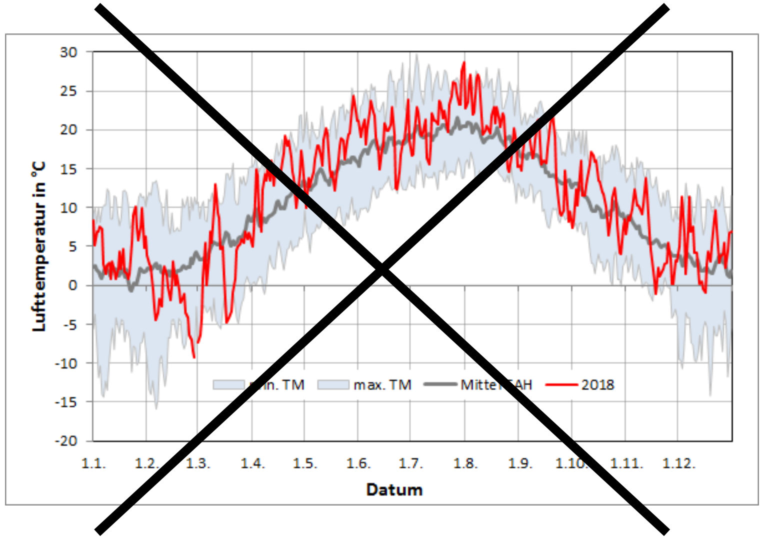
Florian Mock

How to use sequences in Deep Learning

Sequences, could you be even more vague?



MEACCMELVKC

TACCTTGGC…

- Typically:
  - Proteins
  - DNA
  - RNA
- Generally:
  - Order important
  - Text representation

http://wetter.mb.eah-jena.de/station/statistik/rueckblick18.html

What‘s the problem?

NEURAL NETWORKS USE NUMERICAL INPUT

WHAT ARE GOOD NUMBERS FOR SEQUENCES?

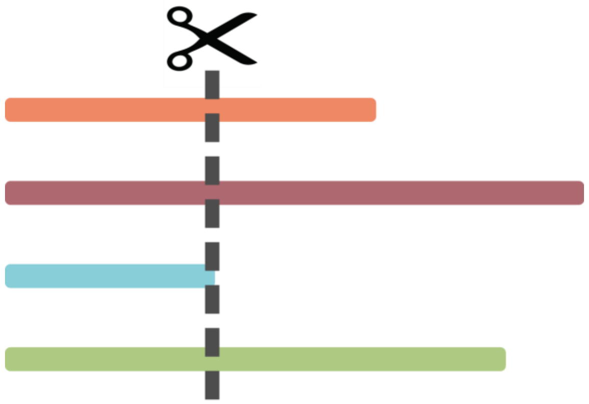
NEURAL NETWORKS EXPECT A CONSISTENT INPUT SIZE

HOW DO WE TREAT UNEQUAL SEQUENCE SIZES?

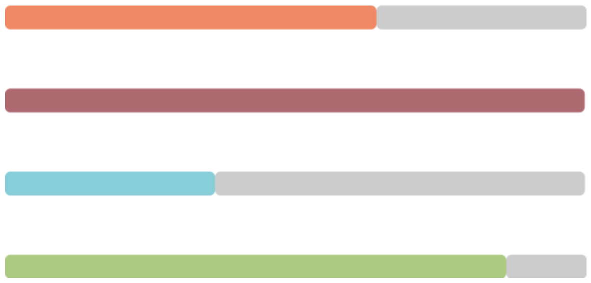
Start simple

```
A = [0,0,0,0,1]
C = [0,0,0,1,0] …
N = [1,0,0,0,0]
```

