

A complex network graph with numerous nodes represented by small circles of varying sizes and colors (white, light orange, pink, purple) connected by a dense web of thin white lines. The background has a warm, orange-to-red gradient.

Classification of Family Domain of Amino Acid Sequences using CNN – LSTM Architecture

Gokul S – 2018103026

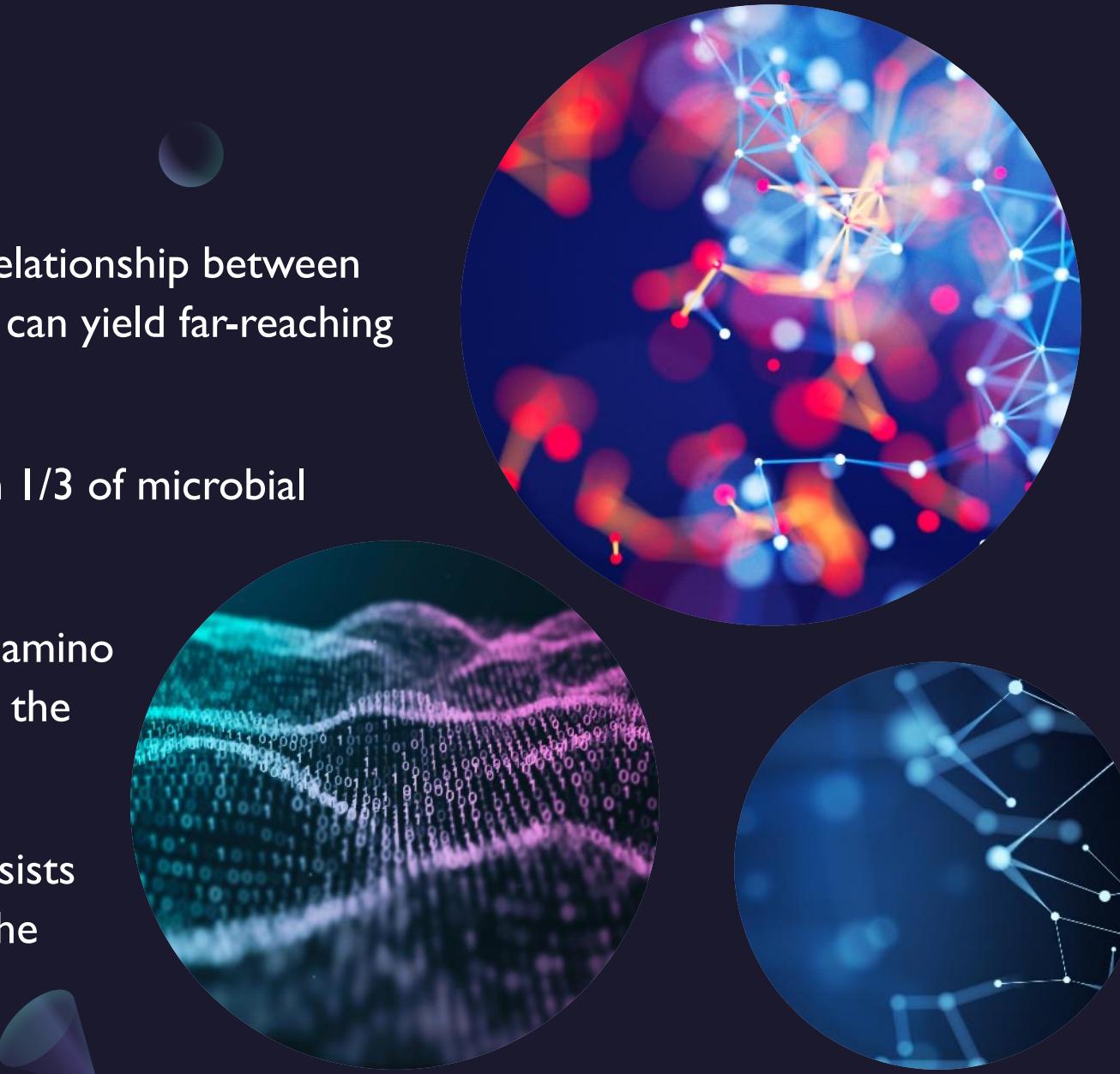
Steven Gilbert – 2018103071

CSE P



Introduction

- In the field of micro biology, understanding the relationship between amino acid sequence and their protein functions can yield far-reaching scientific implications.
- But current lab techniques cannot annotate even 1/3 of microbial protein sequences.
- Predicting the function of a protein from its raw amino acid sequence is a critical step for understanding the relationship between genotype and phenotype.
- We propose a Deep-learning method which consists of a CNN-LSTM hybrid architecture to classify the families of the amino protein sequences.



