Image Captioning

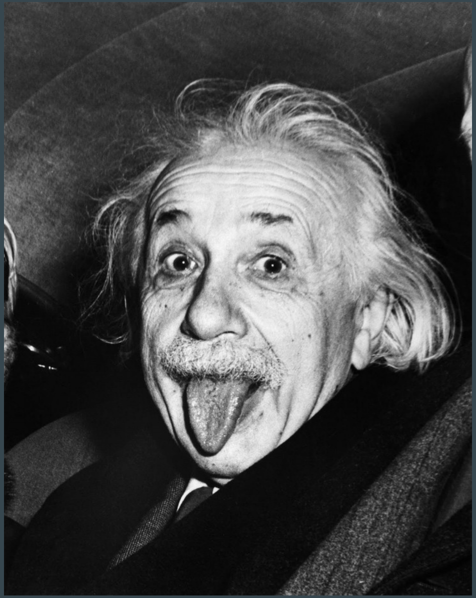
Basic Definitions

Computer Vision

Natural Language Processing

Image Captioning

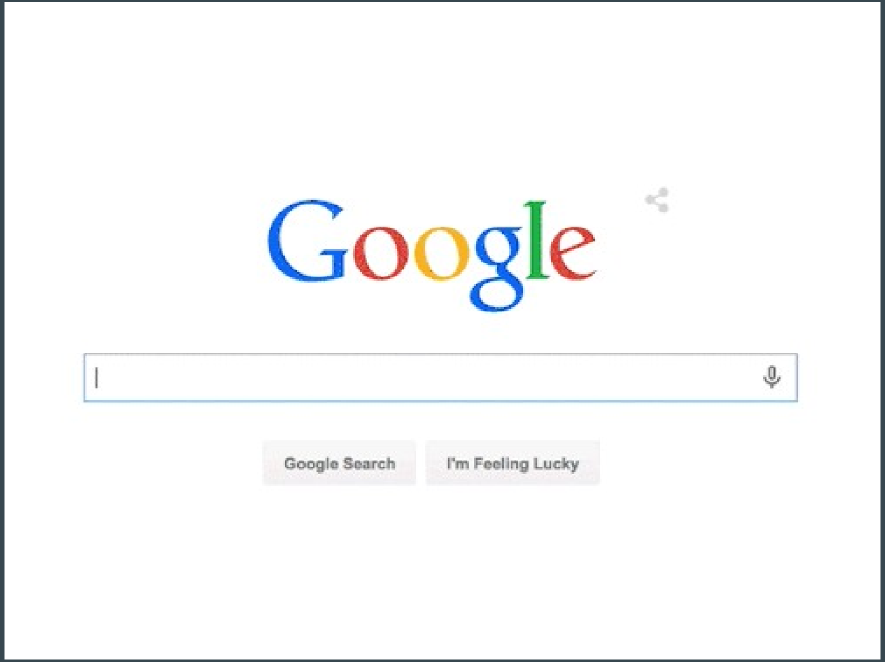
Let me tell you about myself.



::Motivations::

