An RNN based pipeline for the classification of HIV-1 genomes

Rayhan Rashed (1505006) Nishat Anjum Bristy (1505007)

Bangladesh University of Engineering and Technology

October 5, 2021

Problem description

The classification of infections among the subtypes of human immunodeficiency virus type 1 (HIV-1) is a routine component of clinical management, and there are now many classification algorithms available for this purpose [1].

Our Goal

We propose an alignment-free subtyping method using Recurrent Neural Networks that operates on both time series analysis of the individual character and k-mer frequencies in HIV-1 sequences.

Problem description

The classification of infections among the subtypes of human immunodeficiency virus type 1 (HIV-1) is a routine component of clinical management, and there are now many classification algorithms available for this purpose [1].

Our Goal

We propose an alignment-free subtyping method using Recurrent Neural Networks that operates on both time series analysis of the individual character and k-mer frequencies in HIV-1 sequences.

Dataset description

- The dataset is taken from https://www.hiv.lanl.gov/components/sequence/HIV/search/185search.html
- The dataset consists of 243 subtypes of HIV-1 genomes. We had a total of 12,938 sample genomes, with these 243 subtypes.
- We took 25 subtypes among them, which consisted of the heighest samples.
- The subtype with heighest samples was 'B', with 5727 samples. And the subtype '35_AD', with 22 samples.

Dataset description

- The dataset is taken from https://www.hiv.lanl.gov/components/sequence/HIV/search/185search.html
- The dataset consists of 243 subtypes of HIV-1 genomes. We had a total of 12,938 sample genomes, with these 243 subtypes.
- We took 25 subtypes among them, which consisted of the heighest samples.
- The subtype with heighest samples was 'B', with 5727 samples. And the subtype '35_AD', with 22 samples.

Dataset description

- The dataset is taken from https://www.hiv.lanl.gov/components/sequence/HIV/search/185search.html
- The dataset consists of 243 subtypes of HIV-1 genomes. We had a total of 12,938 sample genomes, with these 243 subtypes.
- We took 25 subtypes among them, which consisted of the heighest samples.
- The subtype with heighest samples was 'B', with 5727 samples. And the subtype '35_AD', with 22 samples.

Dataset description

- The dataset is taken from https://www.hiv.lanl.gov/components/sequence/HIV/search/185search.html
- The dataset consists of 243 subtypes of HIV-1 genomes. We had a total of 12,938 sample genomes, with these 243 subtypes.
- We took 25 subtypes among them, which consisted of the heighest samples.
- The subtype with heighest samples was 'B', with 5727 samples. And the subtype '35_AD', with 22 samples.

Subtype information

```
subtypes = {'B': 5727, 'C': 2077, '01_AE': 1426, 'A1': 498, '01B
    ': 210, '02_AG': 168, 'BF1': 143, 'A6': 117, 'A1C': 111, 'G
    ': 96, 'BC': 95, 'A1D': 94, 'AD': 94, 'D': 87, 'F1': 82, '
    A1CD': 62, 'CD': 61, 'O': 57, '0107': 57, '01BC': 50, '07
    _BC': 41, '08_BC': 35, '02A1': 29, '11_cpx': 25, '35_AD':
    22}
```

Listing 1: Code snippet for subtype information

Model Implementation and Comparison

- According to the k-mer value, we first split the sequences and tokenize them. This way, each k-mer is represented with an integer.
- These tokens are passed to an embedding layer, to reduce the dimension of One-Hot encoded vector space to a smaller 24 dimension.
- The next layers are consecutively a bidirectional LSTM layer, a dropout layer, a dense layer with 'relu' activation and a final dense layer with softmax.

