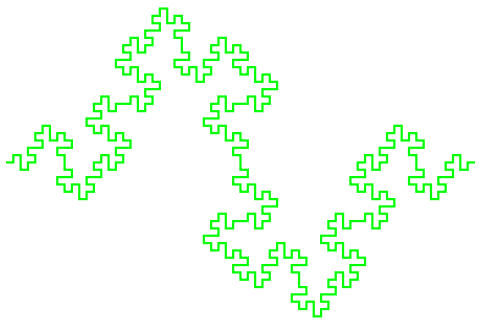
Vector representation of large DNA and protein strings

Bogdan Kirillov Phan Duc Evgenij Baraboshkin

Skolkovo Institute of Science and Technology



April 25, 2018

The key papers to use

- Dna2vec: Consistent vector representation of variable-length k-mers, Ng,P – the one we are to discuss here;
- Distributed Representations for Biological Sequence Analysis, Kimothi et al., ;
- Continuous Distributed Representation of Biological Sequences for Deep Genomics and Deep Proteomics, Asgari, E. and Mofrad, M;

Bioinformatics 101

Biological data are complex!

- Direct Data Augmentation is not really applicable;
- Working with really large (1Mb and more) DNA sequences is unclear;
- Unlike basic CV, there usually are no clue about how well we can solve the problem with given training/testing data;
- A lot of experiments are low-throughput so sample sizes can be limited and the samples may overlap;

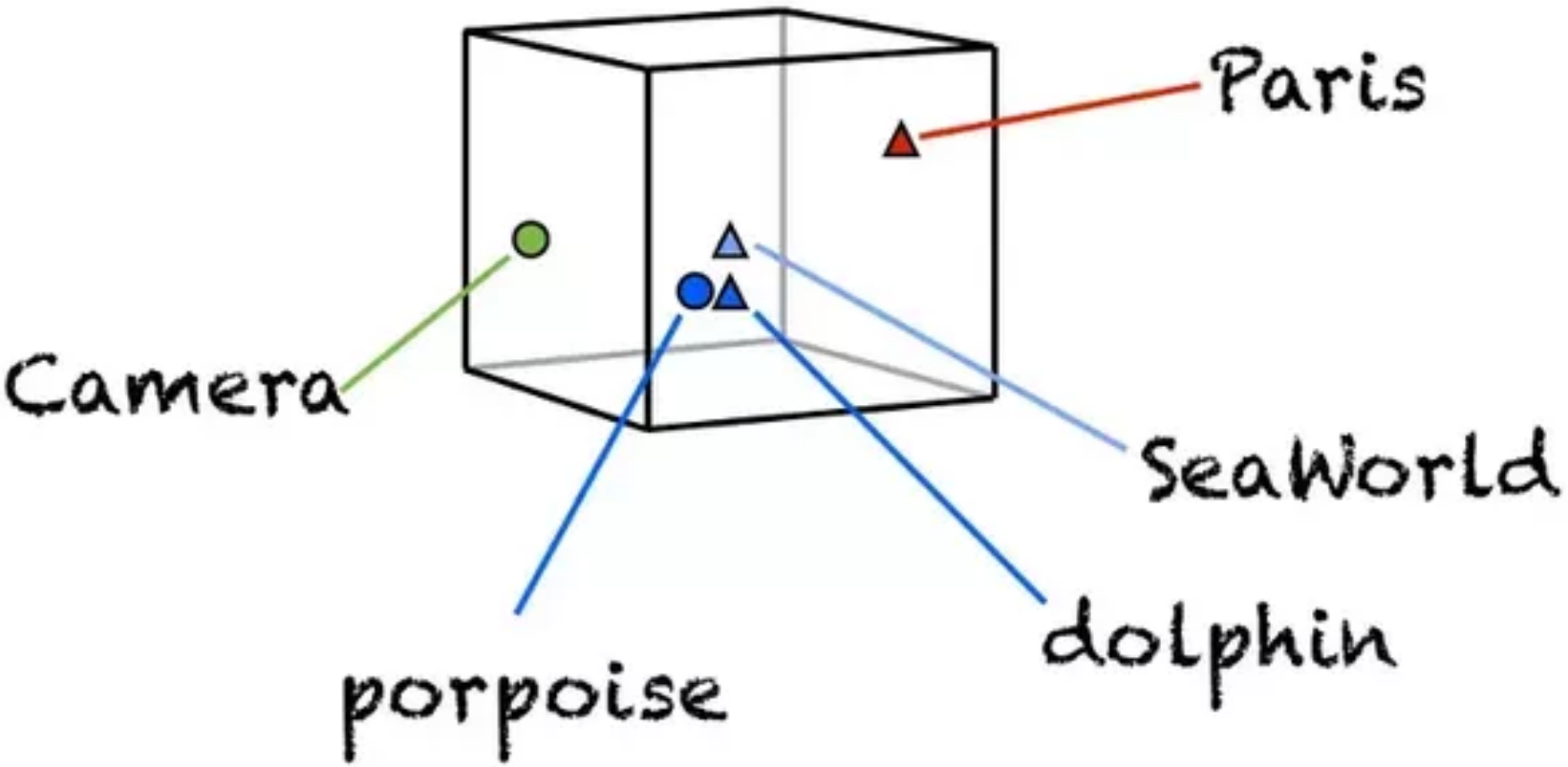
Augmentation in Computer Vision

Cropped and mirrored cat pic remains a cat pic

Biological sequence case

You can't say the same here

Word embeddings and their use for DNA/Proteins



Example of word embedding

Problem of k-mer length choice

dna2vec training procedure

Stage 1: Long non-overlapping DNA fragments

Stage 2: Overlapping variable-length k-mers

Stage 3: Two-layer neural network

Stage 4: Decompose aggregated model by k-mer lengths

Alignment similarity

Possible applications

What can be improved?