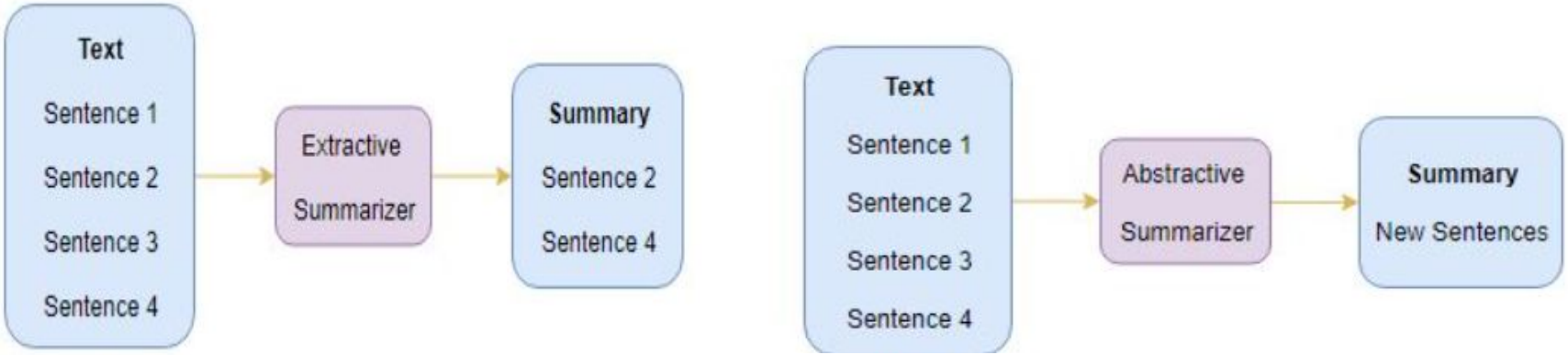
Assignment 2
Deep Learning

Text Summarization

Parth Saboo 2019A4PS0457P
Udbhav Tripathi 2019A4PS0323P

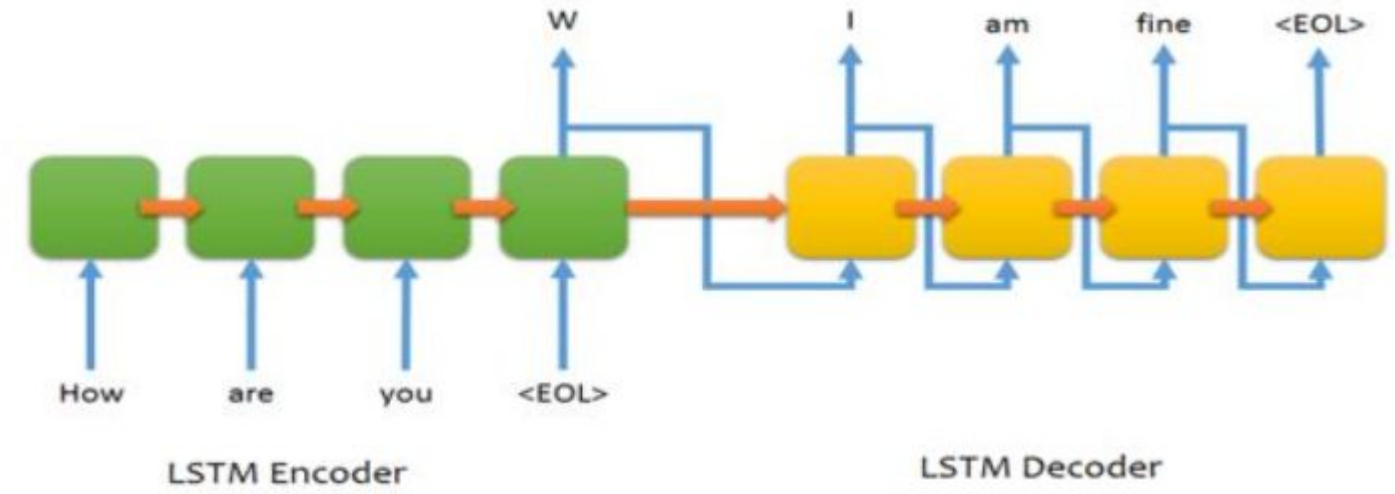
Extractive and Abstractive summarization



