Comparison of Neural Network Performance for Predicting Transcription Factor Binding

SS21 Research Internship

Minie Jung

Contents.

01 Introduction

02 Methods

03 Results

04 Conclusion

01. Introduction

01.

Transcription Factor

- Transcription factor binds specific site of DNA and regulates gene expression
  - Transcription factor binding site have a specific motif

01.

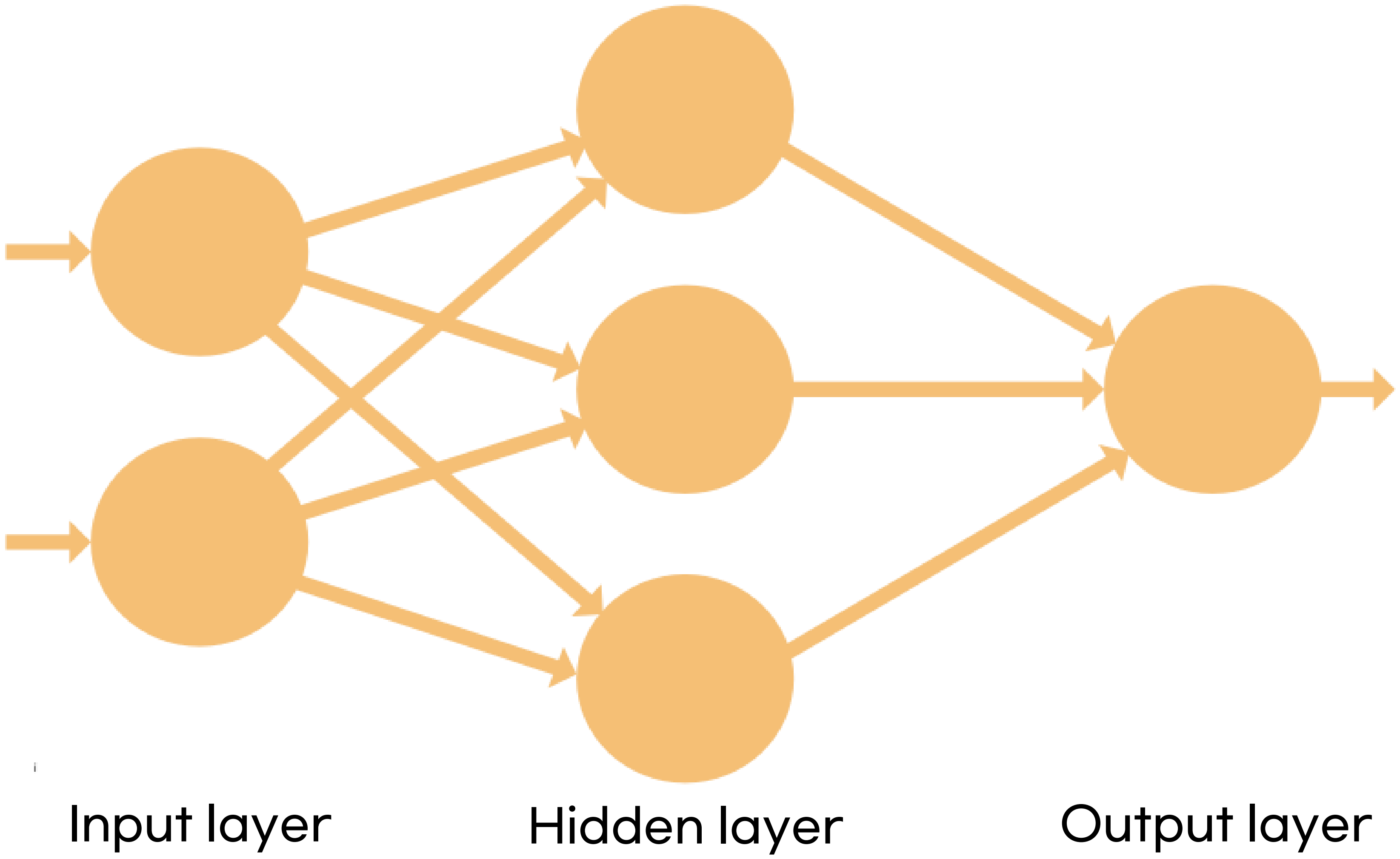
Transcription Factor

- Transcription factor binds specific site of DNA and regulates gene expression
  - Transcription factor binding site have a specific motif

Neural network is one of the powerful tools for predicting transcription factor binding sites

01.

Neural Networks



01.

Convolutional Neural Networks

CNN

- Handle data in the form of multiple arrays
- Image classification
- Convolutional layer, pooling layer, and fully-connected layer

RNN

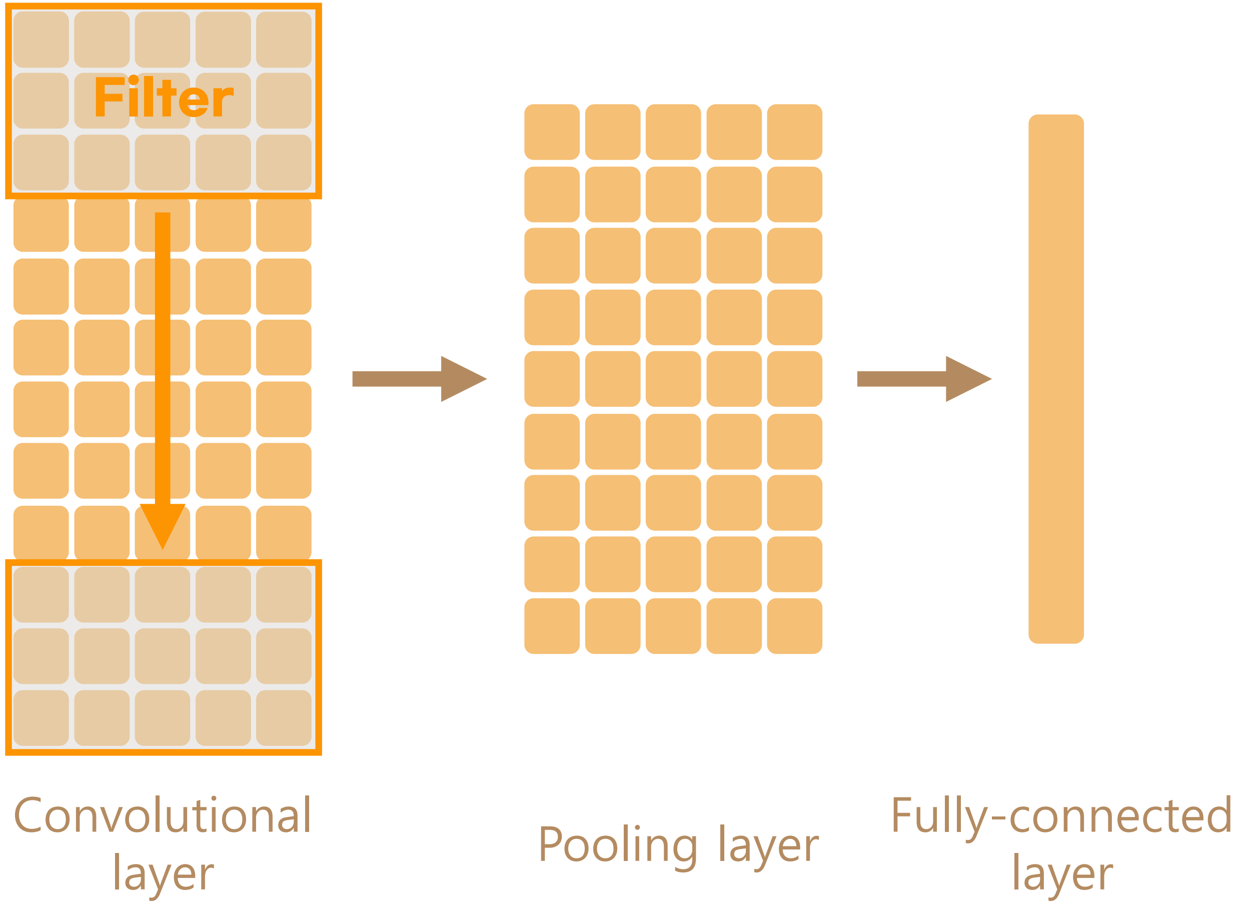
CNN + RNN

Convolutional Neural Networks

CNN

RNN

CNN + RNN



01.

