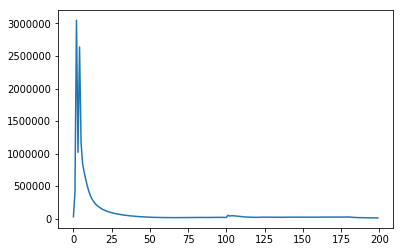
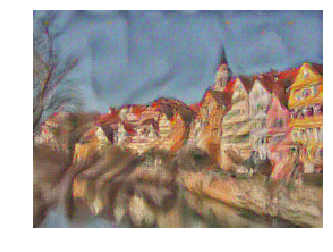
1.5. validation accuracy:

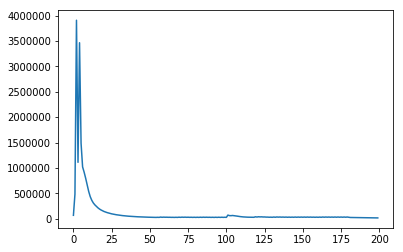
|  |  |  |
| --- | --- | --- |
|  | Finetune | freeze |
| On pretrain | 0.3529 | 0.3529 |
| After training | 0.9281 | 0.9477 |

2.4

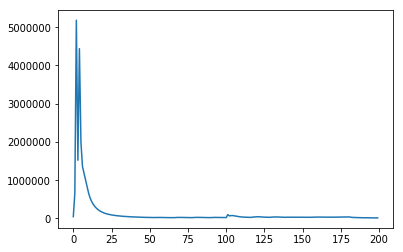
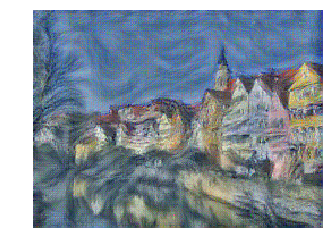
Composition VII + Tubingen:



Scream + Tubingen:



Starry Night + Tubingen:



3.2 Let denote a vector with the i-th element equals to .

\* means element wise product.

. So,

So,

So,

So,

So,

3.4 For each timestep:

Recursively define:

.

So at timestep t, Π is element wise product here.

So

4.2

4.4

For each timestep:

Recursively define:

.

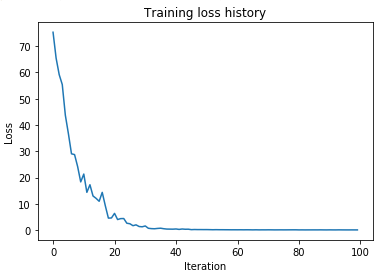
So at timestep t

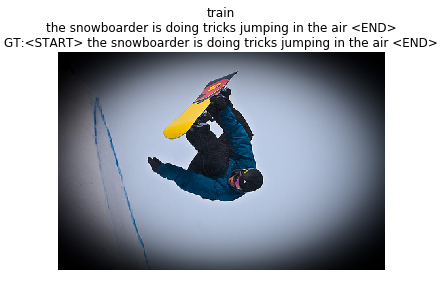
So at timestep t, Π is element wise product here.

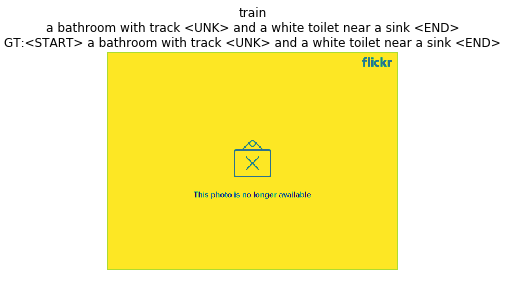
5.4.

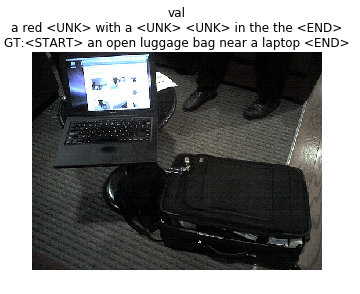
In the prediction, I am ***not*** including <START> inside the predicted caption, because it is only initial state. And according to the code rnn.py, “The first element of captions should be the first sampled word, not the <START> token.”

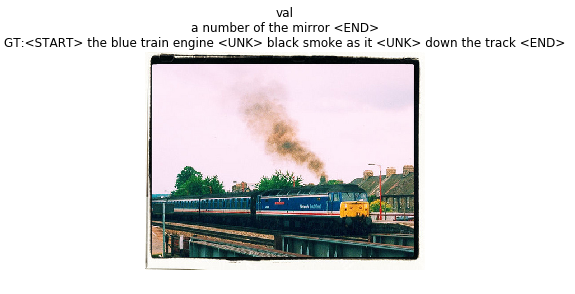
RNN:



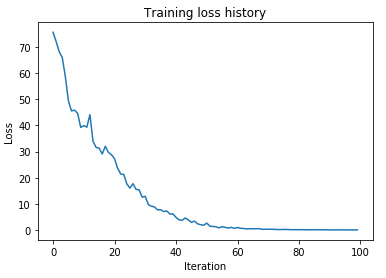


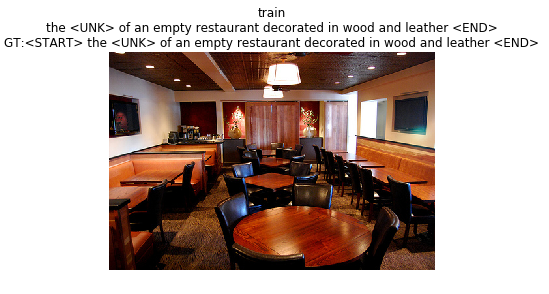


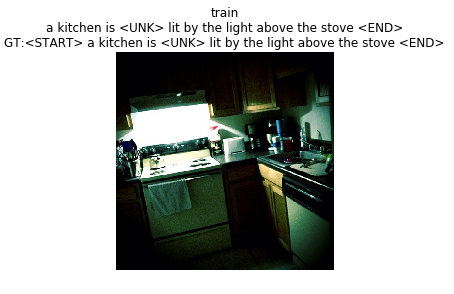


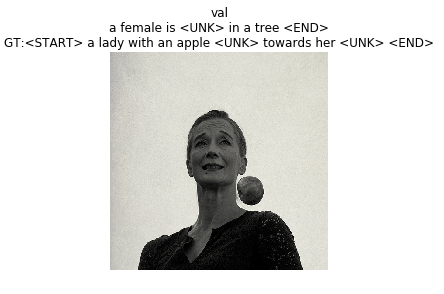


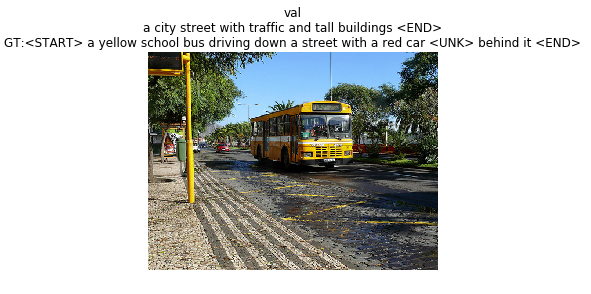
LSTM:











5.5

I got a better model in image\_captioning\_better.py. With BLEU validation scores higher than 0.25 (in 5.4, the BLEU validation scores is around 0.15).

1. test performances:
2. 92.14%
3. 94.76%
4. 94.84%
5. 95.88%
6. 92.98%