One hot encoding

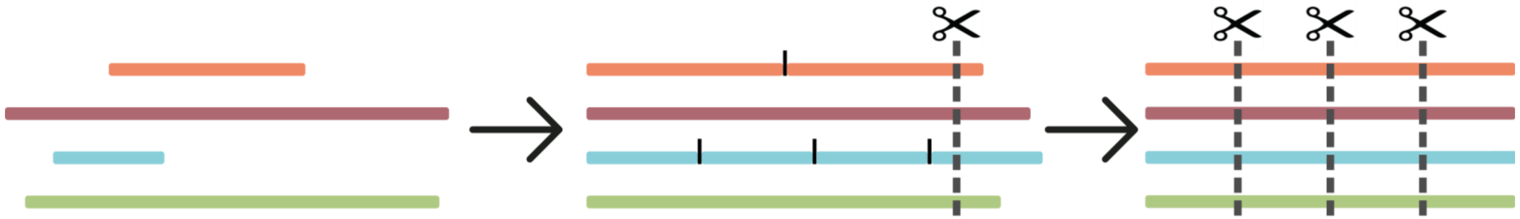


Trimming



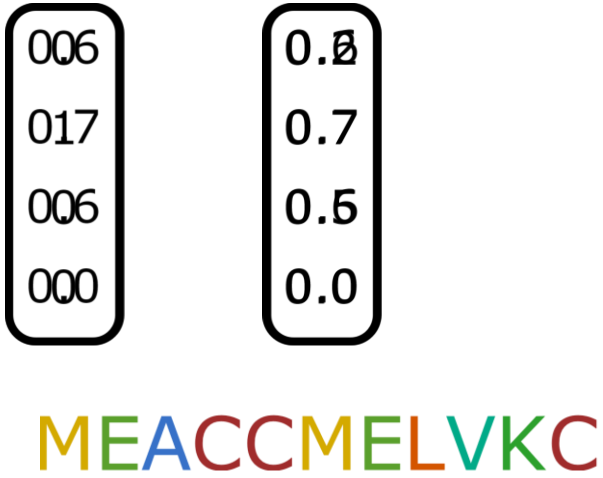
Appending

How to improve length unification?



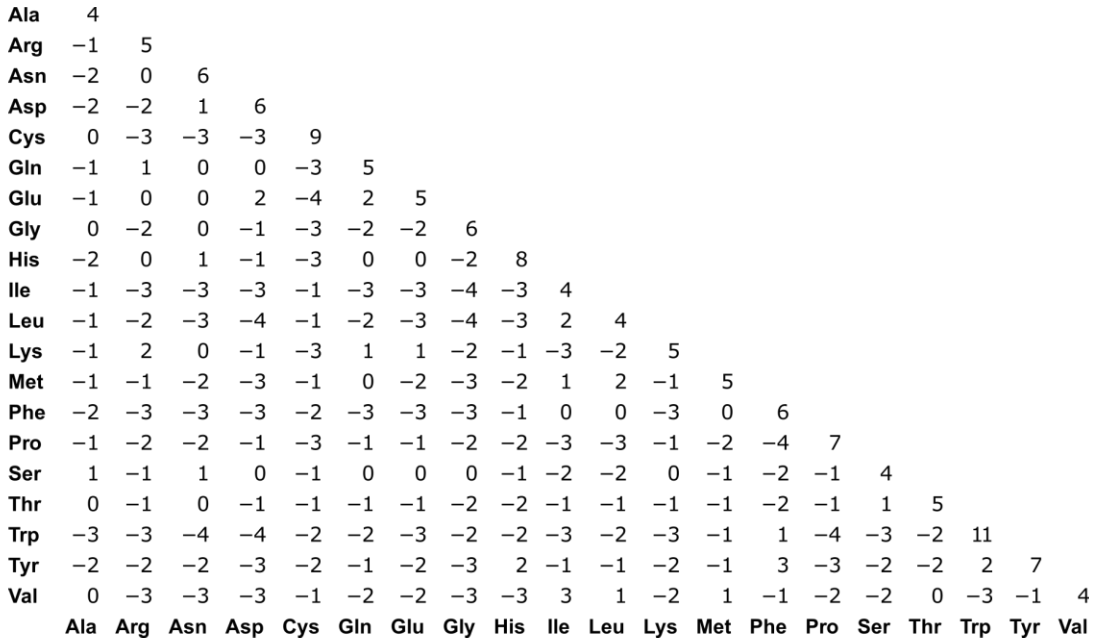
- Select sequence length depended of model layers
  - E.q. LSTM < 400
- Append with content
- Truncate now, join later

How to improve embedding?



- Problems:
  - When using one hot encoding, network needs to learn properties of each letter
  - Each letter has no context information

How to improve embedding?