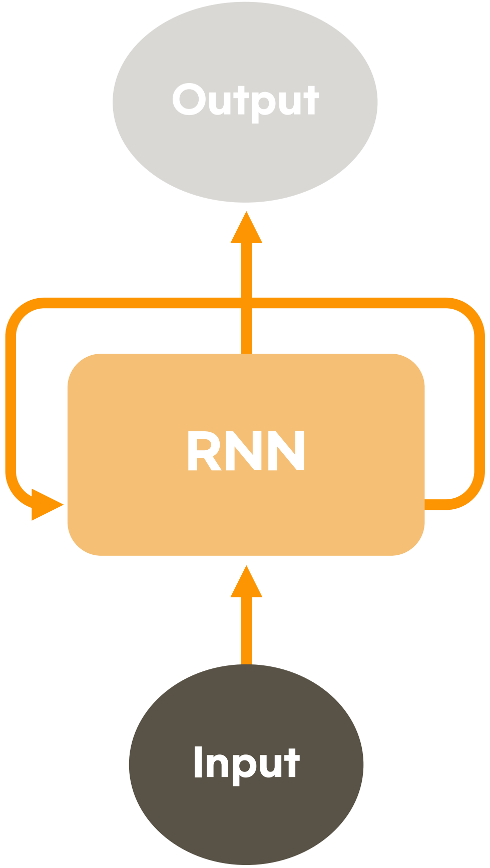
Recurrent Neural Networks

CNN

RNN

CNN + RNN

- Recurrent connection of neuron
- Take its output as its input
- Sequential data such as text, time-series, etc.



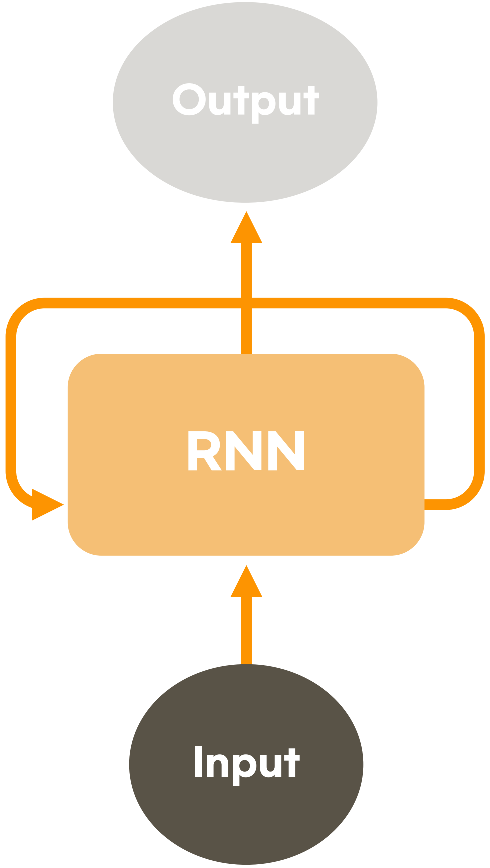
01.

Recurrent Neural Networks

CNN

RNN

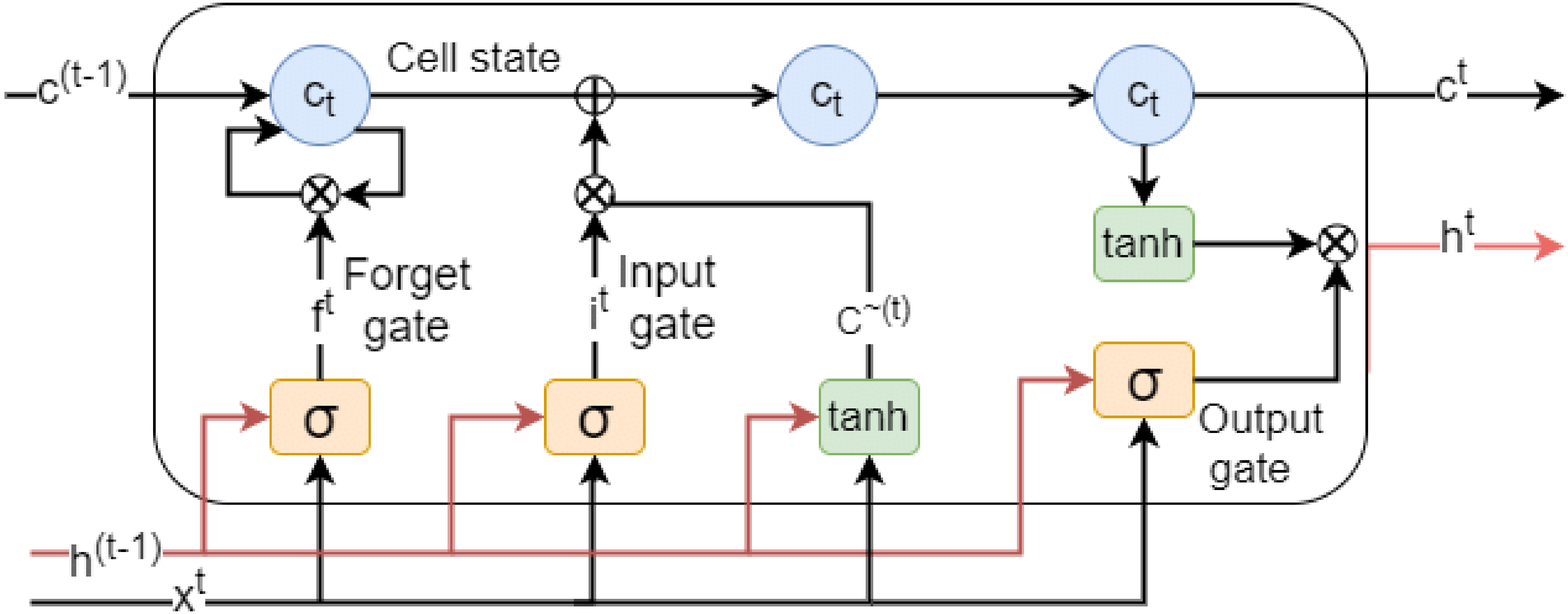
- Recurrent connection of neuron
- Take its output as its input
- Sequential data such as text, time-series, etc.
- **Vanishing gradient problem**



CNN + RNN

01.