Google Image Search Results

Reverse searching images using their captions



Aid to Blind

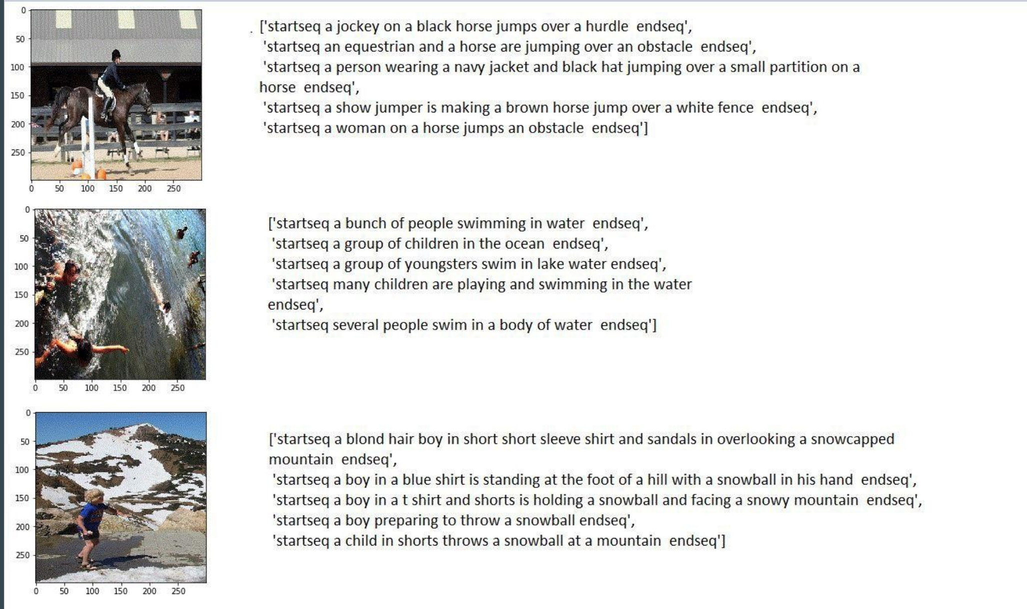
Describing
Surroundings

Self Driving Vehicles

Logging Journeys

Flickr8K

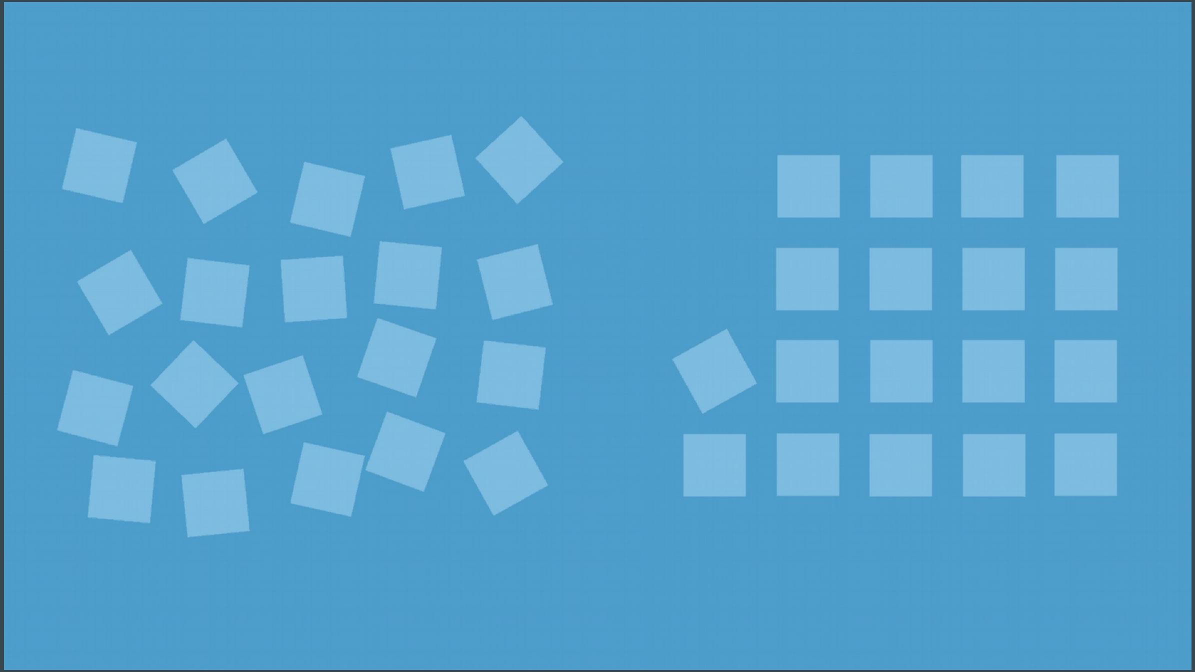
More than 8000 images with 5 captions per image



```
['startseq a jockey on a black horse jumps over a hurdle  endseq',
 'startseq an equestrian and a horse are jumping over an obstacle  endseq',
 'startseq a person wearing a navy jacket and black hat jumping over a small partition on a
horse  endseq',
 'startseq a show jumper is making a brown horse jump over a white fence  endseq',
 'startseq a woman on a horse jumps an obstacle  endseq']
```

```
['startseq a bunch of people swimming in water  endseq',
 'startseq a group of children in the ocean  endseq',
 'startseq a group of youngsters swim in lake water endseq',
 'startseq many children are playing and swimming in the water
endseq',
 'startseq several people swim in a body of water  endseq']
```

```
['startseq a blond hair boy in short short sleeve shirt and sandals in overlooking a snowcapped
mountain  endseq',
 'startseq a boy in a blue shirt is standing at the foot of a hill with a snowball in his hand  endseq',
 'startseq a boy in a t shirt and shorts is holding a snowball and facing a snowy mountain  endseq',
 'startseq a boy preparing to throw a snowball endseq',
 'startseq a child in shorts throws a snowball at a mountain  endseq']
```