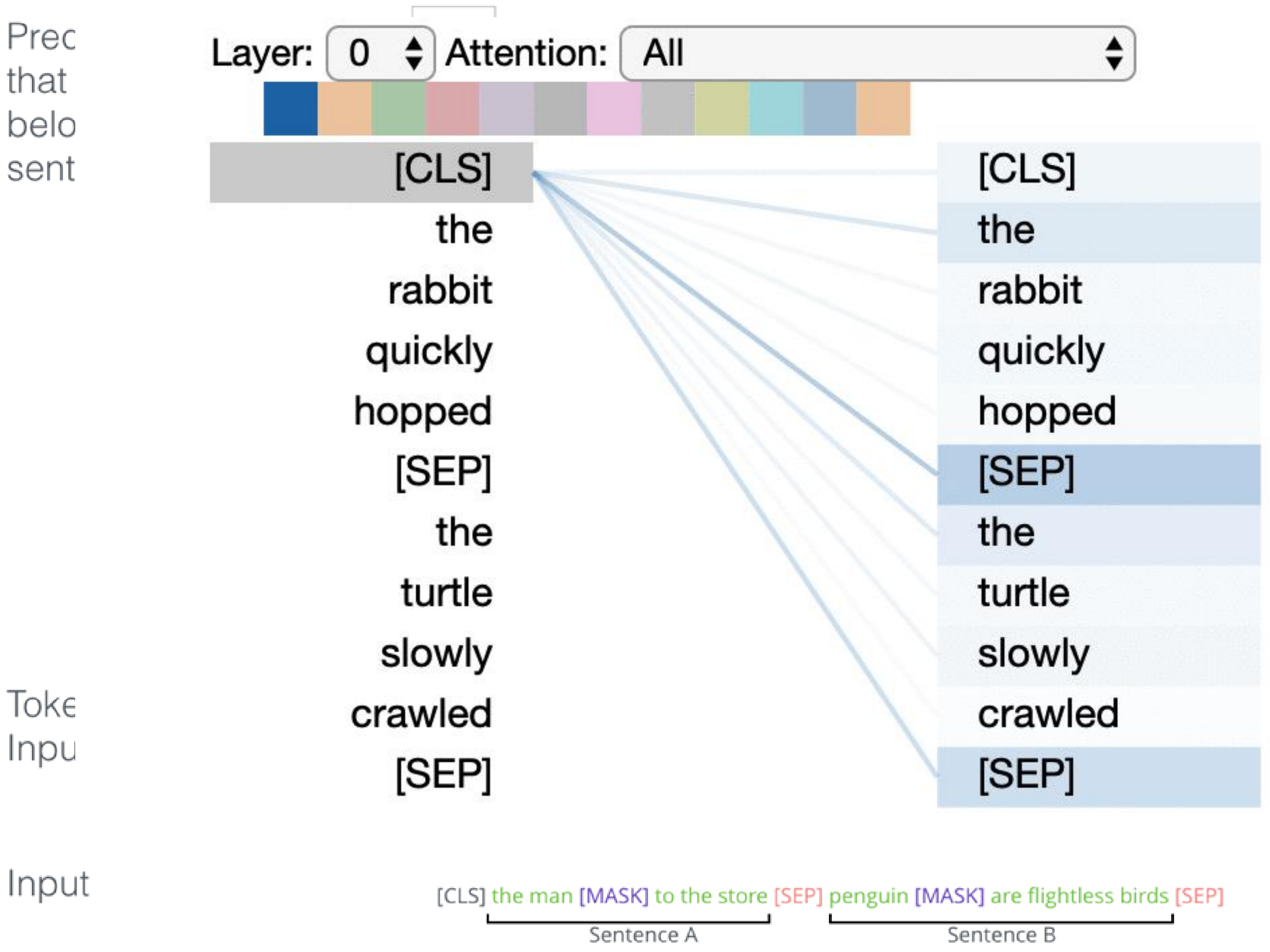


Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

- Potential solutions:
  - Encode properties (measurements, Blosum62 …)
  - Use different embedding per letter dependent on context

https://en.wikipedia.org/wiki/BLOSUM#/media/File:BLOSUM62.png

How to generate context sensitive embeddings?