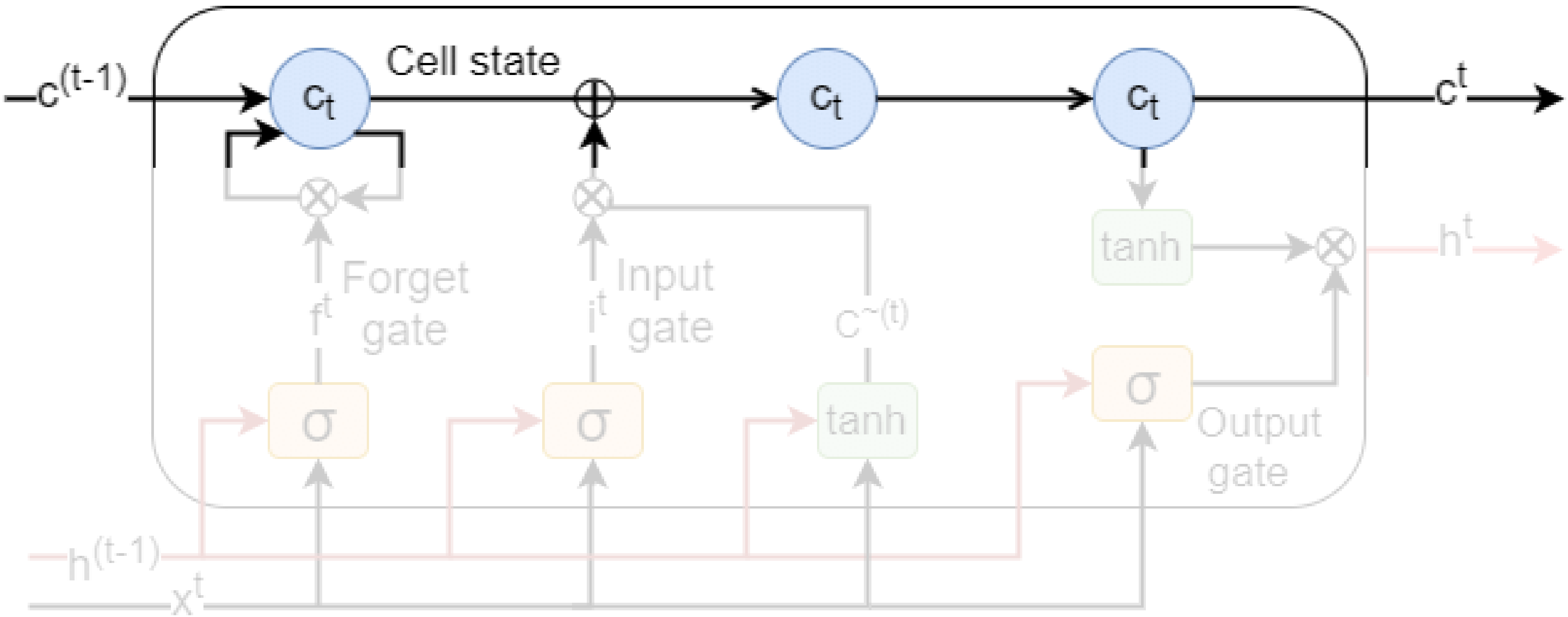
Long Short Term Memory



Jenkins, Ian & Gee, Ludvig & Knauss, Alessia & Yin, Hang & Schroeder, Jan. (2018). Accident Scenario Generation with Recurrent Neural Networks. 3340-3345. 10.1109/ITSC.2018.8569661.

01.

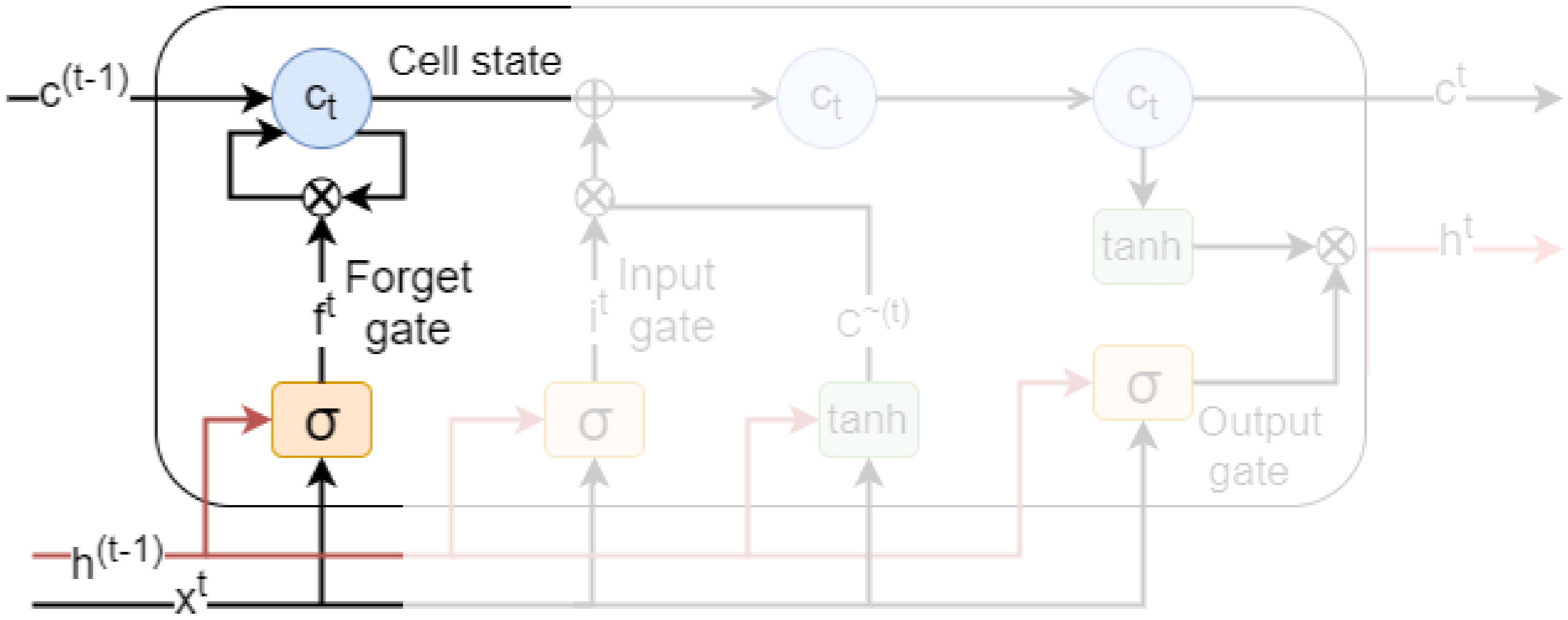
Long Short Term Memory



Jenkins, Ian & Gee, Ludvig & Knauss, Alessia & Yin, Hang & Schroeder, Jan. (2018). Accident Scenario Generation with Recurrent Neural Networks. 3340-3345. 10.1109/ITSC.2018.8569661.

01.

