Promoter Sequence Classification

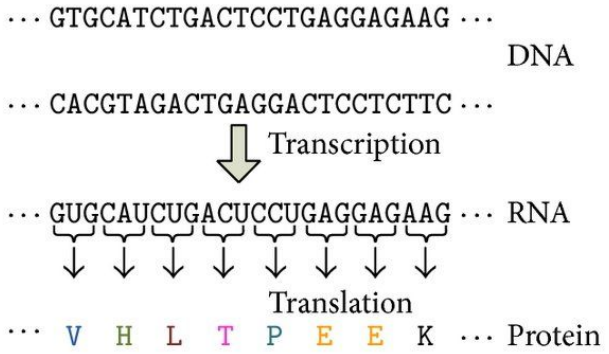
using CNN, GRU, BERT based models

2019-13773 Kyungjin Kim
2018-18574 Junyoung Park
2018-12018 Sungmin Song

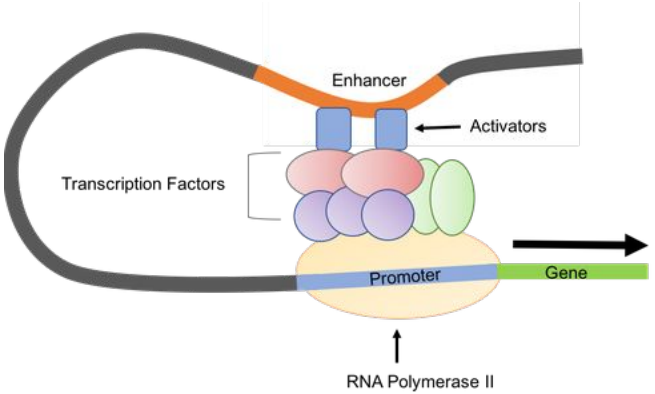
Contents

Contents

Motivation



Central Dogma



Promoter Gene

- DNA, especially the non-coding region, indeed exhibits great similarity to human language, ranging from alphabets and lexicons to grammar and phonetics
- We chose promoter gene classification as a sample task

https://www.researchgate.net/publication/262055850
https://www.addgene.org/mol-bio-reference/promoters

Contents

Idea

Task: Classifying the sequence whether it is a promoter gene

Models: CNN+GRU(baseline), CNN, GRU and BERT … (+some variations)

Goal 1: To improve the accuracy by implementing other appropriate model

Goal 2: To practice applying various machine learning models into bioinformatic task

Dataset

	sequence	label
0	TTAATTTGTCCTTATTTGATTAAGAAGAATAAATCTTATATATAGA...	1
1	ATAGCTCAAATTGCTTTATTAGTATTAGAATCAGCTGTAGCTATAA...	1
2	AAGCTTCCCTTTAATGTGCTCCTTGTGAATACAGCATTACAATGCC...	1
3	TATGTAGAATCTGTACAAGTATCTGTGTTTGGACAATGGCATGTGT...	1
4	ACATATTACTGCATACAGGTCTCAAATTATAAAATGACACTCGTGG...	1
...	...	...
11295	CGACAAAGTTTGATCCATGTGCATTCTTGGCGCCTTATCGATAGCT...	1
11296	CATATCTACATCTCGCTTGCTCCTTCCCTTTCGCTGCGTGTGTGTG...	1
11297	ATACCGCGGAAGCGCAAAAGTACCAGAATTTCCCTGGTATCGCGCT...	1
11298	ATTATTCCGAATTCTTTTATCAGATTTAAATATGGGAAACACTTTA...	1
11299	AATTCATTTATACCTGCATTTGTAACTGTACTAAATCTTCAACCAA...	1

[11300 rows x 2 columns]
(11300, 2)

	sequence	label
0	TAATTACATTATTTTTTTATTTACGAATTTGTTATTCCGCTTTTAT...	0
1	ATTTTTACAAGAACAAGACATTTAACTTTAACTTTATCTTTAGCTT...	0
2	AGAGATAGGTGGGTCTGTAACACTCGAATCAAAAACAATATTAAGA...	0
3	TATGTATATAGAGATAGGCGTTGCCAATAACTTTTGCGTTTTTTGC...	0
4	AGAAATAATAGCTAGAGCAAAAAACAGCTTAGAACGGCTGATGCTC...	0
...	...	...
11295	TGGTAAAAAATTGTACACCTAACTAGTGCCTTCATGTATACCACCA...	0
11296	AGTGCAACTGGAGCCGTGCCGTGACCCACAGAGATCGCCCACTCGA...	0
11297	GCATGGATTTCATATTATCTTAATCGACTTGCTTTTATAAAATAGG...	0
11298	GTGACCAGGTTTTGCTCTAATGCGAAGTACGGATTGGGTAGAGATA...	0
11299	TCATATTGAAAATTGATAAGATTGATATTAACGTAGCAAAGAAAGC...	0