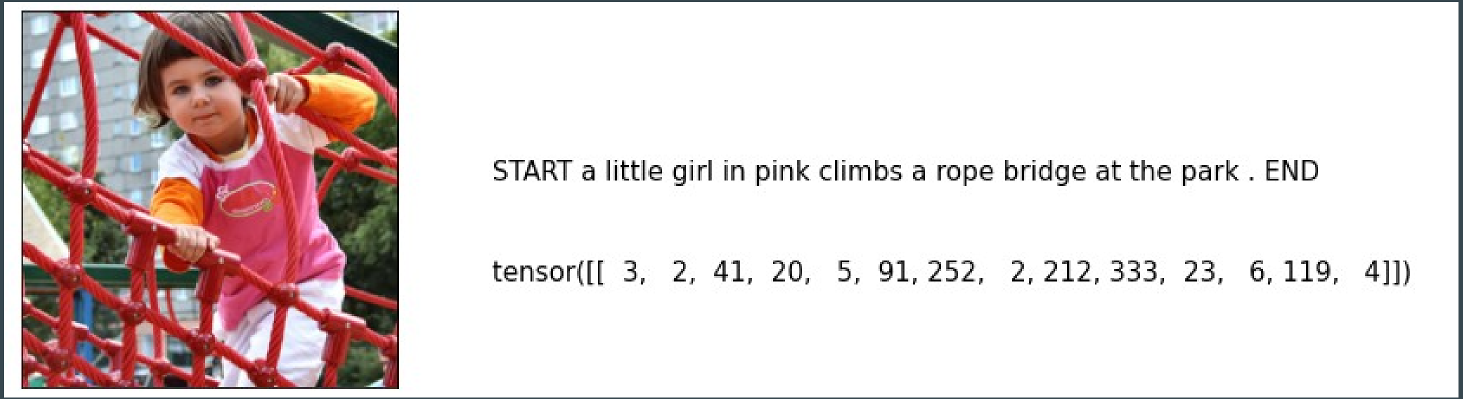
Preprocessing

Processing captions

- Turned into lowercase.
- Numbers are removed.
- Start/End tags are added

Tokenization

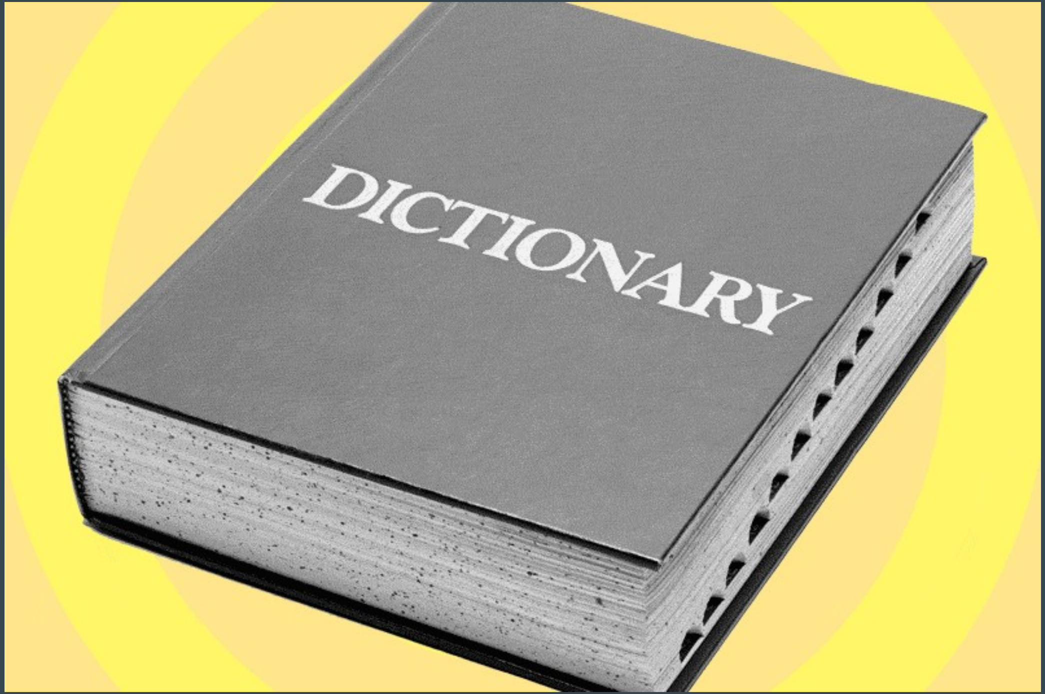
- Assigning numbers to unique words.