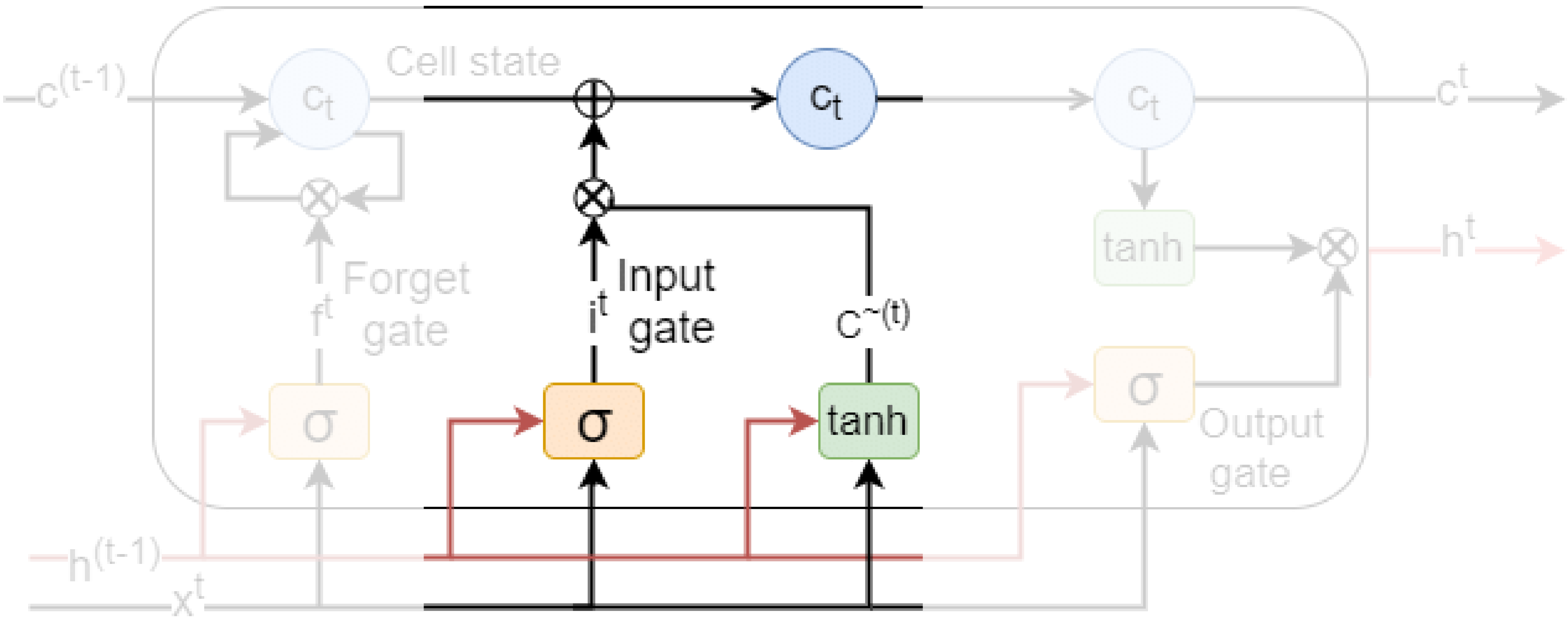
Long Short Term Memory



Jenkins, Ian & Gee, Ludvig & Knauss, Alessia & Yin, Hang & Schroeder, Jan. (2018). Accident Scenario Generation with Recurrent Neural Networks. 3340-3345. 10.1109/ITSC.2018.8569661.

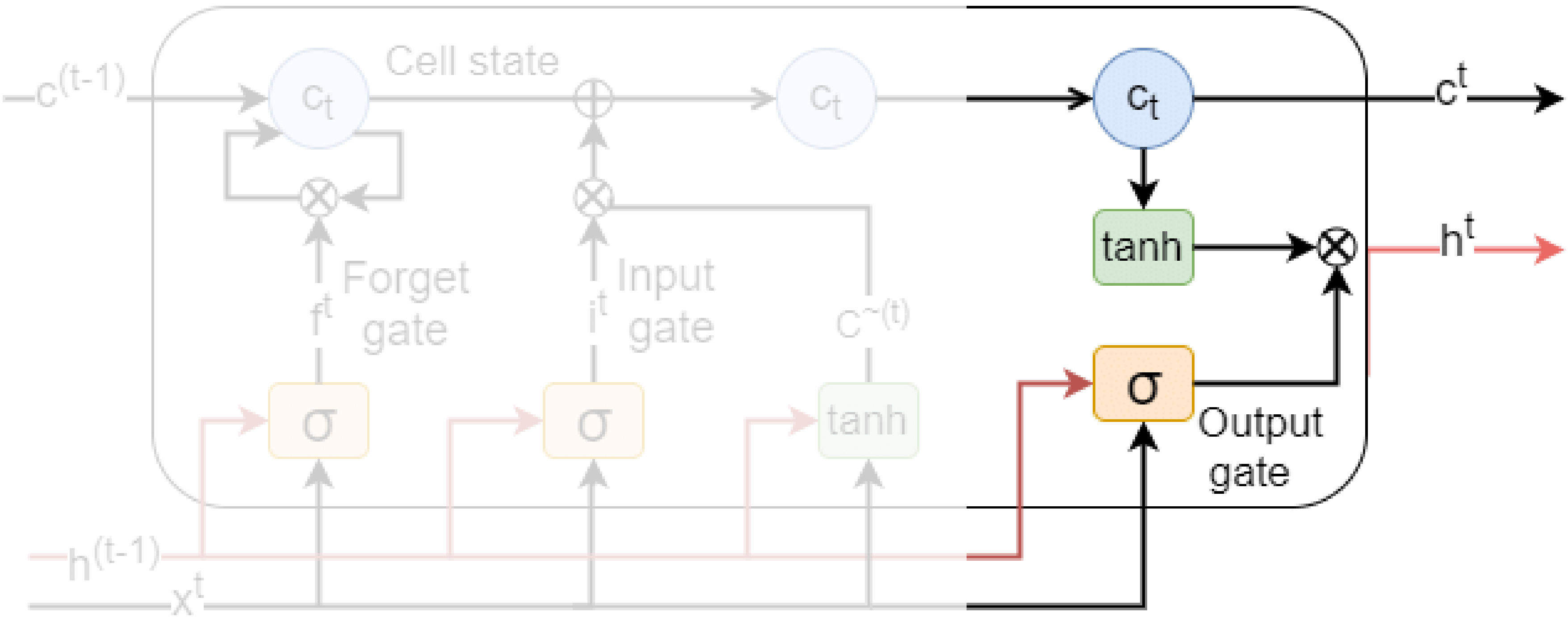
01.

Long Short Term Memory



Jenkins, Ian & Gee, Ludvig & Knauss, Alessia & Yin, Hang & Schroeder, Jan. (2018). Accident Scenario Generation with Recurrent Neural Networks. 3340-3345. 10.1109/ITSC.2018.8569661.