[11300 rows x 2 columns]
(11300, 2)

- Sequences: same length (about 300)
- Label: 1: promotor gene, 0: non-promotor gene

Contents

Baseline: CNN+GRU

- 2 CNN + 1 bidirectional GRU + 5 Linear
- Train 115 epoch with 100 minutes
  - Accuracy : 87.6%
  - Precision : 86%
  - Recall : 89%

Contents

CNN

- Check the contribution of CNN part in Baseline
- Use only CNN layers and one more dense layer

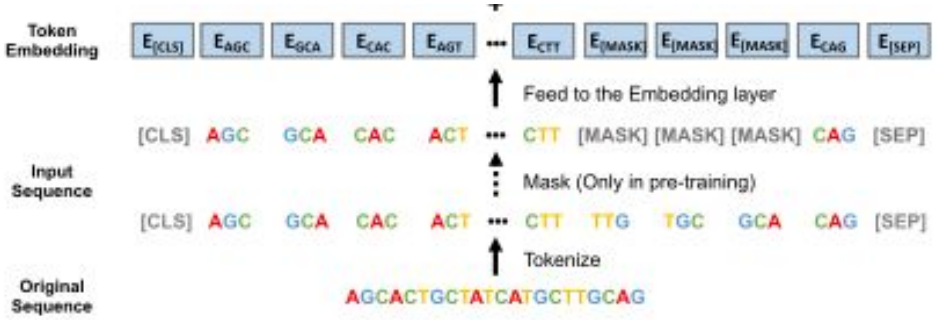
Contents

GRU: only GRU

- Check the contribution of GRU part in Baseline
- Use only Bidirectional GRU and Linear Layers

GRU: Window + GRU

- Idea of DNABERT
  - Window-based tokenization
- Wanted to replace CNN by Window-based tokenization



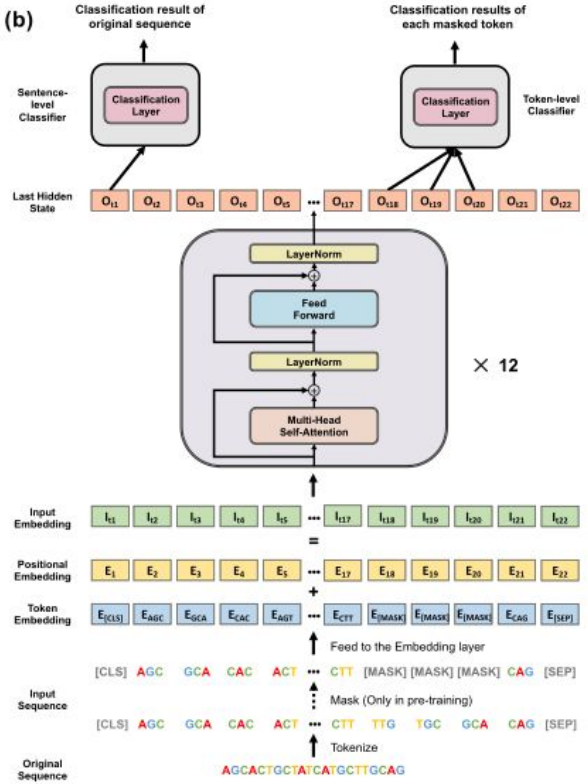
GRU: Result

- Slow Train Speed
  - Limitation of RNN architecture
  - Long Input Sequence
  - only 40 epochs with 120 minutes
- Low Performance
  - only GRU accuracy : 82%
  - Window + GRU accuracy : 73%