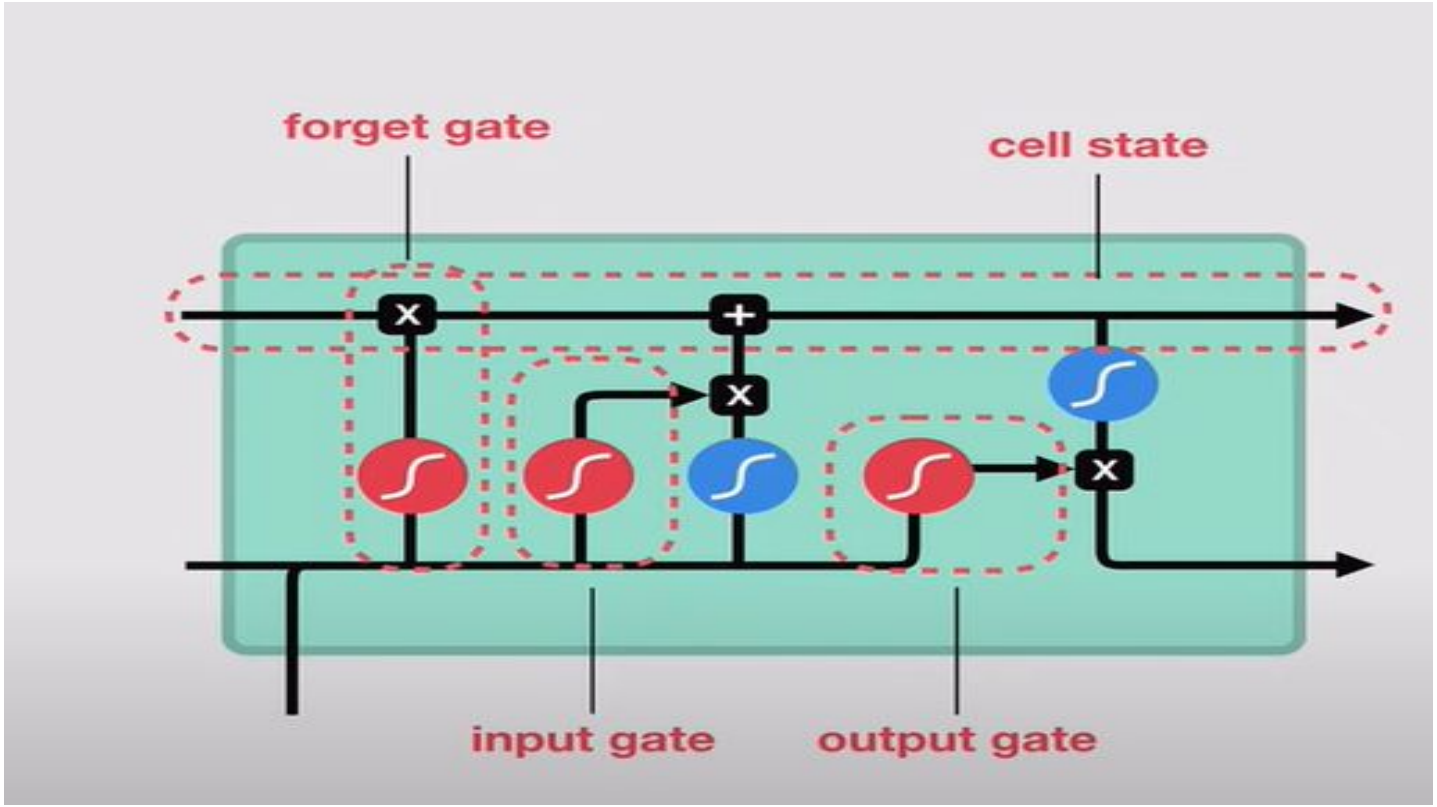
Seq2Seq Modelling

There are two major components of a Seq2Seq model:

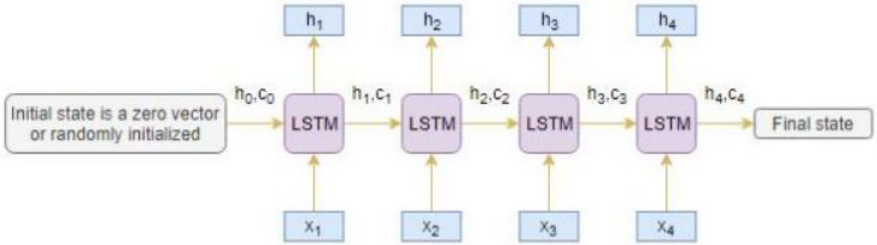
- Encoder
- Decoder

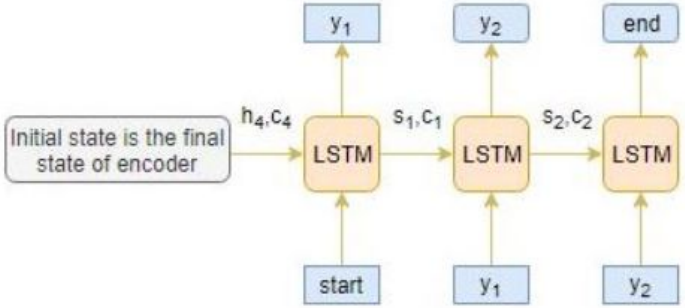


LSTM Architecture



Training and Inference Phase

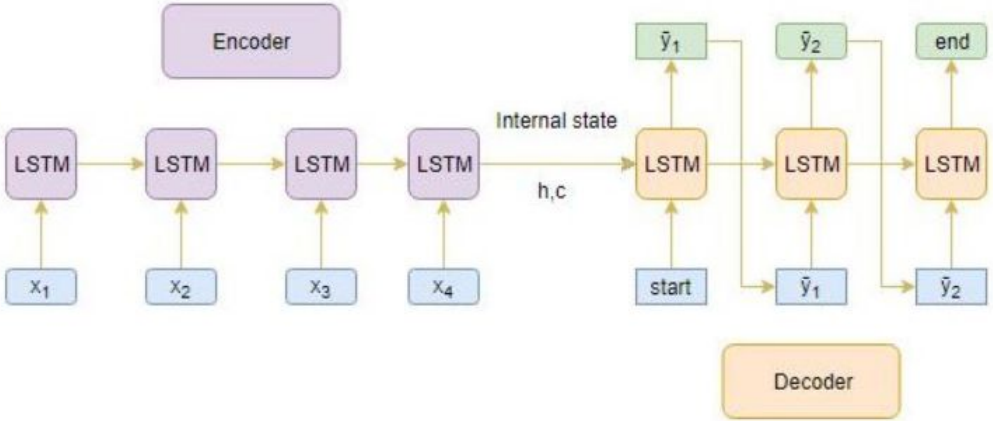




An Encoder Long Short Term Memory model (LSTM) reads the entire input sequence wherein, at each timestep, one word is fed into the encoder. It then processes the information at every timestep and captures the contextual information present in the input sequence.

The decoder is also an LSTM network which reads the entire target sequence word-by-word and predicts the same sequence offset by one timestep. The decoder is trained to predict the next word in the sequence given the previous word

Training and Inference Phase

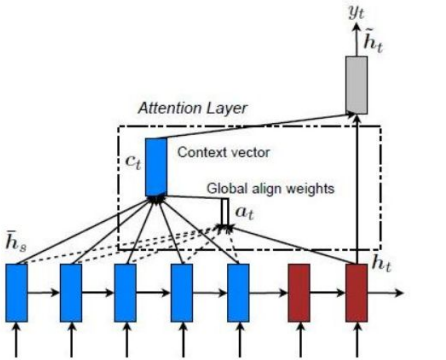


1. Encode the entire input sequence and initialize the decoder with internal states of the encoder
2. Pass <start> token as an input to the decoder
3. Run the decoder for one timestep with the internal states
4. The output will be the probability for the next word. The word with the maximum probability will be selected
5. Pass the sampled word as an input to the decoder in the next timestep and update the internal states with the current time step
6. Repeat steps 3 – 5 until we generate <end> token or hit the maximum length of the target sequence

