CIML Summer Institute:

4.1 Writing and Sharing Computational Analyses in Jupyter Notebooks

June 24, 2021

Peter Rose
SDSC

EXPANSE
COMPUTING WITHOUT BOUNDARIES

SAN DIEGO SUPERCOMPUTER CENTER

NSF Award 1928224

Tools and Infrastructure

Computational notebooks: combine documentation, code, and results

Scalable compute infrastructure

Open cloud environment to run computational notebooks

Open-source package and environment management system

Version-control system for tracking changes in source code

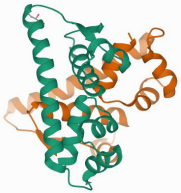
Source code repository

Classification Problem: Predict Protein Fold Class

Protein Sequence

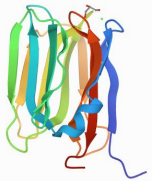
Fold Class

TNKELQAIRKLLMLDVSEAAEHIGRVSARSWQYWESGRSAVPDDVEQEML
DLASVRIEMMSAIDKRLADGERPKLRFYNKLDEYLADNPDHNVIGWRLSQS
VAALYYTEGHADLI



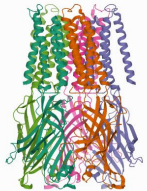
alpha

GARSSSYSGEYGSGGGKRFSHSGNQLDGPITALRVRVNTYYIVGLQVRYG
KVWSDYVGGRNGDLEEIFLHPGESVIQVSGKYKWYLKKLVFVTDKGRYLSF
GKDSGTSFNAVPLHPNTVLRFISGRSGSLIDAIGLHWDVYPSSCSRC



beta

APADNAADARPVDVSVSIFINKIYGVNTLEQTYKVDGYIVAQWTGKPRKTPGD
KPLIVENTQIERWINNGLWVPALEFINVVGSPDTGNKRLMLFPDGRVIYNARFL
GSFSNDMDFRLFPFDRQQFVLELEPFSYNNQQLRFSDIQVYTENIDNEEIDEW
WIRGKASTHISDIRYDHLSSVQPNQNEFSRITVRIDAVRNPSYYLWSFILPLGLII
AASWSVFWLESFSERLQTSFTLMLTVVAYAFYTSNILPRLPYTTVIDQMIIAGYG
SIFAAILLIIFAHHRQANGVEDDLLIQRCRLAFPLGFLAIGCVLVIRGITL



alpha+beta

N-grams and Word2Vec Models

Word-level unigrams

Text

One Two Three Four
One Two Three Four
One Two Three Four
One Two Three Four

Word-level bigrams

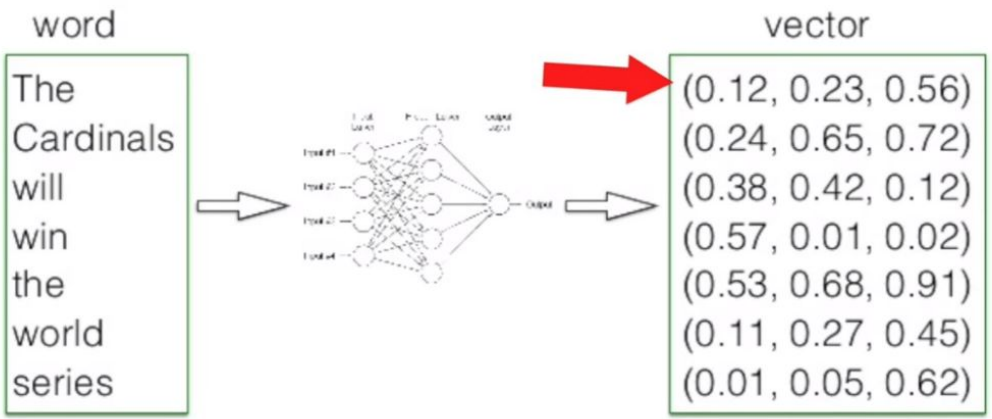
Text

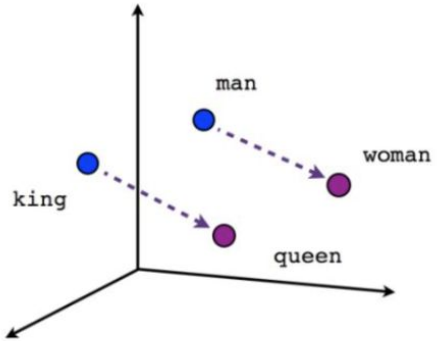
One Two Three Four
One Two Three Four
One Two Three Four

