Sequence Modeling Using Recurrent Neural Networks

Task I – Recurrent Neural Network Design

RNN Description

The architecture of the Recurrent Neural Network is a LSTM-RNN to learn long-term dependencies. The model has 4 Sequential layers. First layer is an embedding layer that maps the input to a 10-dimensional vector. The following 2 layers consists of 2 LSTM cells, in between is a dropout of 0.2 to prevent overfitting. Lastly is a dense layer that outputs the probability for each character i.e. amino acid using softmax.

Hyper-Parameters

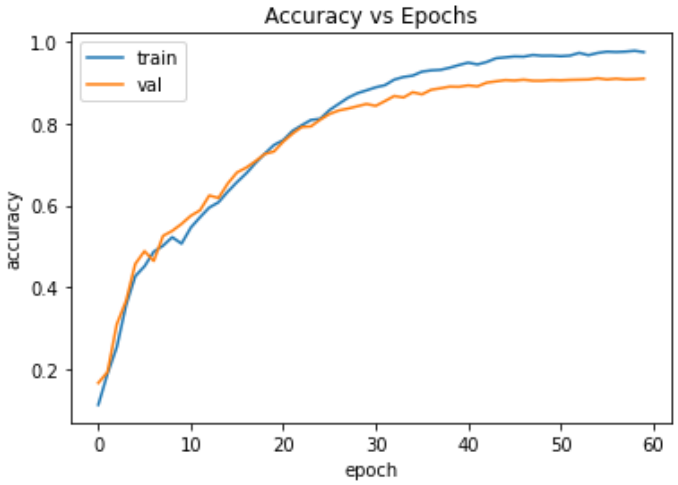
A dropout layer of 0.2 to prevent overfitting is used. The loss used is the *categorical cross-entropy* because the problem is a multi-class problem. Number of epochs used is 60.

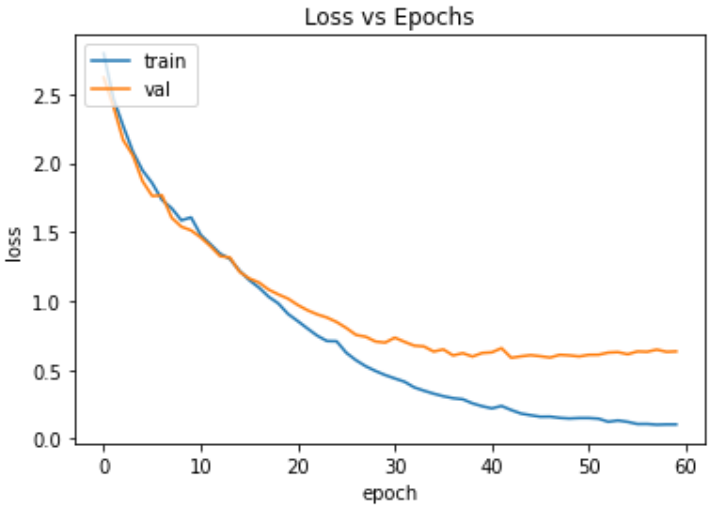
The table below shows the summary of the model.

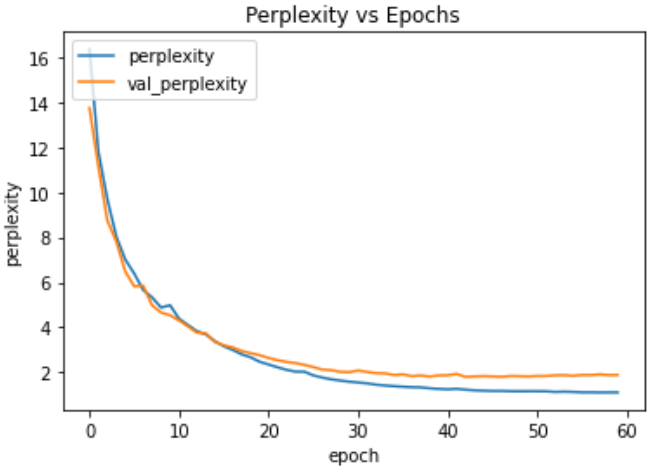
Layer (type)	Output Shape	Param #
embedding_1 (Embedding)	(None, 39, 10)	210
lstm_2 (LSTM)	(None, 39, 150)	96600
dropout_1 (Dropout)	(None, 39, 150)	0
lstm_3 (LSTM)	(None, 100)	100400
dense_1 (Dense)	(None, 21)	2121

Task II– Language Models for Protein Sequences and Evaluation

Default keras weight values were used. The optimizer used is Adam as it makes use of momentum and can handle sparse gradients on noisy problems. The number of epochs used is 60 as anything greater than 60 does not give significant increase in accuracy.


