Generating amino acid sequences

Jitian Zhao

University of Wisconsin Madison

September 9, 2019

Outline

- Goal
- Char-RNN
- Transformer

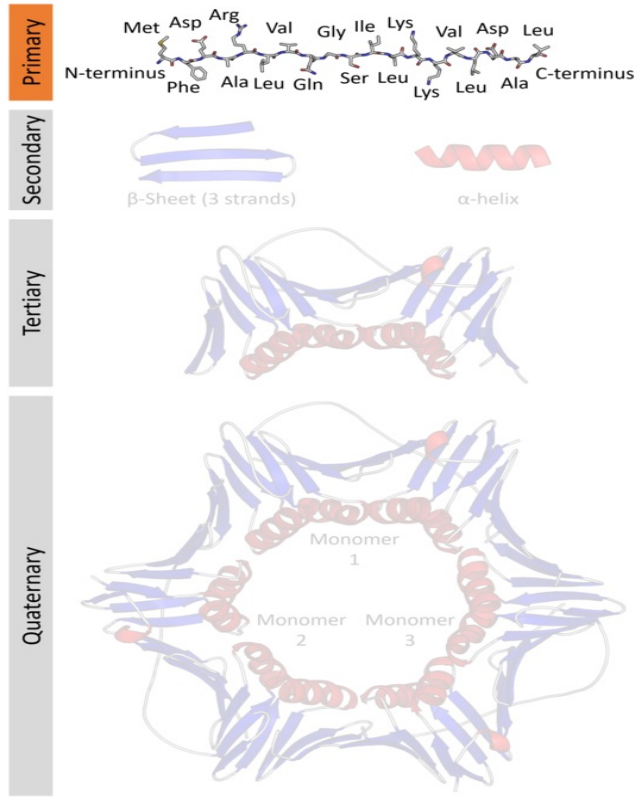
Goal

- generating a.a. sequences from certain protein family
- similar to character-level text generation problem

20 natural amino acid notation

Amino Acid	3–Letter[4]	1–Letter[4]
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

(a) natural amino acid



(b) protein structure

Source: https://en.wikipedia.org/wiki/Protein_primary_structure#Biological

char-RNN:trainning process

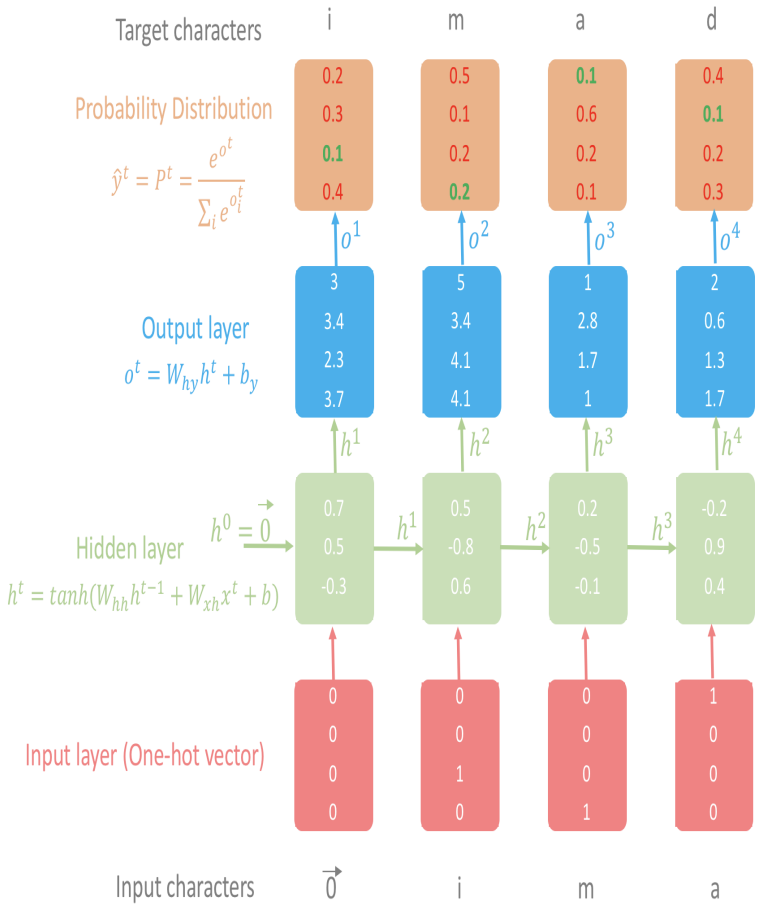


Figure: trainning process

Source: `https://towardsdatascience.com/character-level-language-model-1439f5dd87fe`