Limitations of the Encoder – Decoder Architecture

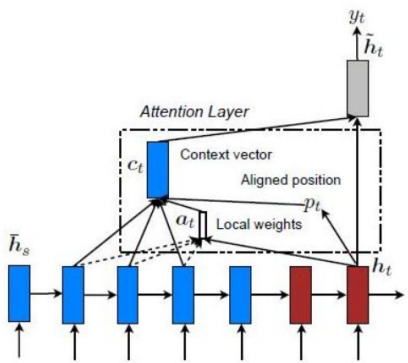
- The encoder converts the entire input sequence into a fixed length vector and then the decoder predicts the output sequence. This works only for short sequences since the decoder is looking at the entire input sequence for the prediction
- Here comes the problem with long sequences. It is difficult for the encoder to memorize long sequences into a fixed length vector

Attention Mechanism



Global Attention

Instead of looking at all the words in the source sequence, we can increase the importance of specific parts of the source sequence that result in the target sequence



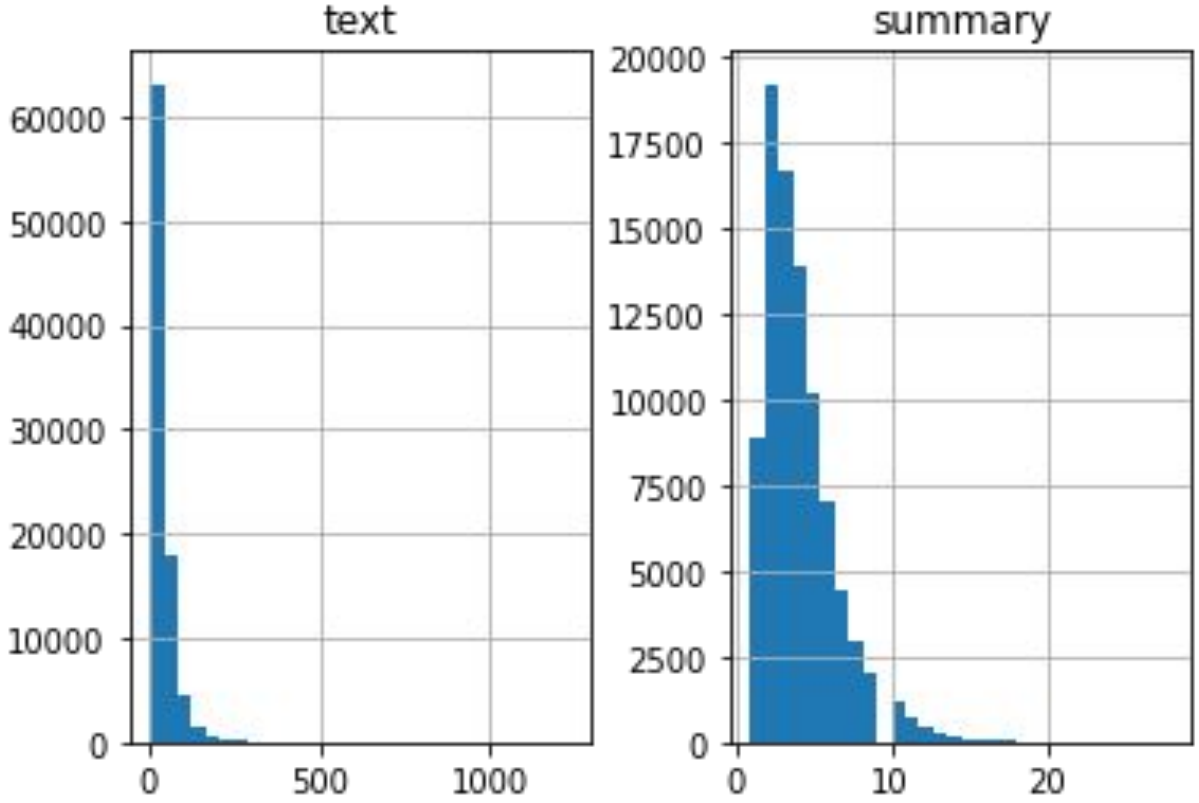
Local Attention

- The alignment score is computed from the source hidden state h_j and target hidden state s_i using the score function. This is given by:

$$e_{ij} = \text{score}(s_i, h_j)$$

- We compute the linear sum of products of the attention weights a_{ij} and hidden states of the encoder h_j to produce the attended context vector (C_i)

