LUND UNIVERSITY

Sequence to Sequence Machine Translation

LU LINFENG

Machine Translation (MT)

- The process of converting a source sentence sequence into the target sentence sequence of same/different length.

Recurrent Neural Networks (RNNs) In MT – Encoder Decoder Approach

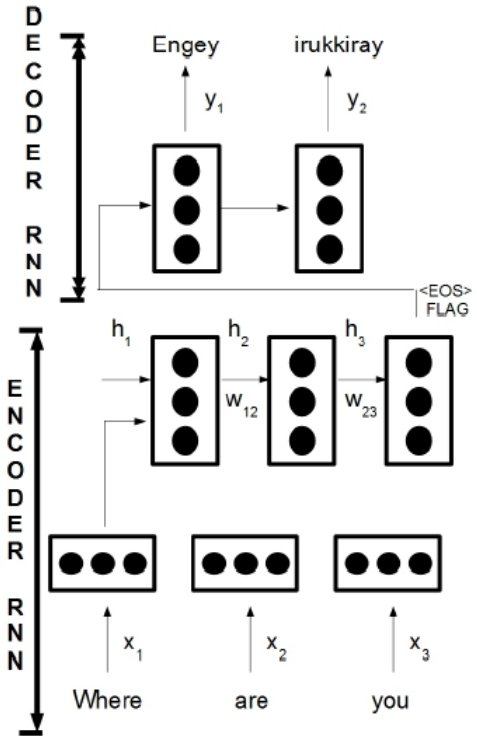


Figure 1 Encoder Decoder Approach

- Machine translation has been worked on for decades now and the recent advancements of using Recurrent Neural Networks (RNNs) have propelled the field to a new height.
- vanishing gradient problem

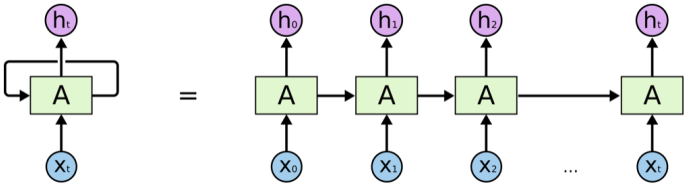


Figure 2 An unrolled recurrent neural network

LUND UNIVERSITY

Long Short Term Memory (LSTM)