char-RNN:generating process

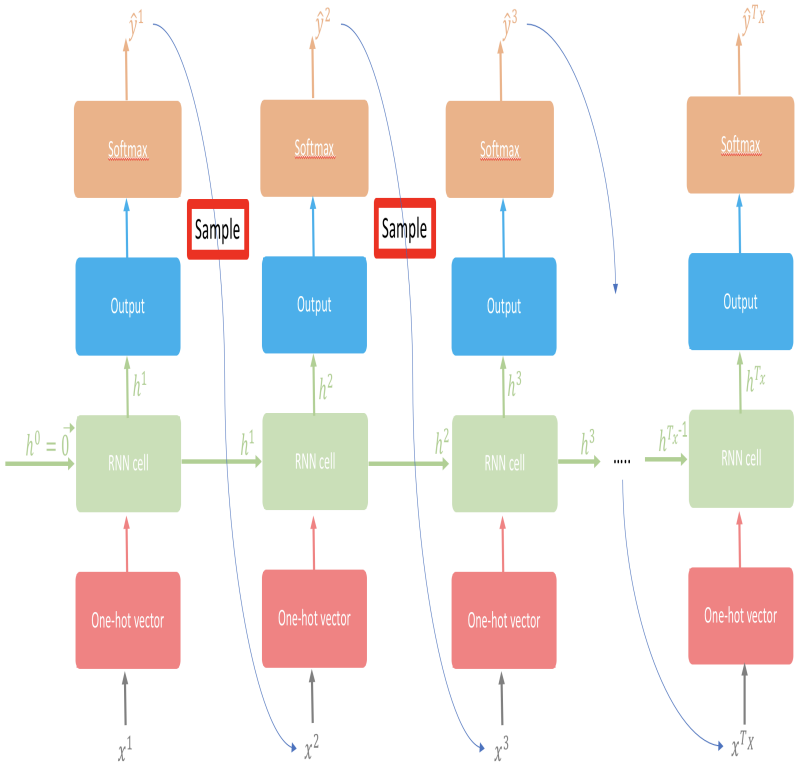


Figure: generating process

Source: `https://towardsdatascience.com/character-level-language-model-1439f5dd87fe`

char-RNN:generating process

```
(array([1.090e+02, 1.799e+03, 2.791e+03, 4.691e+03, 3.016e+03, 1.394e+03,
        4.900e+02, 1.300e+02, 2.000e+01, 1.000e+00]),
 array([116. , 123.6, 131.2, 138.8, 146.4, 154. , 161.6, 169.2, 176.8,
        184.4, 192. ]),
 <a list of 10 Patch objects>)
```