Contents

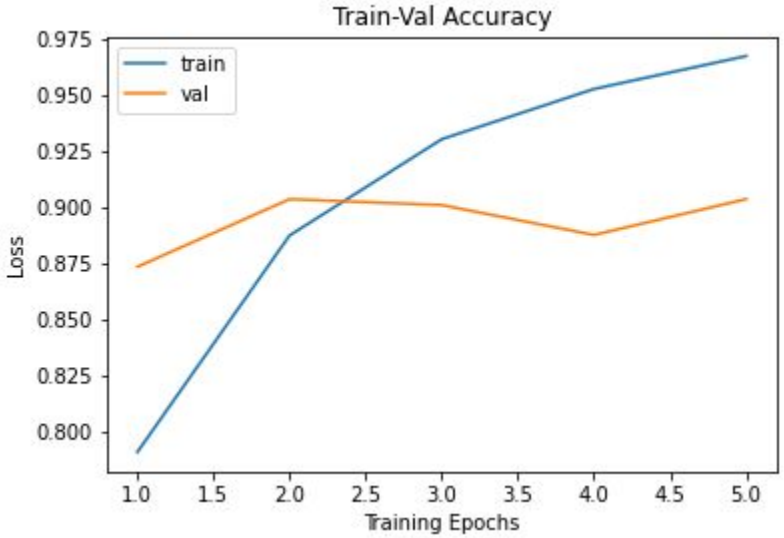
DNABERT finetuning

- BERT pretrained with DNA sequences
- SOTA in DNA classification Area
- MLM with window-based tokenization



DNABERT + 1 Classification Layer

- BERT + 1 Linear layer
- Finetuning with 5 epoch
  - take about 180 minutes
- Good Performance, but Unstable

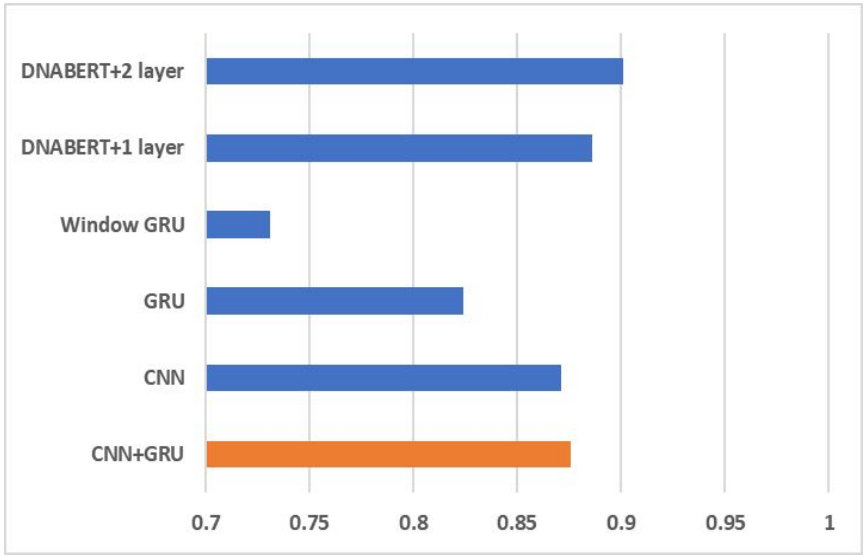


DNABERT + 2 Classification Layer

- BERT + 2 Linear layer
- Better Performance
- Best Result
    - Accuracy : 90%

Contents

Accuracy



- DNABERT + 2 layers had the best result
- DNABERT + 1 layer also outperformed CNN + GRU
- Only CNN had slightly low performance than CNN+GRU

Model Size

Model	Parameters
CNN+GRU	158,207
CNN	377,784
GRU	149,298
Window GRU	151,602
DNABERT + 1 layer	89,192,450
DNABERT + 2 layer	89,388,290

- DNABERT >> Others
- CNN > CNN+GRU > GRU

Training time

- Training time
  - DNABERTs: 30min/epoch
  - CNN+GRU: 40s/epoch
  - GRUs: 3min/epoch
  - CNN: 20s/epoch
- DNABERT models can be improved with more epochs

Contents

Conclusion

- Pretrained Language Model can be applied on Promoter classification
  - Outperformed the baseline, and possibility for better results
- CNN itself has a nice performance
  - Good choice for limited computation source
- RNN structure is inappropriate
  - Slow training + Low performance
  - Low merit for combining with CNN

Thank you!