Word-level trigrams

Text

One Two Three Four
One Two Three Four





Male-Female

https://spark.apache.org/docs/latest/mllib-feature-extraction.html#word2vec

Embedding a Protein Sequence

Sequence:
TNKELQAIRKLL…

3-grams (“words”):
TNK, NKE, KEL, ELQ, …

Word2Vec (100-dimensional vector):
[-2.23197367481583, -0.465980592717598, …]

Pre-trained Word2Vec model trained on 546,790 protein sequences: **ProtVec**

Asgari E, Mofrad MR (2015) Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics, PLoS One. 10(11):e0141287. doi: [10.1371/journal.pone.0141287](https://doi.org/10.1371/journal.pone.0141287).

Transfer Learning

Sequence

TNKELQAIRKLL…

3-grams

TNK, NKE, KEL, ELQ, …

ProtVec Model

Feature Vector (embedding) 100-dimensional

[-2.23197367481583, -0.465958059271759 8, …]

Downstream Classification Models

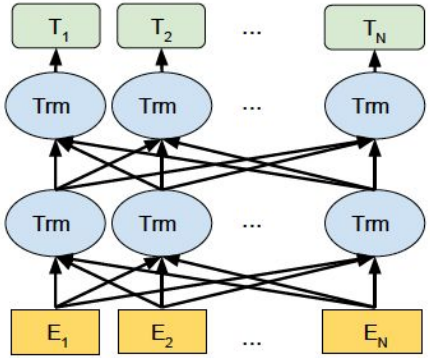
- SVM
- Logistic Regression
- Neural Network

Pretrained BERT Models

Protein sequences

Mask amino acids in protein sequence

Pretrained BERT Model



Input: DP[MASK1]KDSKAQVSAAE[MASK2]GIT…

Labels: [MASK1] = S; [MASK2] = A

For small datasets use embeddings as feature vectors.

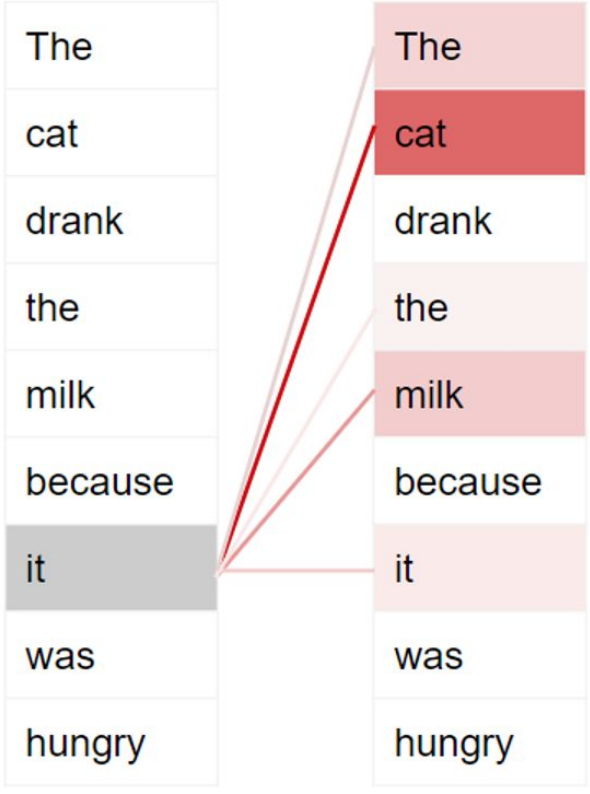
Embeddings (weights) as feature vectors for ML models