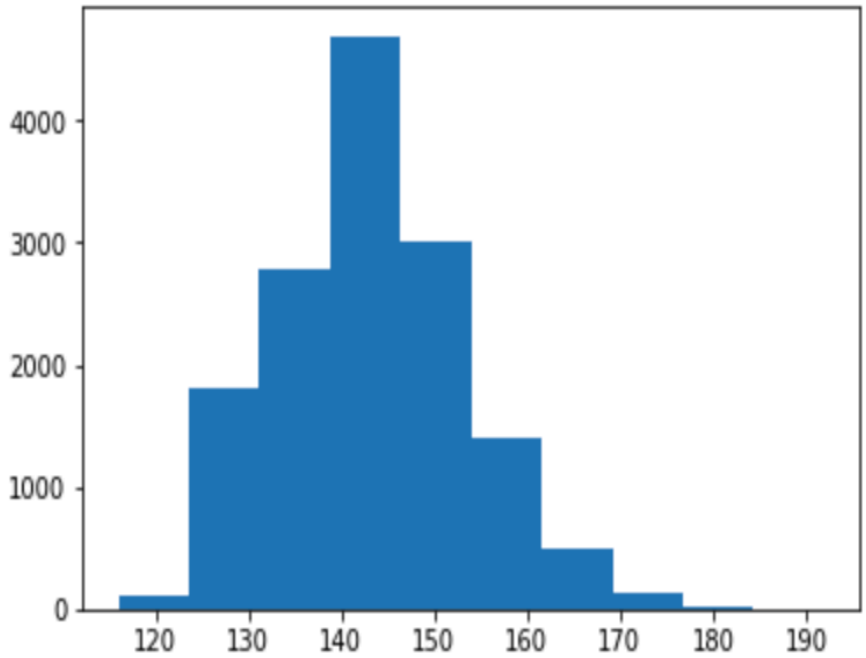


Figure: distance distribution

Transformer

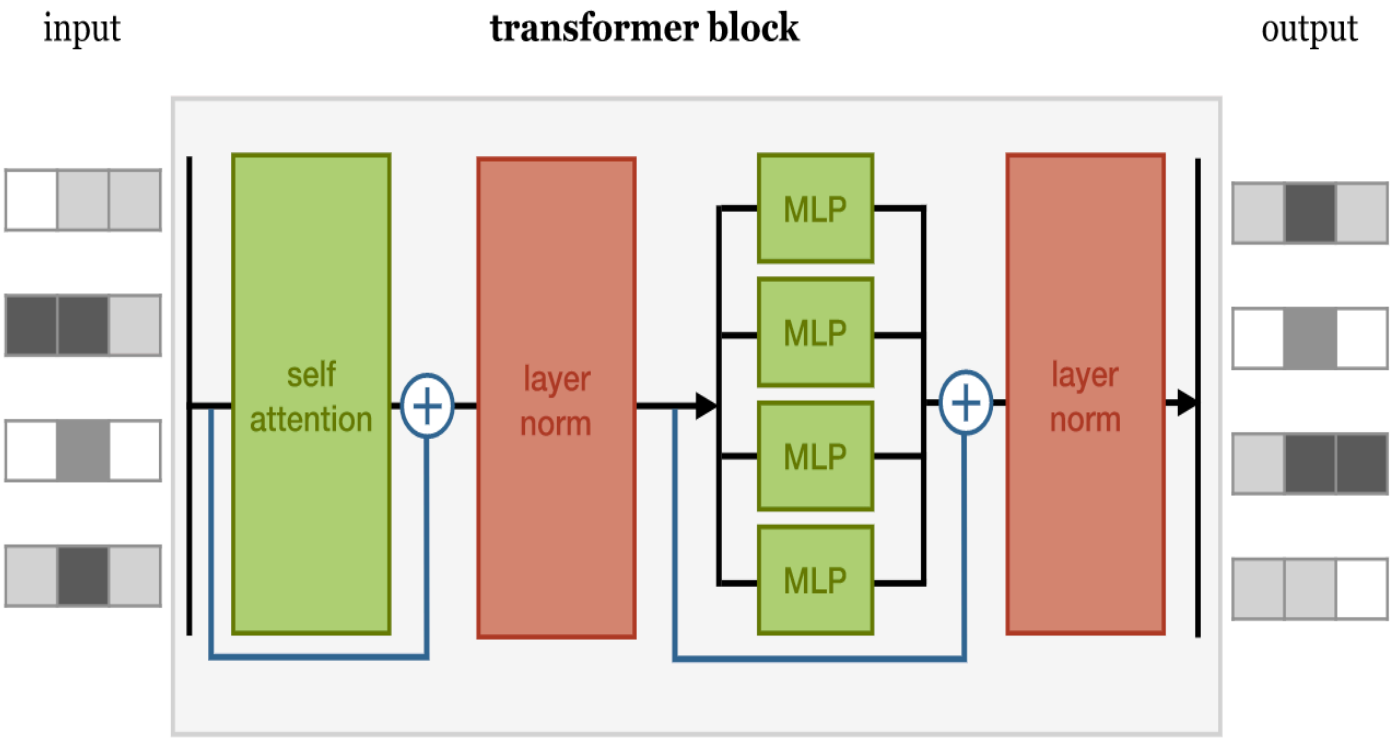


Figure: structure of transformer

Source: `http://peterbloem.nl/blog/transformers`

self-attention mechanism

- more like embedding procedure (no learning parameter)
- take relationship between words into consideration
- input:$x_1, ..., x_t$
- output:$y_i = \sum_j w_{ij} x_j$
- weight:w_{ij}softmax$(w'_{ij} = x_i^T x_j)$