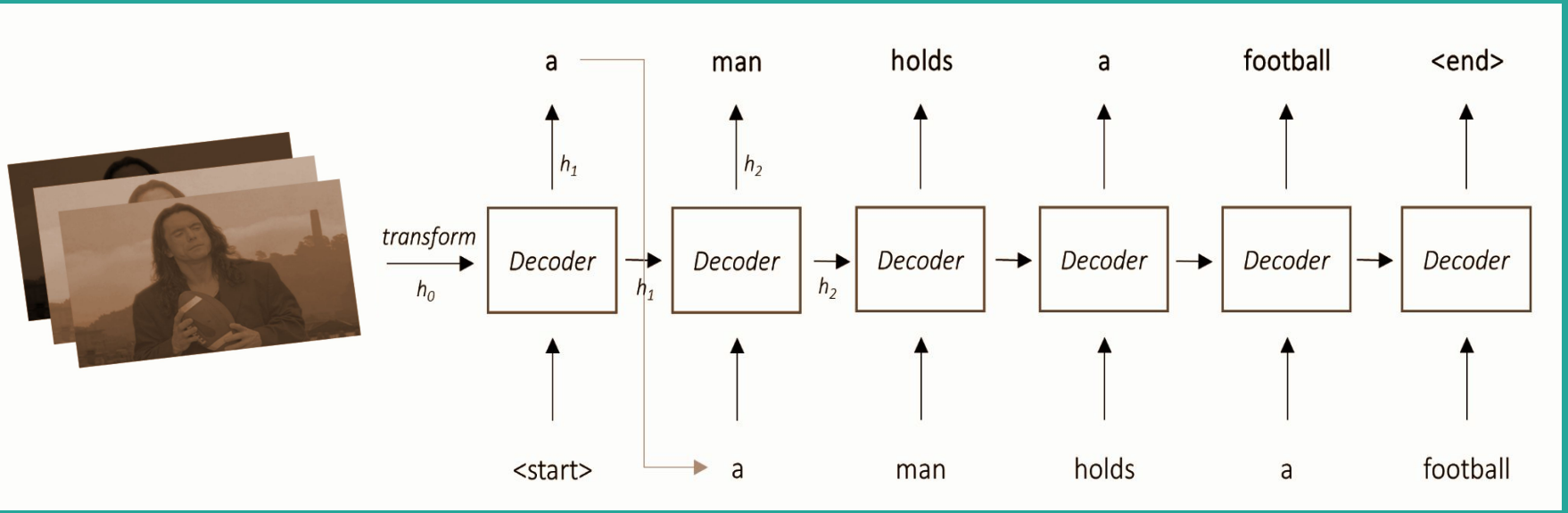


(c) Par-inject: The RNN accepts two inputs at once in every time step: a word and an image.