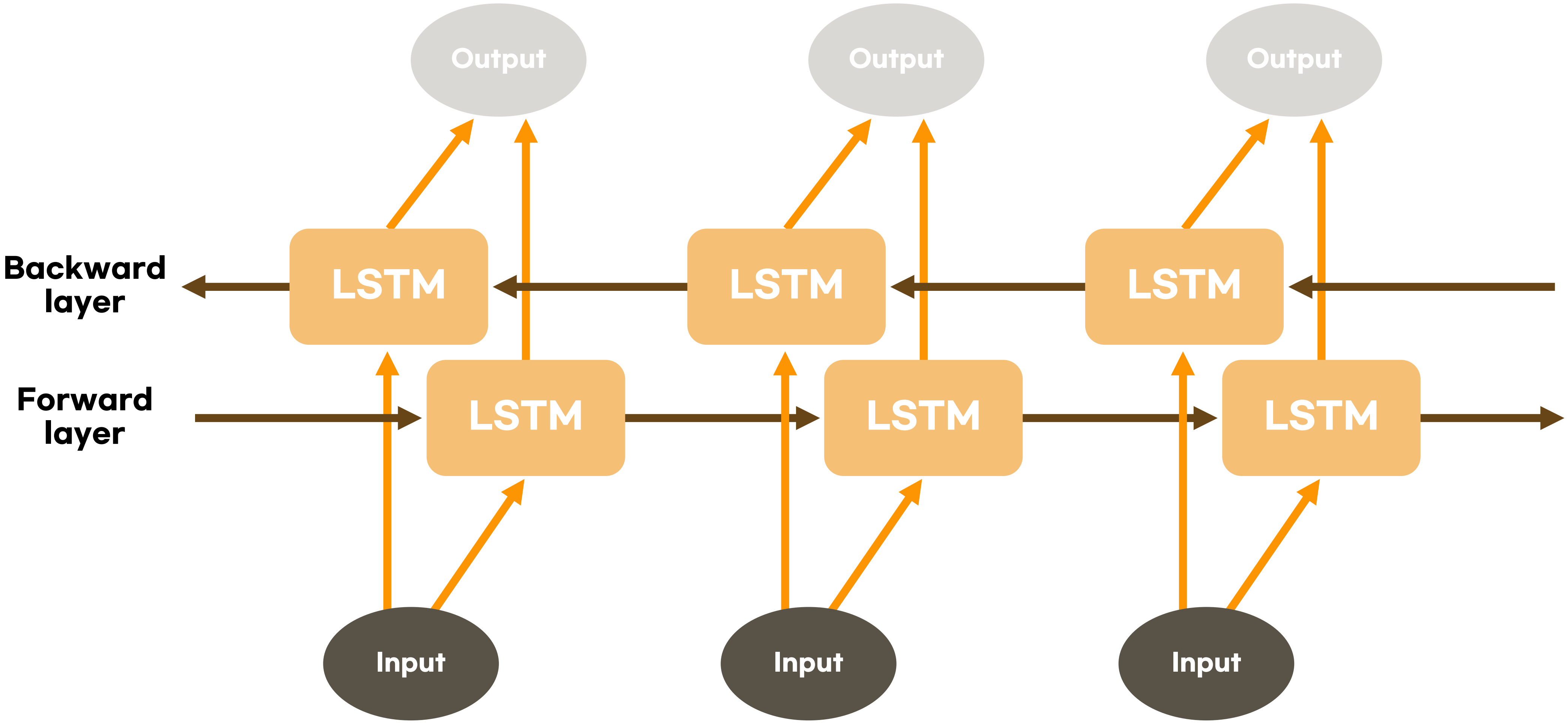
Long Short Term Memory



Jenkins, Ian & Gee, Ludvig & Knauss, Alessia & Yin, Hang & Schroeder, Jan. (2018). Accident Scenario Generation with Recurrent Neural Networks. 3340-3345. 10.1109/ITSC.2018.8569661.

01.

Bidirectional LSTM



01.

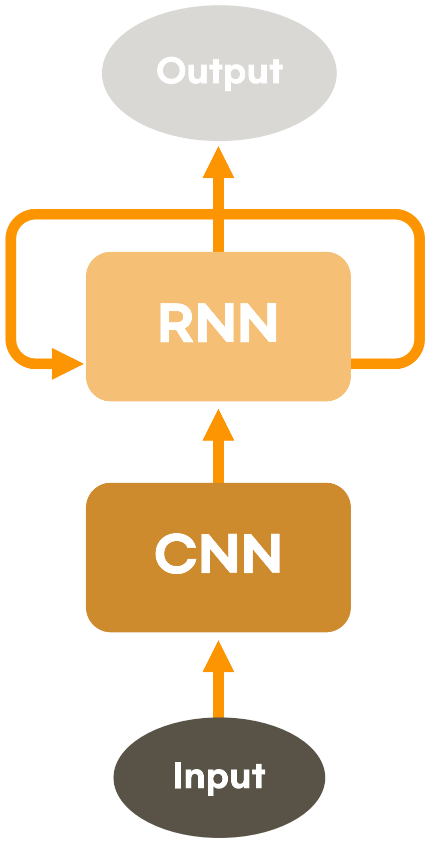
CNN + RNN

CNN

RNN

CNN + RNN

- CNN captures specific pattern of data
- RNN learns feature information and dependencies between data



01.

Introduction

CNN

RNN

CNN
+RNN

The models have different advantages and characteristics.

Introduction

CNN

RNN

CNN
+RNN

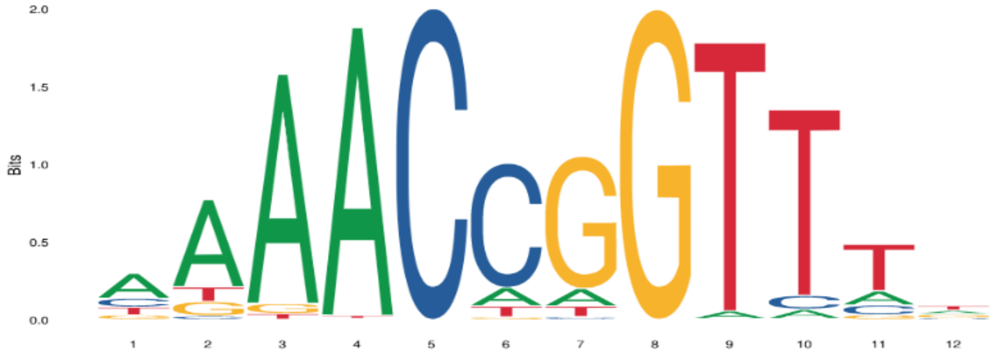
Compare the performance of the model
to see which model handles the task better.

02. Methods

Data

Grainyhead-like 1

- Transcription factor related to wound healing, tubulogenesis, and cancer
- Binds to the consensus DNA sequence 5'-AACCGGTT-3'

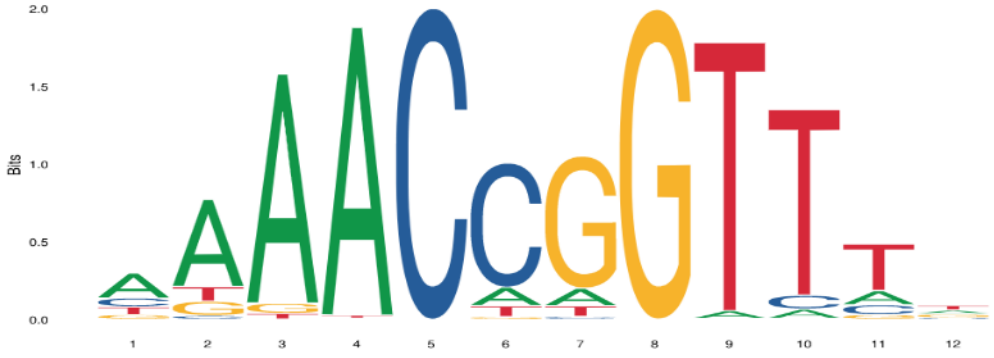


http://jaspar.genereg.net/matrix/MA0647.1/

Data

Grainyhead-like 1

- Transcription factor related to wound healing, tubulogenesis, and cancer
- Binds to the consensus DNA sequence 5'-AACCGGTT-3'