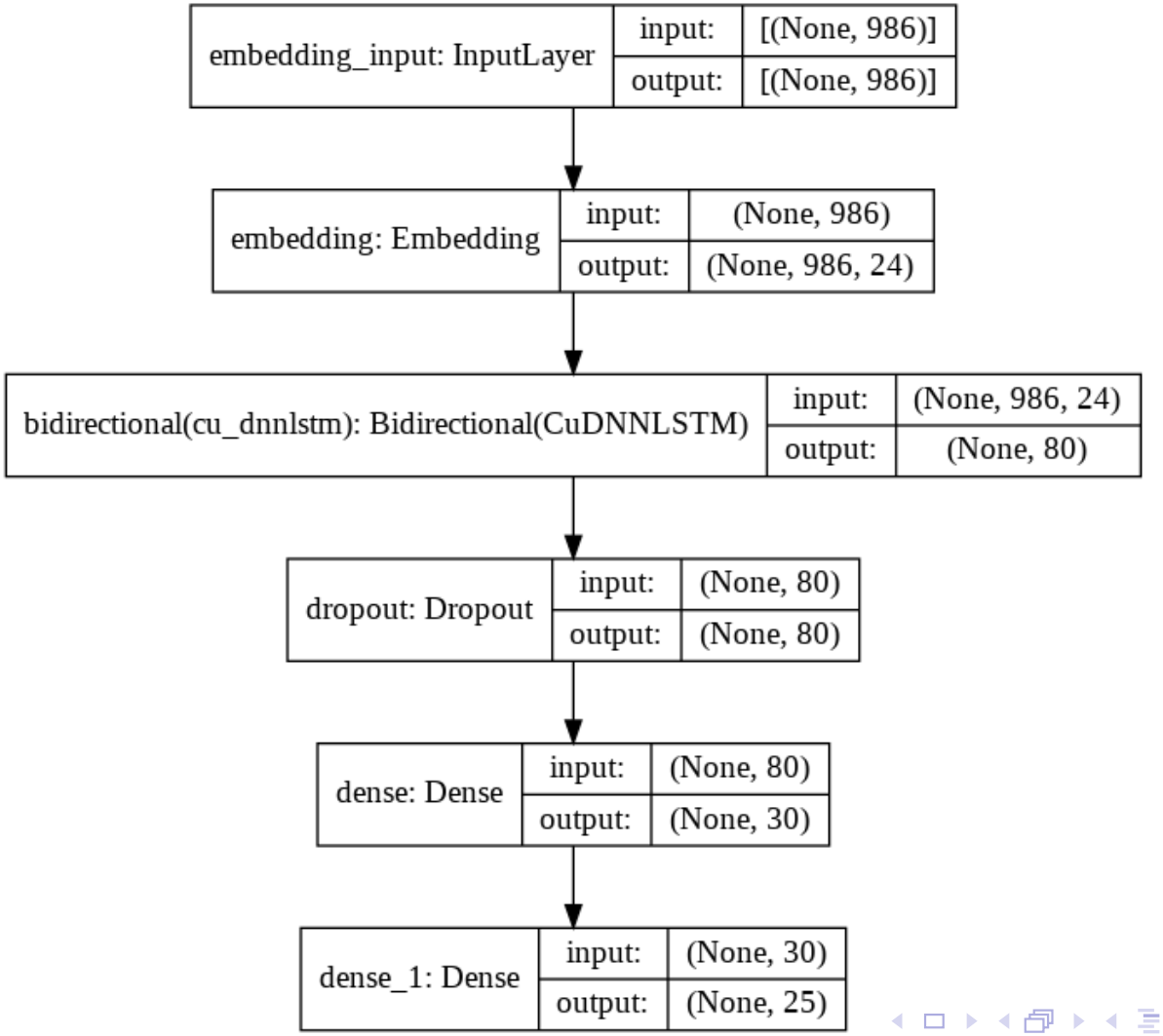
Model Implementation and Comparison

- According to the k-mer value, we first split the sequences and tokenize them. This way, each k-mer is represented with an integer.
- These tokens are passed to an embedding layer, to reduce the dimension of One-Hot encoded vector space to a smaller 24 dimension.
- The next layers are consecutively a bidirectional LSTM layer, a dropout layer, a dense layer with 'relu' activation and a final dense layer with softmax.

Model Implementation and Comparison

- According to the k-mer value, we first split the sequences and tokenize them. This way, each k-mer is represented with an integer.
- These tokens are passed to an embedding layer, to reduce the dimension of One-Hot encoded vector space to a smaller 24 dimension.
- The next layers are consecutively a bidirectional LSTM layer, a dropout layer, a dense layer with 'relu' activation and a final dense layer with softmax.

Model Summary



Comparison of different model conditions

The following table lists the accuracy of training and validation dataset accuracy of all our experiments.

k	Read Length	Training Accuracy	Validation accuracy
1(One pass)	7500	72.69	70.26
1(Sliding Window)	7500	95.54	93.03
1	Whole sequence	82.64	82.42
15	200	96.35	91.76
15	1000	99.78	92.23
15	Whole sequence	92.07	85.48
21	200	99.65	93.32
21	1000	94.04	88.02
21	Whole sequence	96.97	83.04

Summary of different model conditions

- With character level time series analysis (k = 1), a sliding window approach seems to be the most accurate. This is because sliding window gives the opportunity of increasing samples in the training dataset, as well as ensures that every portion of the genome is being considered.
- Both k = 15 and k = 21 performs very well. This is something we can further investigate.

Summary of different model conditions

- With character level time series analysis (k = 1), a sliding window approach seems to be the most accurate. This is because sliding window gives the opportunity of increasing samples in the training dataset, as well as ensures that every portion of the genome is being considered.
- Both k = 15 and k = 21 performs very well. This is something we can further investigate.

References

Stephen Solis-Reyes, Mariano Avino, Art Poon, and Lila Kari.
An open-source k-mer based machine learning tool for fast and accurate subtyping of hiv-1 genomes.
PLoS One, 13(11):e0206409, 2018.