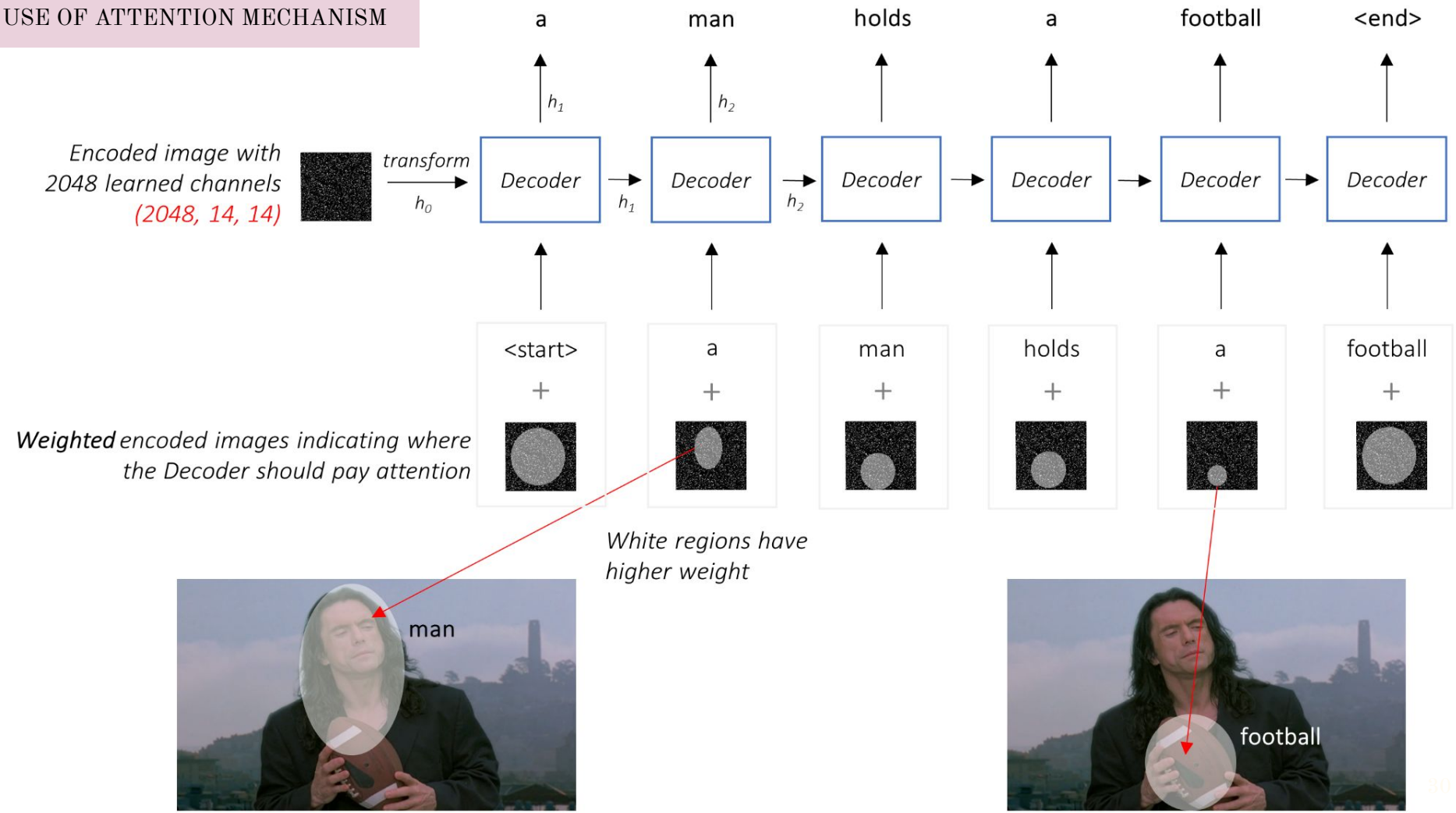


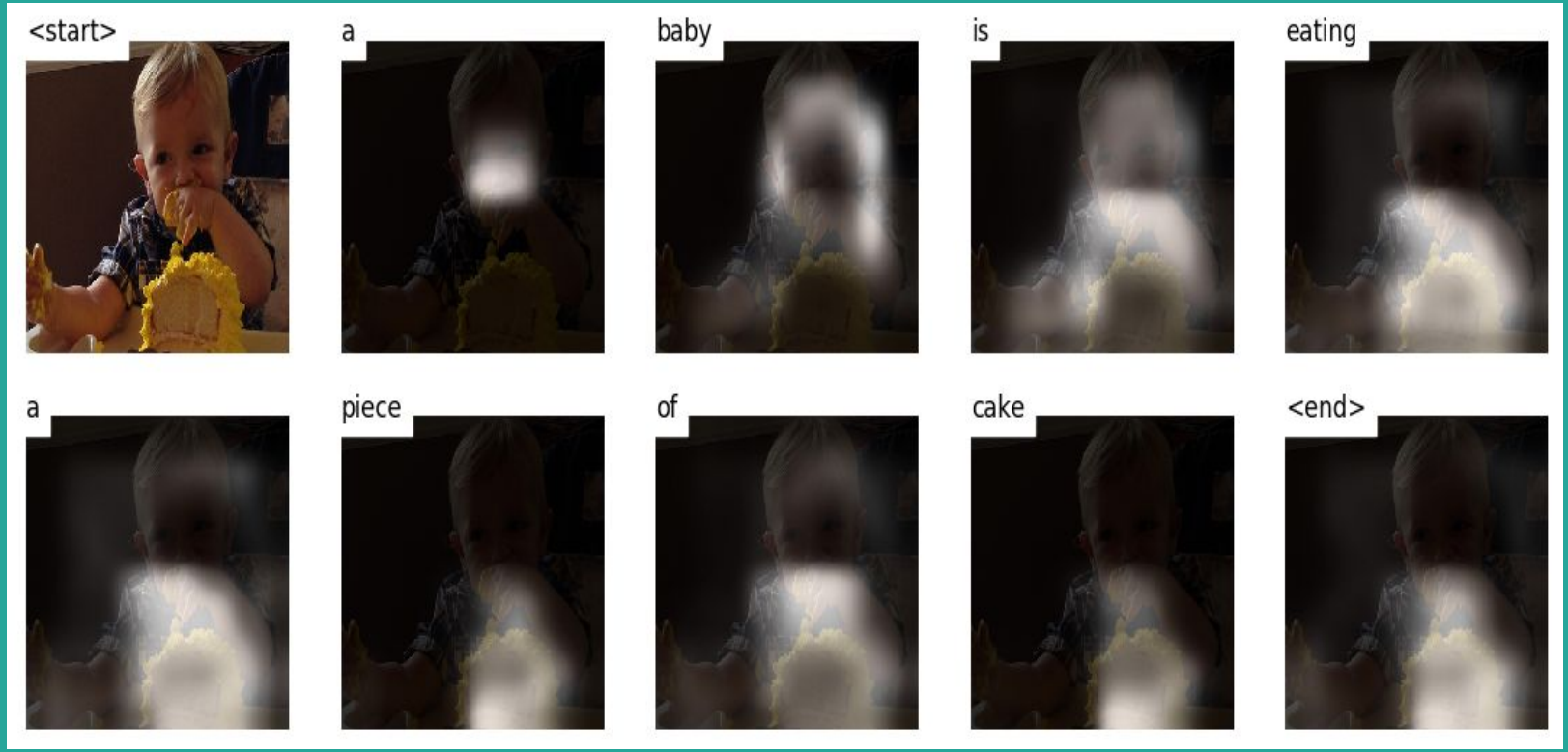
(d) Merge: The image vector is merged with the prefix outside of the RNN.

Init-inject :


a
man
holds
a
football
<end>
h_1
h_2
transform
h_0
Decoder
Decoder
Decoder
Decoder
Decoder
Decoder
h_1
h_2
<start>
a
man
holds
a
football

USE OF ATTENTION MECHANISM




<start>
a
baby
is
eating
a
piece
of
cake
<end>

References

- https://medium.com/@raman.shinde15/image-captioning-with-flickr8k-dataset-bleu-4bcba0b52926#:~:text=BLEU%20stands%20for%20Bilingual%20Evaluation,quality%20of%20our%20generated%20caption.&text=It%20is%20easy%20to%20compute.
- https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1
- https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning
- https://github.com/yudi09/pytorch-image-captioning
- https://github.com/yurayli/image-caption-pytorch
- https://github.com/tatwan/image-captioning-pytorch
- https://github.com/ruotianluo/ImageCaptioning.pytorch/tree/bart
- https://github.com/MITESHPUTHRANNEU/Image-Caption-Generator
- https://github.com/nalbert9/Image-Captioning

THANK YOU

To our departmental head Resp. Swami Dhyangamyananda and our instructor Sri Sujoy Biswas .