Systematic evolution of ligands by exponential enrichment (SELEX)

- Analyze transcription factors binding specificity
- Provide sequences with high affinity to a specific transcription factor

http://jaspar.genereg.net/matrix/MA0647.1/

Data

Positive set

- Grainyhead-like 1 transcription factor binding site sequences obtained by SELEX experiment

Data

Positive set

- Grainyhead-like 1 transcription factor binding site sequences obtained by SELEX experiment

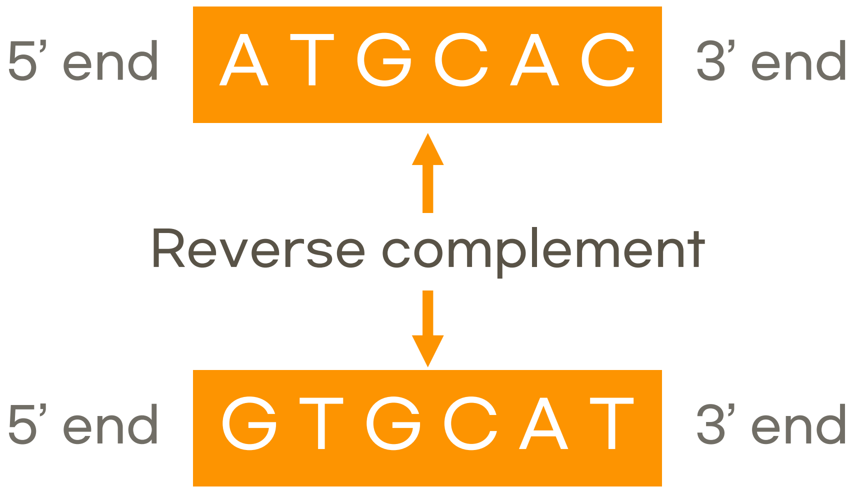
Negative set

- Generated by applying dinucleotide-preserving shuffle to the positive sequences
- Dinucleotide-preserving shuffle shuffles the sequence preserving number of dinucleotides
- Allow the model to learn TF-specific motifs rather than which sequence is not a binding site

Data

Reverse Complement

- Same pattern can appear equally on a forward strand and its reverse
- Add reverse complement of given sequences to improve model performance



02.

Implementation

STEP 1		STEP 2		STEP 3		STEP 4
Data preprocessing	> >	Hyperparameter tuning	> >	Model training	> >	Model evaluation

02.

Implementation

STEP 1

Data preprocessing

One-hot encoding

- Transform categorical data into more appropriate format for machine learning

A T G C →

A	1	0	0	0
T	0	0	0	1
G	0	0	1	0
C	0	1	0	0

Implementation

STEP 2

Hyperparameter tuning

GridSearchCV

- search the best combination of parameters

Implementation

STEP 2

Hyperparameter tuning

GridSearchCV

- search the best combination of parameters

Loss-epoch curves

- detect overfitting

02.

Implementation

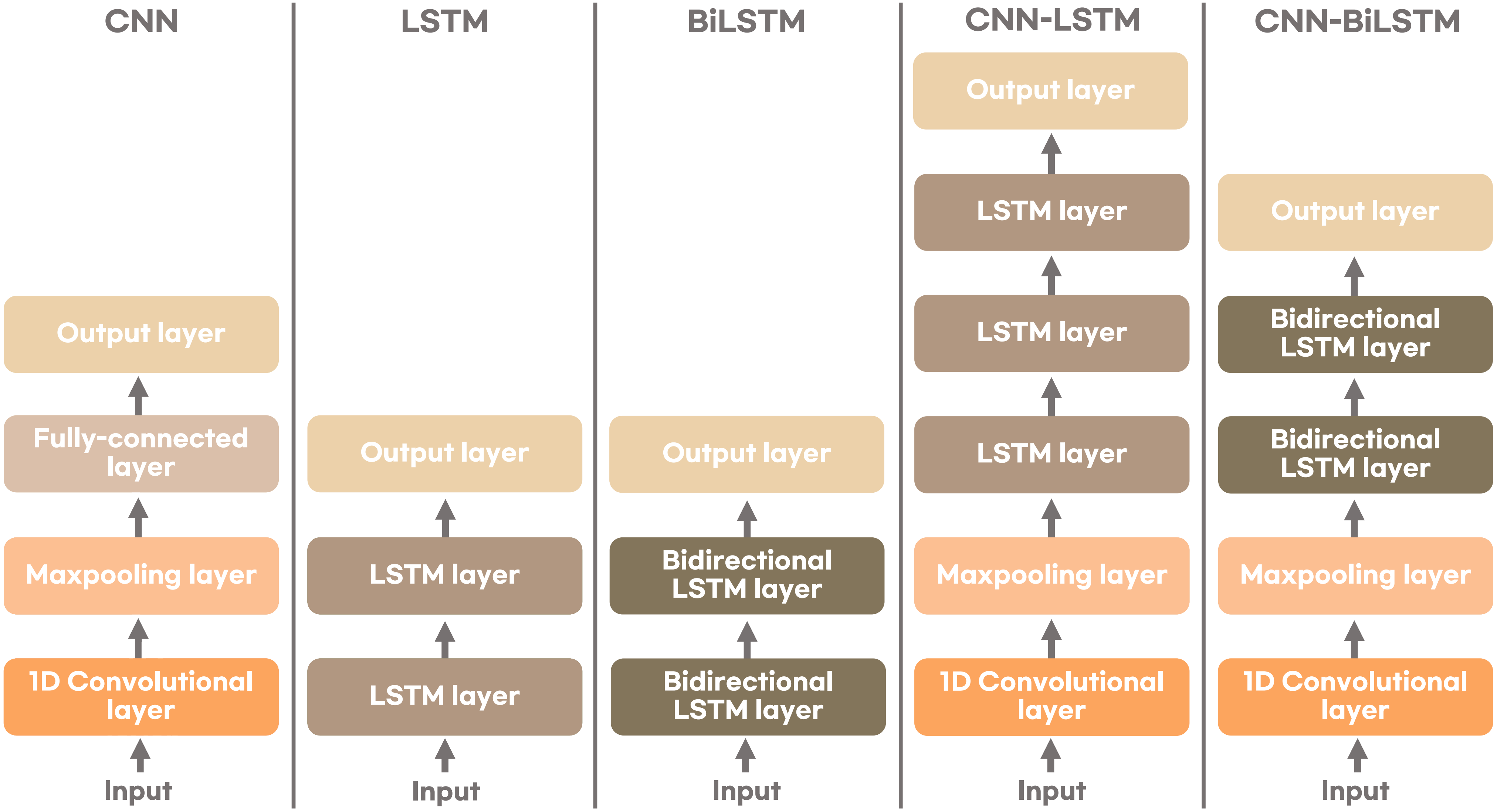
STEP 3

Model training

- Train set pass through the model 50 times
- Applying **earlystopping** to terminate training early if there is no improvement

02.

Implementation



Implementation

STEP 3

Model evaluation

Accuracy represents how the model correctly predict the class

Implementation

STEP 3

Model evaluation

Accuracy represents how the model correctly predict the class

Loss-epoch curve represents how well-trained the model is

02.