Understanding Distribution of Sequences

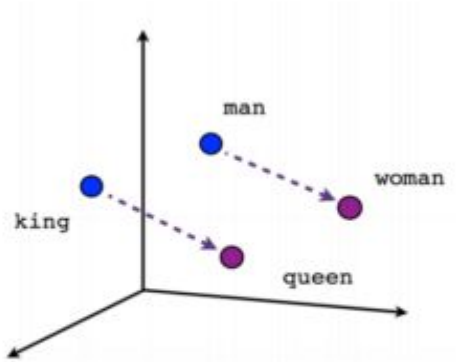


Word Embeddings

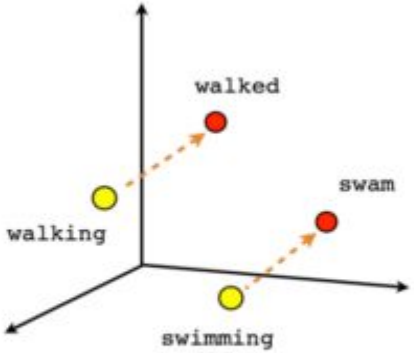
“ You shall know a word by the company it keeps”

- J.R. Firth, "Studies in linguistic analysis", 1957

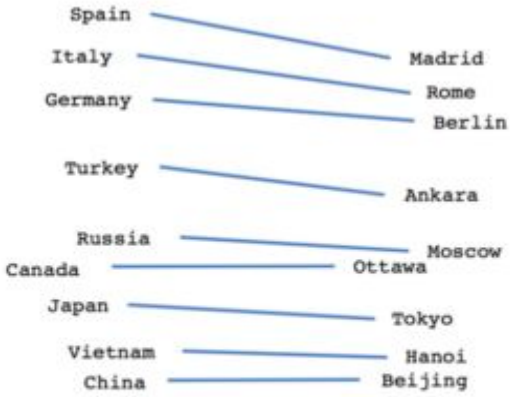
Word Embeddings



Male-Female



Verb tense



Country-Capital

- A relatively low-dimensional space into which you can translate high-dimensional vectors.
- Captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space.

Word Embeddings

```
      dtype=float32),
 'discontent': array([-2.6740e-01,  5.9682e-01,  2.0791e-01, -3.3717e-01, -1.5903e-01,
        4.4978e-01, -7.9726e-01, -1.0069e+00,  1.8970e-01,  1.9601e-01,
       -1.1614e+00,  4.4620e-01, -6.3861e-03, -5.3942e-01, -4.5059e-02,
       -2.9735e-01, -5.9177e-01, -6.9177e-02, -2.6065e-01, -5.2627e-01,
       -2.2953e-01,  1.9007e-01,  2.2029e-01, -2.0479e-01,  1.0551e-02,
       -4.9431e-01,  2.7939e-01,  4.9928e-01,  3.4812e-01, -1.7214e-01,
       -1.0167e-01,  2.6358e-01,  5.1627e-01,  6.2668e-04, -6.4836e-01,
        5.6074e-01,  7.6320e-01, -2.2073e-01, -3.5735e-01, -2.3398e-01,
       -4.5567e-01, -4.2571e-01, -2.8577e-02, -2.1825e-01,  5.0149e-01,
       -5.2477e-01,  9.6253e-01, -5.1857e-02,  5.3464e-01, -3.2355e-01,
       -4.3002e-01, -1.6797e-01, -4.6244e-01,  1.2074e+00,  4.9106e-01,
       -6.0517e-01,  8.1095e-01, -2.7462e-02,  1.0032e-01,  3.8506e-02,
        1.6851e-01, -1.3804e-01,  3.0985e-01,  6.3662e-02,  5.6669e-01,
        3.6601e-01,  4.8901e-01, -1.6636e+00,  7.1084e-01, -2.7856e-02,
        4.2597e-01,  2.4089e-01, -1.2681e+00, -4.2287e-01, -2.2623e-02,
        1.0027e+00,  2.5596e-01,  8.7852e-02, -8.3472e-01,  2.8966e-01,
        2.9416e-01,  1.0485e-01, -1.7026e-01, -2.6958e-01, -1.3832e-01,
       -1.0714e+00, -4.5481e-02,  2.3205e-01, -5.5243e-01, -4.6589e-01,
       -3.6207e-01,  2.7359e-01, -1.3839e+00,  5.9248e-01,  3.2265e-01,
       -2.1726e-01,  1.1852e-01, -3.2317e-01,  1.2290e-01,  1.1880e-01],
      dtype=float32),
```