Accuracy vs Epochs
train
val
accuracy
epoch
1.0
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60


Loss vs Epochs
train
val
loss
epoch
2.5
2.0
1.5
1.0
0.5
0.0
0
10
20
30
40
50
60


Perplexity vs Epochs
perplexity
val_perplexity
perplexity
epoch
16
14
12
10
8
6
4
2
0
10
20
30
40
50
60

Measuring Long-Term Dependency

No.	Input Sequence	Predicted Character
1	RVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
2	MRLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
3	MVRSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
4	MVLREGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
5	MVLSRGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
6	MVLSEREWQLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
7	MVLSEGRWQLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
8	MVLSEGERQLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
9	MVLSEGEWRLVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
10	MVLSEGEWQRVLHVWAKVEADVAGHGQDILIRLFKSHPE	T
11	MVLSEGEWQLRLHVWAKVEADVAGHGQDILIRLFKSHPE	T
12	MVLSEGEWQLVRHVWAKVEADVAGHGQDILIRLFKSHPE	T
13	MVLSEGEWQLVLRVWAKVEADVAGHGQDILIRLFKSHPE	T
14	MVLSEGEWQLVLHRWAKVEADVAGHGQDILIRLFKSHPE	T
15	MVLSEGEWQLVLHVRAKVEADVAGHGQDILIRLFKSHPE	Q

The long term dependency of this model is 40 – 15 = 25, because after 15 generations, the prediction is incorrect.

Task III – Sequence Generation Techniques

1. Completing and Generating Sequences

Input	**Result**	**Sequence (input + result)**
“”	“PPYTVVYFPVRGRCAA LRMLLADQGQSWKEEV VTVETWQE”	PPYTVVYFPVRGRCAAL RMLLADQGQSWKEEVV TVETWQE
MVLSEGEWQL	VLSPADKTNVKAAWGK VGAHAGEYGAEALE	MVLSEGEWQL VLSPADKTNVKAAWGK VGAHAGEYGAEALE Y
MNIFEMLRIDEGLRL	KVFGRCELAAAMKRHG LDNYRGYSL	MNIFEMLRIDEGLRL KVFGRCELAAAMKRHG LDNYRGYSL
VLSEGEWQLVLHVWAK VEADVAGHG	QLSALEAKGETPSAV	VLSEGEWQLVLHVWAK VEADVAGHGQLSALEA KGETPSAV
PPYTVVYFPVRGRCAAL RMLLADQGQSWKEEVV TV	EKKSI	PPYTVVYFPVRGRCAAL RMLLADQGQSWKEEVV TVEKKSI

2. K-Table

K																				
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19+**
1	1	1	1	1	1	1	0	0	1	0	1	1	1	0	1	1	0	1	1	0
2	1	1	1	1	1	2	1	1	1	0	1	1	1	1	1	1	1	1	1	2
3	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	0
4	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	0
10	2	2	1	1	2	1	1	1	1	0	0	1	1	1	1	1	1	1	1	0

3. 3-gram language model

3-gram Sequence	Probability
MVL	0.06535947712418301
VLS	0.06535947712418301
LSE	0.06535947712418301
SEG	0.06535947712418301
ESW	0.06535947712418301
EWQ	0.06535947712418301
WQL	0.06535947712418301
QLV	0.06535947712418301
LVL	0.06535947712418301
VLH	0.06535947712418301
LHV	0.06535947712418301
HVW	0.06535947712418301
VWA	0.06535947712418301
WAK	0.06535947712418301
AKV	0.06535947712418301
KVE	0.06535947712418301
EAD	0.06535947712418301
ADV	0.06535947712418301
DVA	0.06535947712418301
VAG	0.06535947712418301
DVA	0.06535947712418301
VAG	0.06535947712418301
HGQ	0.06535947712418301
GQD	0.06535947712418301
HGQ	0.06535947712418301
GQD	0.08714596949891068