Implementation

STEP 3

Model evaluation

Accuracy represents how the model correctly predict the class

Loss-epoch curve represents how well-trained the model is

ROC AUC summarizes the performance of model in general

Implementation

STEP 3

Model evaluation

Accuracy represents how the model correctly predict the class

Loss-epoch curve represents how well-trained the model is

ROC AUC summarizes the performance of model in general

Precision-recall curve AUC summarizes the performance of model for positive data

Implementation

STEP 3

Model evaluation

Accuracy represents how the model correctly predict the class

Loss-epoch curve represents how well-trained the model is

ROC AUC summarizes the performance of model in general

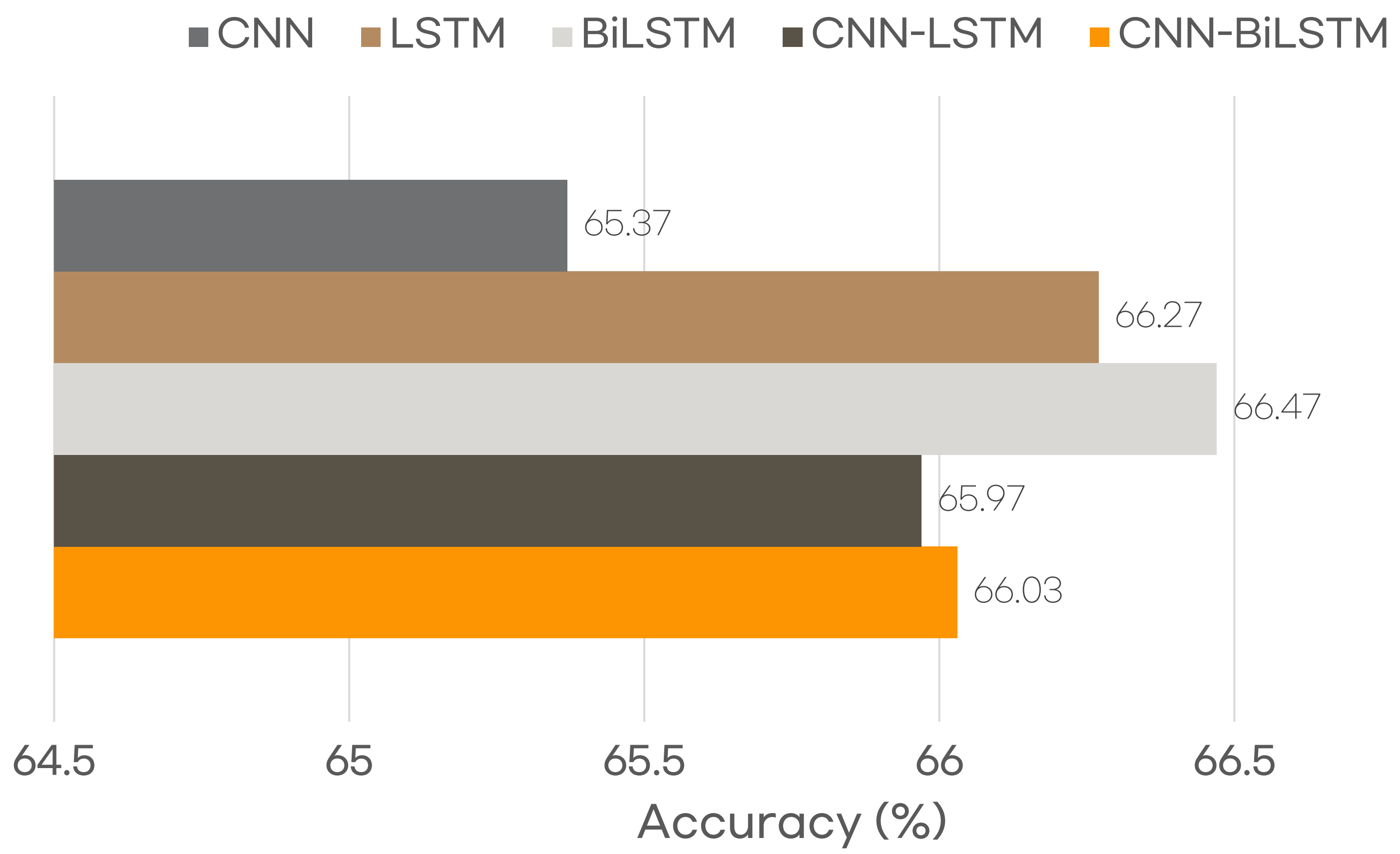
Precision-recall curve AUC summarizes the performance of model for positive data

Visualization shows what the model learns from the data

03. Results

03.

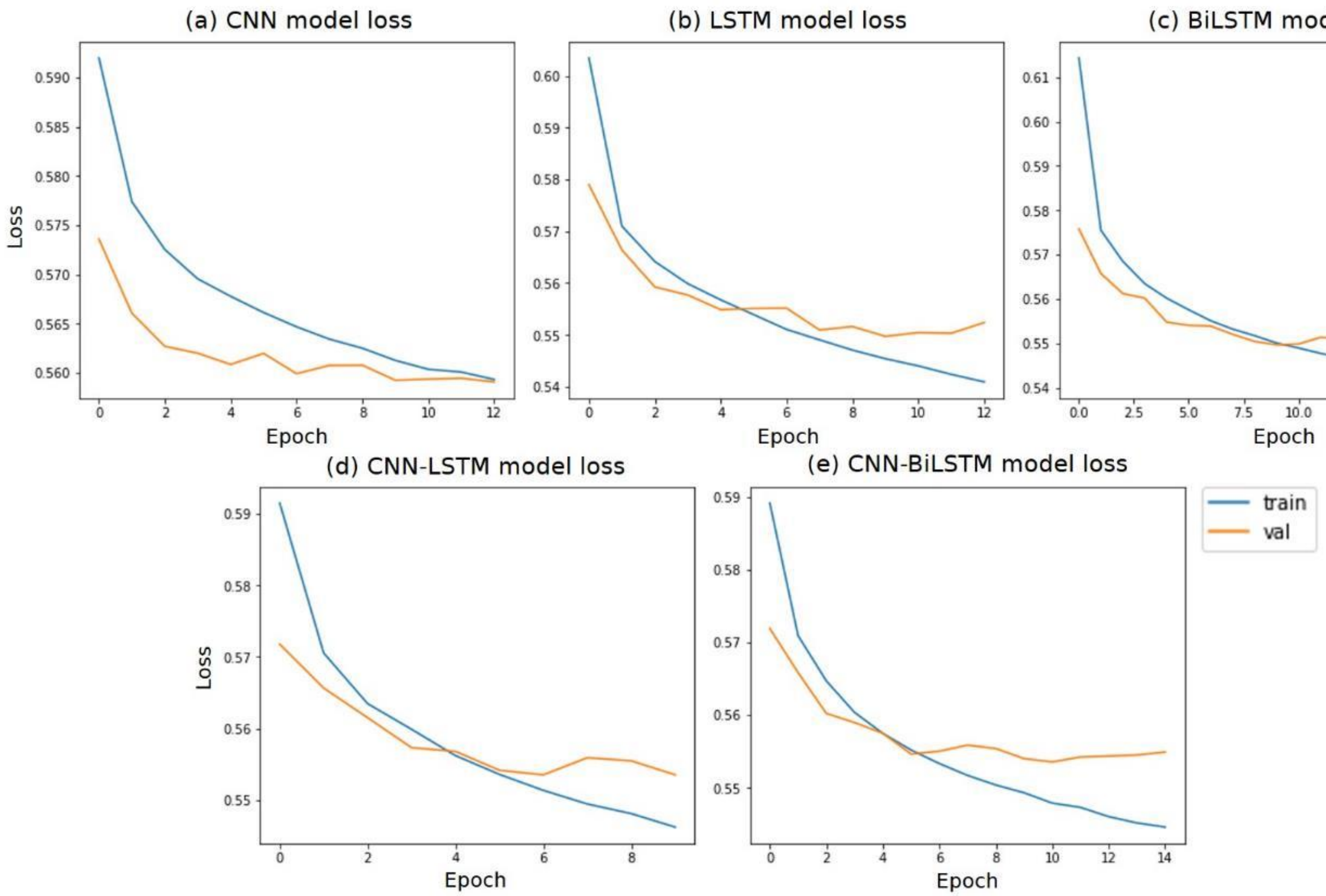
Accuracy



CNN
LSTM
BiLSTM
CNN-LSTM
CNN-BiLSTM
65.37
66.27
66.47
65.97
66.03
64.5
65
65.5
66
66.5
67
Accuracy (%)