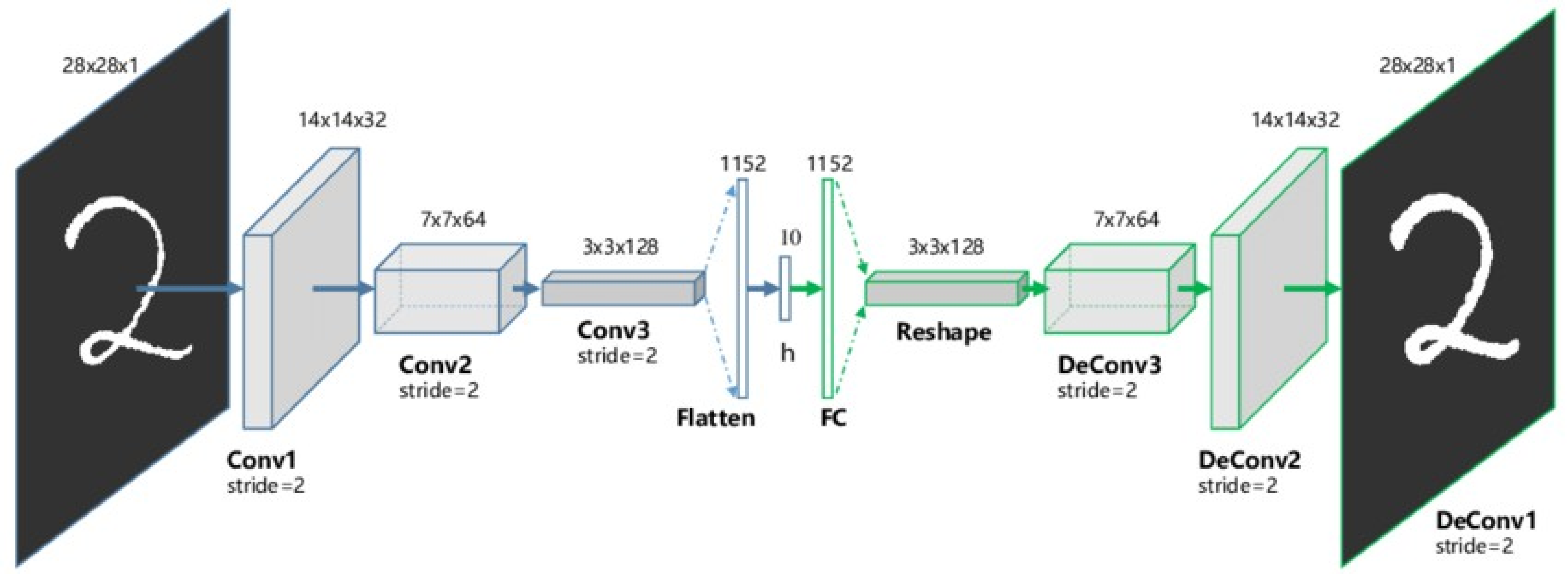
```
START a little girl in pink climbs a rope bridge at the park . END

tensor([[  3,   2,  41,  20,   5,  91, 252,   2, 212, 333,  23,   6, 119,   4]])
```