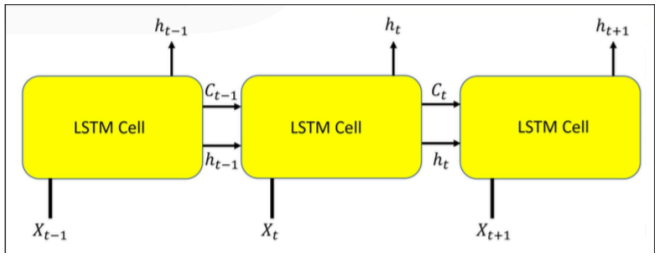


Figure 3 LSTM Layout with cell connections

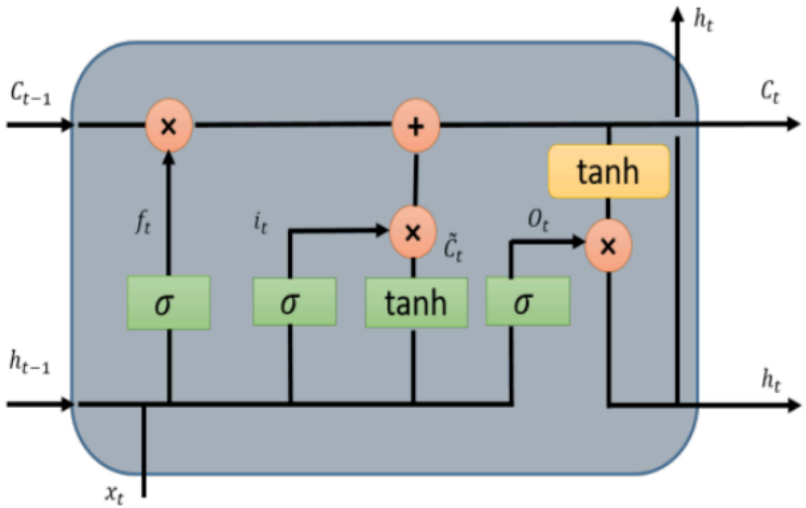


Figure 4 Internal structure of an LSTM cell

LUND UNIVERSITY

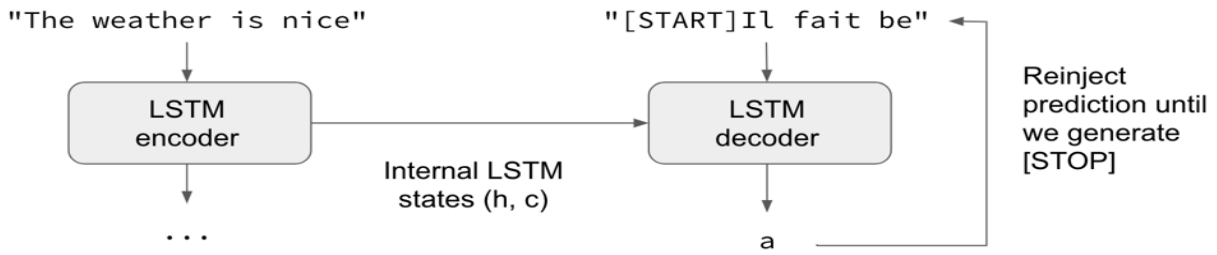
Sequence-to-sequence models in Keras

- Sequence-to-sequence learning is about training models to convert sequences from one domain (e.g. Sentences in Chinese) to sequences in another domain (e.g. The same sentences translated to English).
- Training process:
  - A RNN layer acts as ”encoder”: it processes the input sequence and returns its own internal state.
  - Another RNN layer acts as ”decoder”: It is trained to predict the next characters of the target sequence, given previous characters of the target sequence. Effectively, the decoder learns to generate targets[t+1...] given targets [...t], conditioned on the input sequence.

LUND UNIVERSITY

Sequence-to-sequence models in Keras

- Inference process (decode unknown input sequences)
  - 1) Encode the input sequence into state vectors.
  - 2) Start with a target sequence of size 1 (just the start-of-sequence character).
  - 3) Feed the state vectors and 1-char target sequence to the decoder to produce predictions for the next character.
  - 4) Sample the next character using these predictions (we simply use argmax).
  - 5) Append the sampled character to the target sequence
  - 6) Repeat until we generate the end-of-sequence character or we hit the character limit.



LUND UNIVERSITY

Character level sequence-to-sequence model

- Character level sequence-to-sequence model in Keras
- Collected datasets of pairs of chinese sentences and their english translation
  - manythings.org/anki
  - https://tatoeba.org/eng/downloads
- Parameters
  - Batch_size: Batch size for training, i.e. Total number of training examples present in a single batch.
  - Epochs: Number of epochs to train for, a epoch is defined as a single pass through the entire training set while training a machine learning model.
  - Latent_dim: Latent dimensionality of the encoding space.

The Human Evaluation Results

Test 100 pair of sentences (have not been used for training)
1 (very bad) to 5 (very good)
- Readability,
- Relevance,
- Fluency

Number of Samples =10000

Batch size	64	32	128	64	64	64	64	32	32	128	128	32	32	128	128	64	64	64	64
epochs	100	100	100	100	100	50	200	100	100	100	100	200	50	200	50	50	50	200	200
Latent dim	256	256	256	128	512	256	256	128	512	128	512	256	256	256	256	128	512	128	512
score	241	230	231	238	213	225	228	225	210	227	224	225	220	221	219	218	211	213	212

The Human Evaluation Results

Number of samples	10000	20000	30000
score	241	271	282
我爱我的妻子。 I love my wife.	I like like basies.	I love my life.	I love my life.
他们有12个孩子。 They have twelve children.	They said that.	They have been here engers.	They have the salt or a baby.
请保密。 Please keep this a secret.	Come on, truich.	Please take married.	Please keep to a shory.
會議在五點鐘結束。 The conference ended at five.	The cold is tell hote.	She was a light busy on you.	The meeting ender the man who speaks.
我想用信用卡支付。 I'd like to pay by credit card.	I'll say this begonerald.	I want to be a good singer.	I'd like to know her name.
物價上漲。 Prices went up.	The lause has a rush.	Sill is going out.	Prices went up.
请来吧。 Please come.	Come on, the know.	Please recond.	Please come.
我不知道他什么时候会来。 I don't know when he will come.	I hope poon is weill.	I don't know where Mary was dark.	I don't know when he'll come.
不要擔心這樣的事情。 Don't worry about such a thing.	Don't bey away in the bird.	Don't work in this accident.	Don't worry about your leages.
他們已經結婚十年了。 They have been married for ten years.	They are very big.	They have been called in the world.	They have been married for foot.

Conclusion

- LSTM networks are used for sequence processing problems such as machine translation, since LSTM cells can remember useful information over a long period of time.
- More data produce better translations.

Thank you for your attention!

LUND
UNIVERSITY