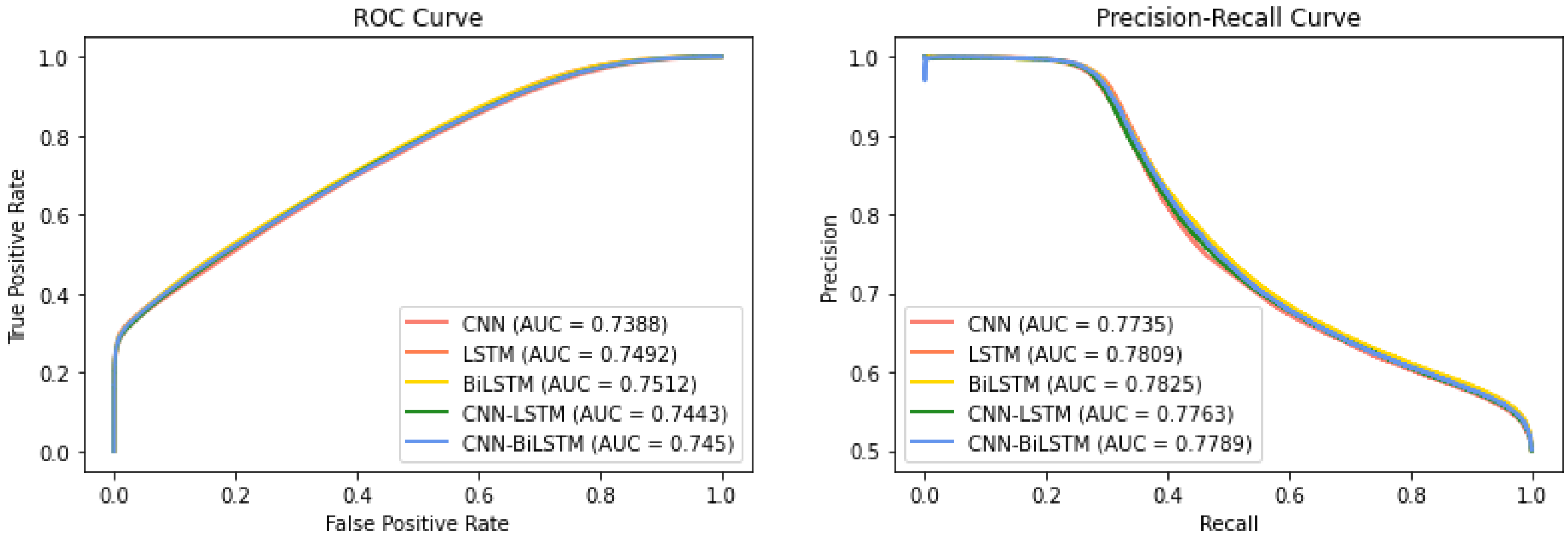
03.

Loss-Epoch Plots



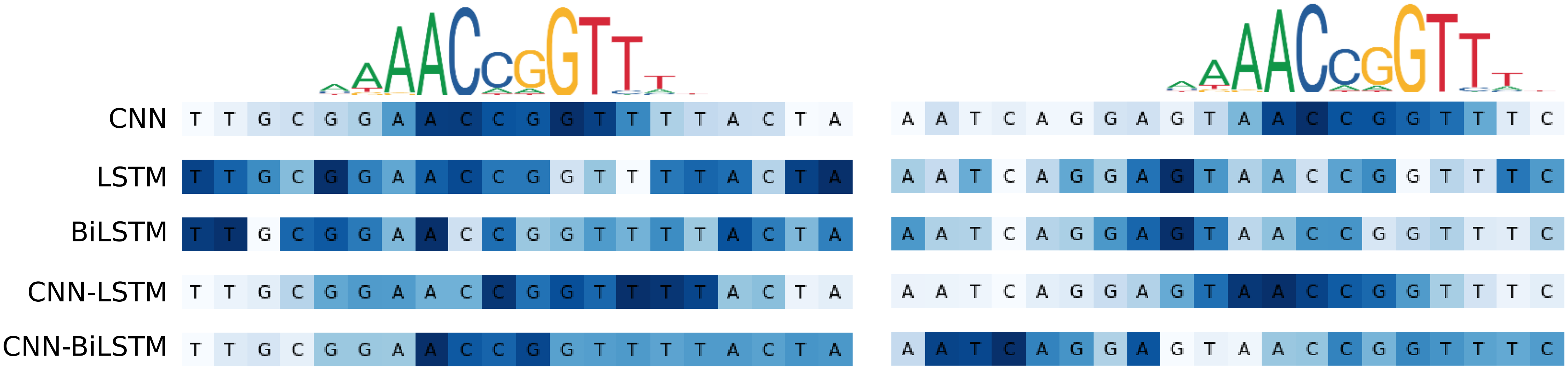
03.

Area Under the Curves



03.

Visualization



04. Conclusion

Conclusion

CNN

- Be able to capture consensus motif
- Lowest accuracy and AUCs

RNN

CNN+ RNN

Conclusion

CNN

- Be able to capture consensus motif
- Lowest accuracy and AUCs

RNN

- Could not capture consensus motif
- Highest accuracy and AUCs

CNN+ RNN

Conclusion

CNN

- Be able to capture consensus motif
- Lowest accuracy and AUCs

RNN

- Could not capture consensus motif
- Highest accuracy and AUCs

CNN+ RNN

- was expected to show the best performance but wasn't
- Evaluation results are similar but worse than RNN models

04.

Conclusion

The performances of the models are similar.

04.

Conclusion

The performances of the models are similar.

Why?

Conclusion

The performances of the models are similar.

Why? 1. The data might be not complex enough to observe the difference of models.

Conclusion

The performances of the models are similar.

Why?
1. The data might be not complex enough to observe the difference of models.
2. There is a potential to improve the performance of the model.

Conclusion

The performances of the models are similar.

Why?

1. The data might be not complex enough to observe the difference of models.
2. There is a potential to improve the performance of the model.

How to improve the performances of models?

Conclusion

1. Improvement of data

- The longer or more complex sequence data
- Better negative data

Conclusion

1. Improvement of data

- The longer or more complex sequence data
- Better negative data

2. Word embedding

- k-mer as a word
- Map k-mer vectors by co-occurance
- Might be able to extract more information (position of k-mer, motif detection, etc.)

Thank you