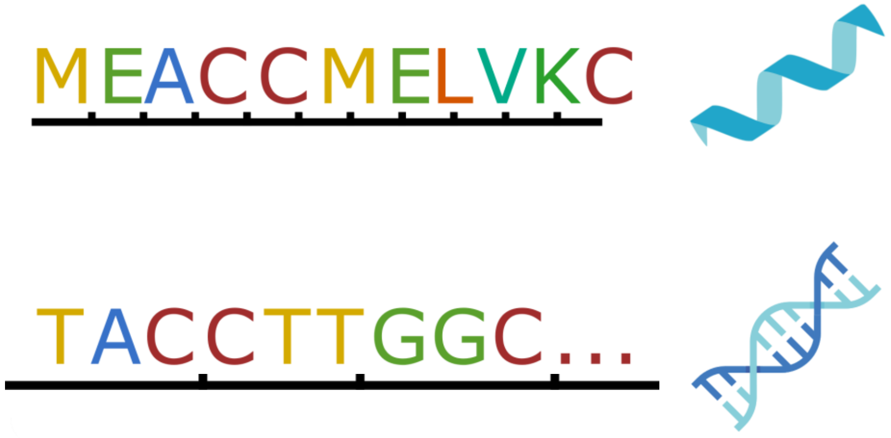


Layer: 0 Attention: All

- Idea we treat our sequence as language
- Words < Sentence < Document

http://jalammar.github.io/illustrated-bert/
https://towardsdatascience.com/deconstructing-bert-part-2-visualizing-the-inner-workings-of-attention-60a16d86b5c1

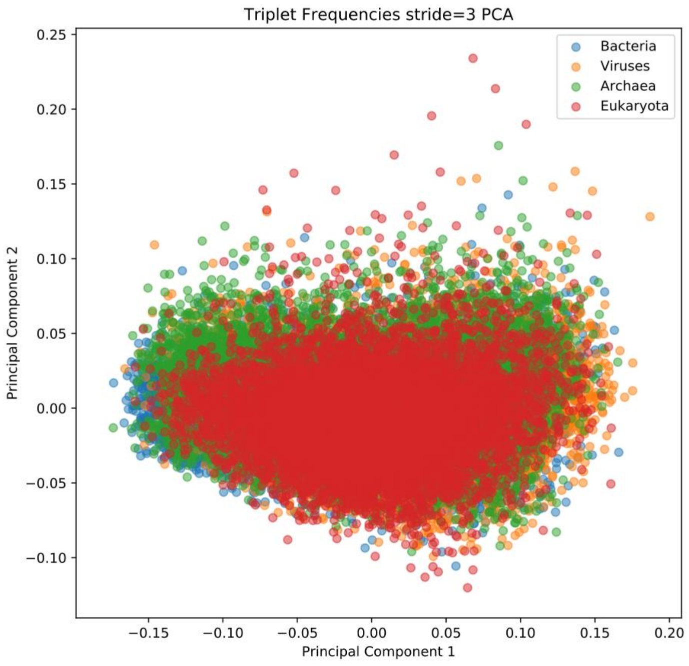
How to generate context sensitive embeddings of sequences?

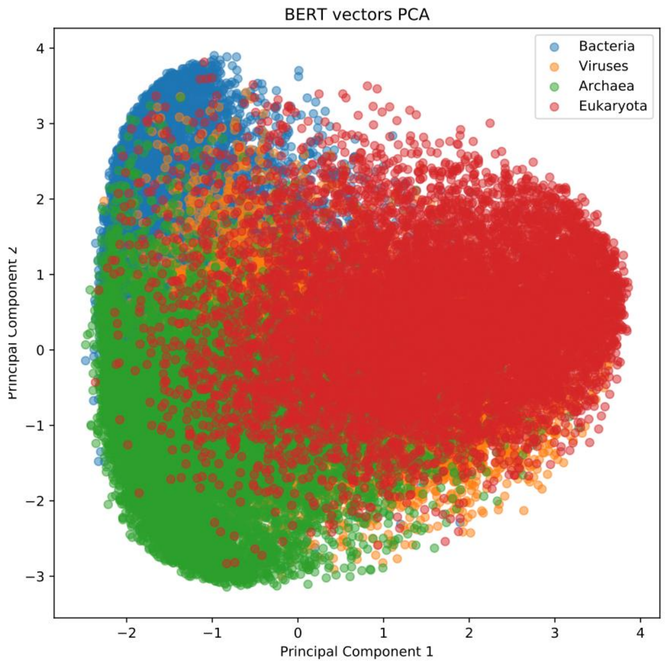


- Idea we treat our sequence as language
- Words < Sentence < Document
- Aminoacid < Proteindomain < Protein
- Triplet < multiple Triplets < Sequence

http://jalammar.github.io/illustrated-bert/
https://towardsdatascience.com/deconstructing-bert-part-2-visualizing-the-inner-workings-of-attention-60a16d86b5c1

Does this work?


Triplet Frequencies stride=3 PCA


BERT vectors PCA

- Trained on UniRef50
  - Word length 1, stride 1, sentence 1024 words
- Trained on DNA of same proteins
  - Word length 3, stride 3, sentence about 80 words

https://github.com/mheinzinger/SeqVec

What comes next?

DNA/RNA/Proteins != human language

Needs to rethink training and architecture

Thank you for your attention