Specific prediction tasks

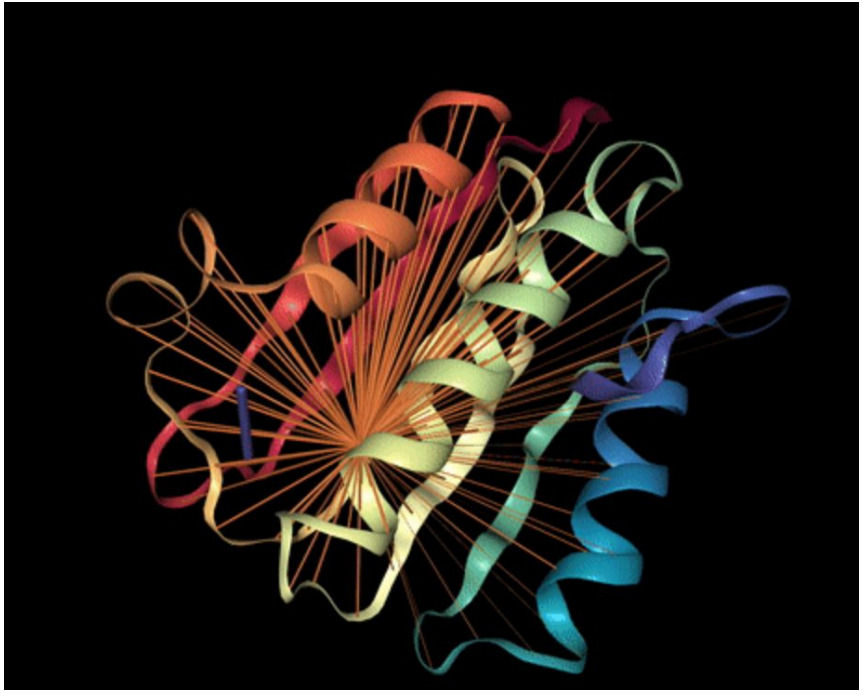
- Protein function
- Protein properties
- Structural features

Attention Mechanism (long range interactions)

Text



Protein Sequence



https://medium.com/deep-learning-digest/bert-model-restores-protein-structure-1171299b963d

Pretrained BERT Models

Supervised downstreams

Model	Input	Pre-training	Params	SSP	Contact
UniRep	Sequence	UR50*	18M	58.4	21.9
SeqVec	Sequence	UR50*	93M	62.1	29.0
TAPE	Sequence	PFAM*	38M	58.0	23.2
ProtBert-BFD	Sequence	BFD*	420M	70.0	50.3
Prot-T5-XL-BFD	Sequence	BFD*	3B	71.4	55.9
LSTM biLM (S)	Sequence	UR50/S	28M	60.4	24.1
LSTM biLM (L)	Sequence	UR50/S	113M	62.4	27.8
Transformer-6	Sequence	UR50/S	43M	62.0	30.2
Transformer-12	Sequence	UR50/S	85M	65.4	37.7
Transformer-34	Sequence	UR100	670M	64.3	32.7
Transformer-34	Sequence	UR50/S	670M	69.2	50.2
ESM-1b	Sequence	UR50/S	650M	**71.6**	**56.9**
ESM-MSA-1	MSA	UR50/S + MSA	100M	**72.9**	Coming Soon

Available as Singularity containers for Expanse (using PyTorch)

Due to high compute and memory demands, we will not use them during this workshop.

https://github.com/facebookresearch/esm

Sharing your Notebooks with MyBinder

https://mybinder.org/

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

New to Binder? Get started with a Zero-to-Binder tutorial in Julia, Python or R.

Build and launch a repository

GitHub repository name or URL

GitHub ▾ | https://github.com/sdsc-hpc-training-org/notebooks-sharing

Git ref (branch, tag, or commit)

HEAD

Path to a notebook file (optional)

Path to a notebook file (optional) | File ▾

launch

Public Cloud Environments

Platform	URL	Memory	Cores	Use for	Comments	Account
MyBinder	https://mybinder.org/	2GB	1	small examples	some ports are blocked	no
Pangeo Binder	https://binder.pangeo.io/	32GB (?)	6 (?)	when exceeding MyBinder limits	open ports, e.g., FTP	no
Google Colab	https://research.google.com/colaboratory/	variable	?	GPU/TPU	software installations using pip in Notebook, share notebooks on Google Drive	yes
CyVerse	https://cyverse.org/discovery-environment	per request	per request	store notebooks, results, and data	100GB storage	yes

Demo

Questions?