Input Dataset

	headlines	text
0	upGrad learner switches to career in ML & AI with 90% salary hike	Saurav Kant, an alumnus of upGrad and IIIT-B's PG Program in Machine learning and Artificial Intelligence, was a Sr Systems Engineer at Infosys with almost 5 years of work experience. The program ...
1	Delhi techie wins free food from Swiggy for one year on CRED	Kunal Shah's credit card bill payment platform, CRED, gave users a chance to win free food from Swiggy for one year. Pranav Kaushik, a Delhi techie, bagged this reward after spending 2000 CRED coi...
2	New Zealand end Rohit Sharma-led India's 12-match winning streak	New Zealand defeated India by 8 wickets in the fourth ODI at Hamilton on Thursday to win their first match of the five-match ODI series. India lost an international match under Rohit Sharma's capt...
3	Aegon life iTerm insurance plan helps customers save tax	With Aegon Life iTerm Insurance plan, customers can enjoy tax benefits on your premiums paid and save up to Ã¢Â□Â¹46,800^ on taxes. The plan provides life cover up to the age of 100 years. Also, c...
4	Have known Hirani for yrs, what if MeToo claims are not true: Sonam	Speaking about the sexual harassment allegations against Rajkumar Hirani, Sonam Kapoor said, "I've known Hirani for many years...What if it's not true, the [#MeToo] movement will get derailed." "I...

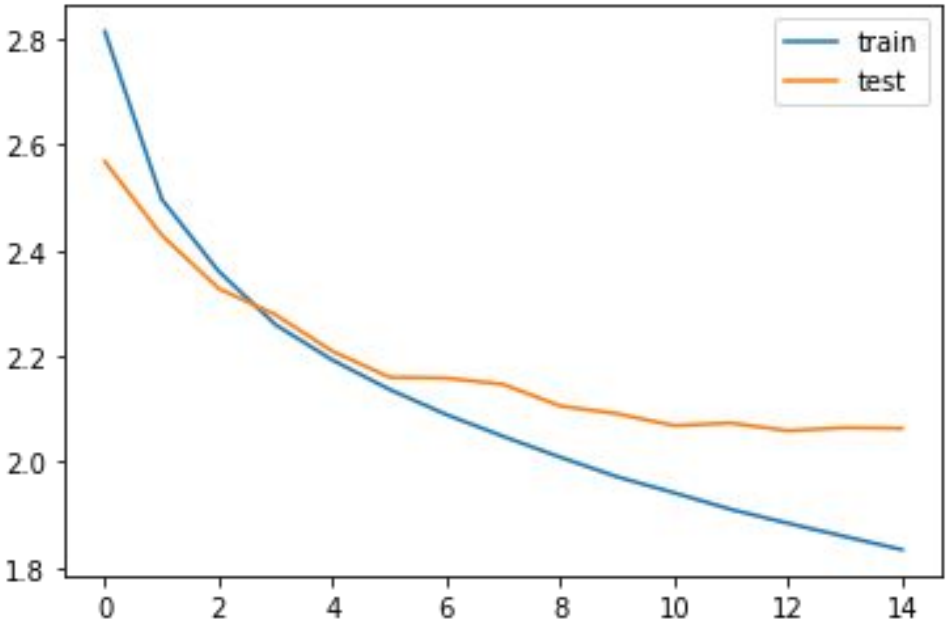
Input to the Decoder for training

```
Review: saurav kant alumnus upgrad iiit program machine learning artificial intelligence
Summary: _START_ upgrad learner switches to career in ml al with salary hike  _END_
```

Word Embeddings extracted from Glove

```
Loaded 400000 word vectors.
all the words in the corpus 519804
the unique words in the corpus 32816
The number of words that are present in both glove vectors and our corpus are 28532 which is nearly 87.0%
word 2 vec length 28532
```

RESULTS



From the plot, we can infer that validation loss has increased after epoch 13 for 2 successive epochs.
-> Training is stopped at epoch 15.