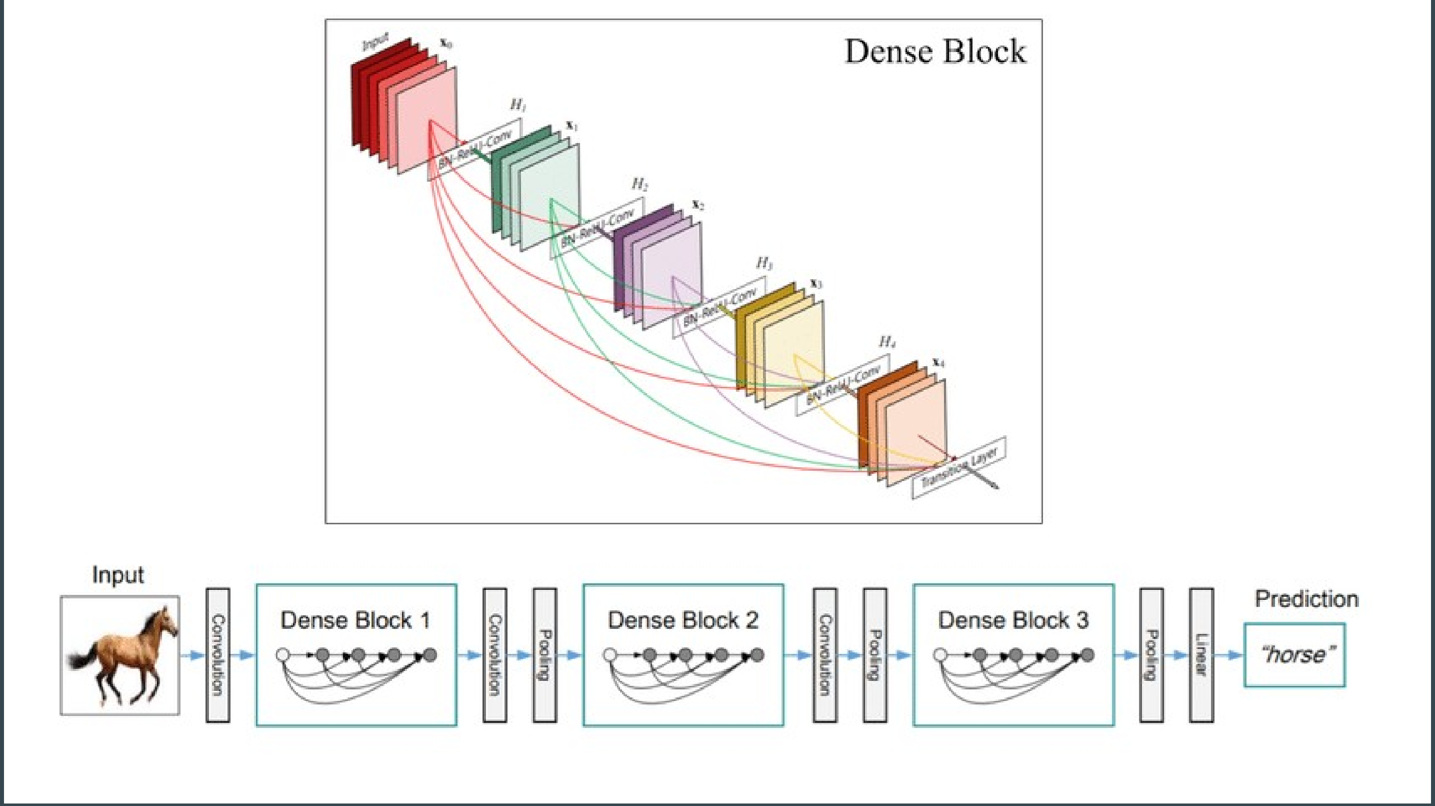
Architecture



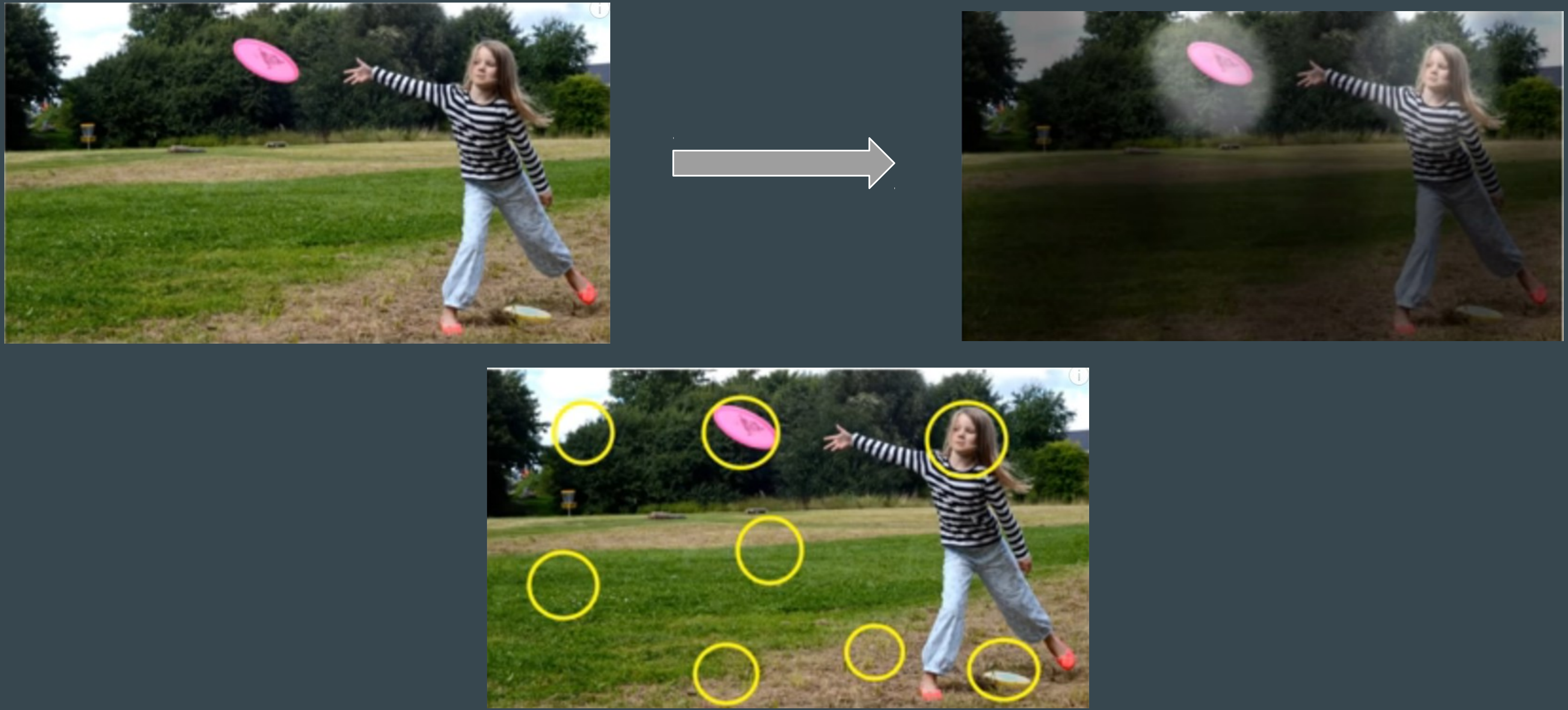
AutoEncoder



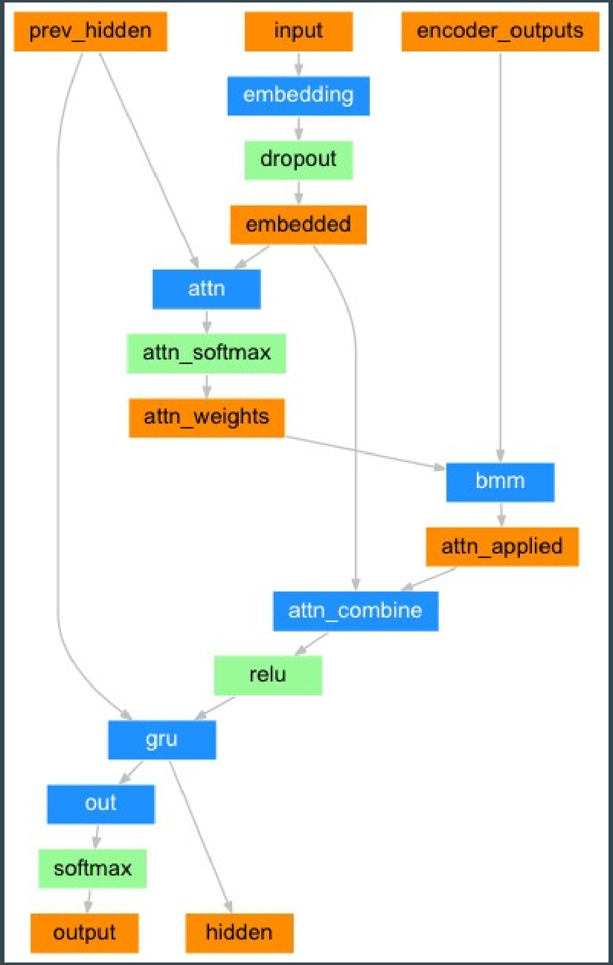
Encoder - DenseNet



Attention

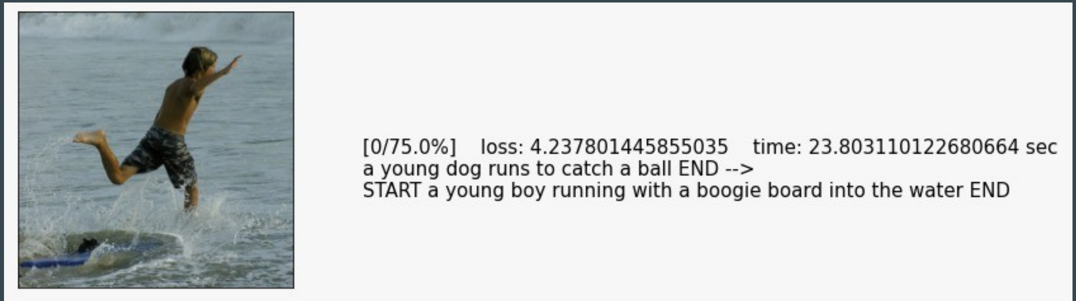


Decoder - GRU with Attention

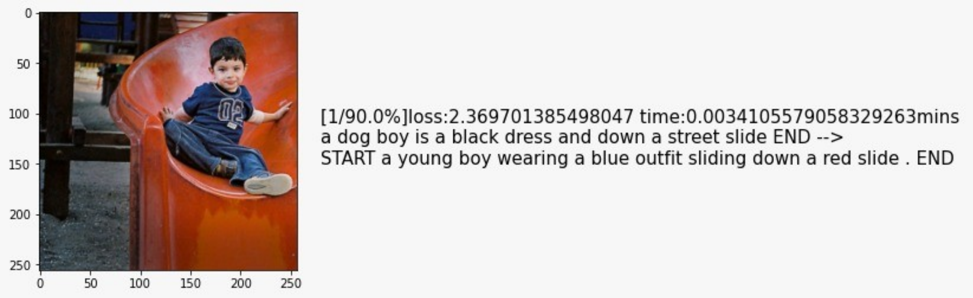


<code/>

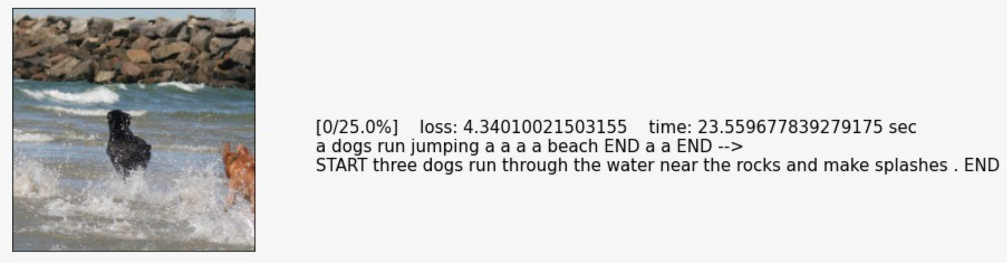
Some results from our Model



```
[0/75.0%]    loss: 4.237801445855035    time: 23.803110122680664 sec
a young dog runs to catch a ball END -->
START a young boy running with a boogie board into the water END
```



```
[1/90.0%]loss:2.369701385498047 time:0.0034105579058329263mins
a dog boy is a black dress and down a street slide END -->
START a young boy wearing a blue outfit sliding down a red slide . END
```



```
[0/25.0%]    loss: 4.34010021503155    time: 23.559677839279175 sec
a dogs run jumping a a a a beach END a a END -->
START three dogs run through the water near the rocks and make splashes . END
```

Conclusion

- Results may differ
- Accuracy ~
  - Larger Dataset
  - Model architecture
  - Hyperparameter Tuning
  - Using Cross Validation