Pfam

- The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models(HMM).
- For each family in Pfam :
 - Family Description
 - Multiple alignments
 - Protein domain architecture
 - Species Distribution
 - Known protein structures

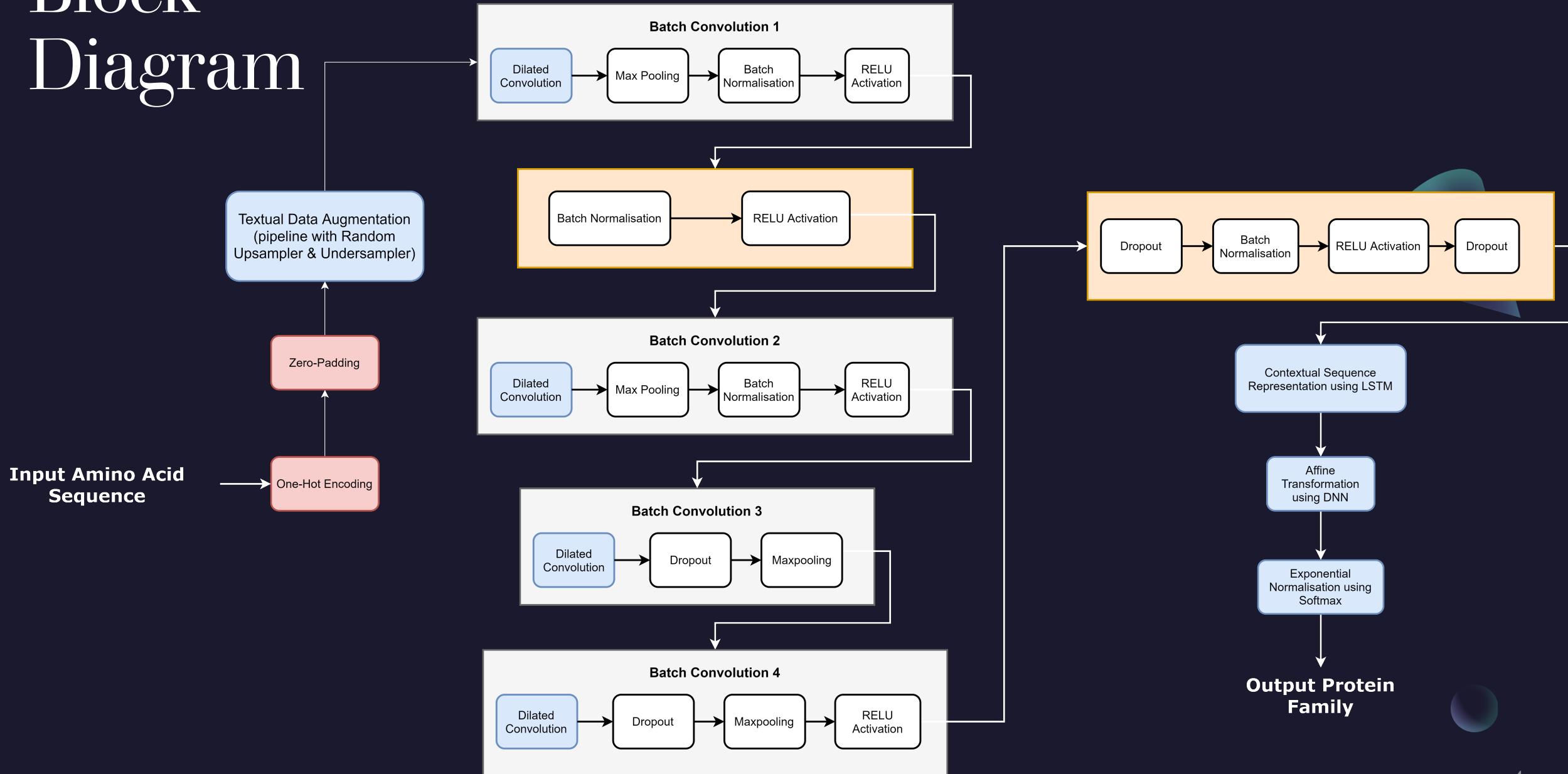
Pfam

```
sequence: HWLQMRDSMNTYNNMVNRCFATCIRSFQEKKVNAEEMDCTKRCVTKFVGYSQRVALRFAE  
family_accession: PF02953.15  
sequence_name: C5K6N5_PERM5/28-87  
aligned_sequence: ....HWLQMRDSMNTYNNMVNRCFATCI.....RS.F....QEKKV  
family_id: zf-Tim10_DDP
```

The Dataset contains nearly :

- I million sequence examples
- Around 18,000 family output classes

Block Diagram

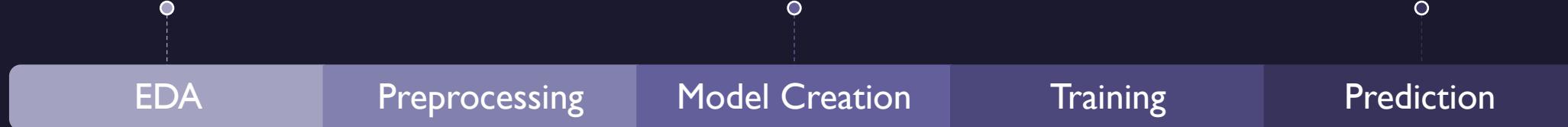


Methodology

In this module, various plots are used to analyze the structure and distribution of the PFAM dataset in order to check the balance among families

The overall CNN-LSTM architecture is concretely defined here and compiled along with predetermined hyperparameters.

The trained model is loaded to perform Amino acid sequence classification of a completely new and unseen input sequence



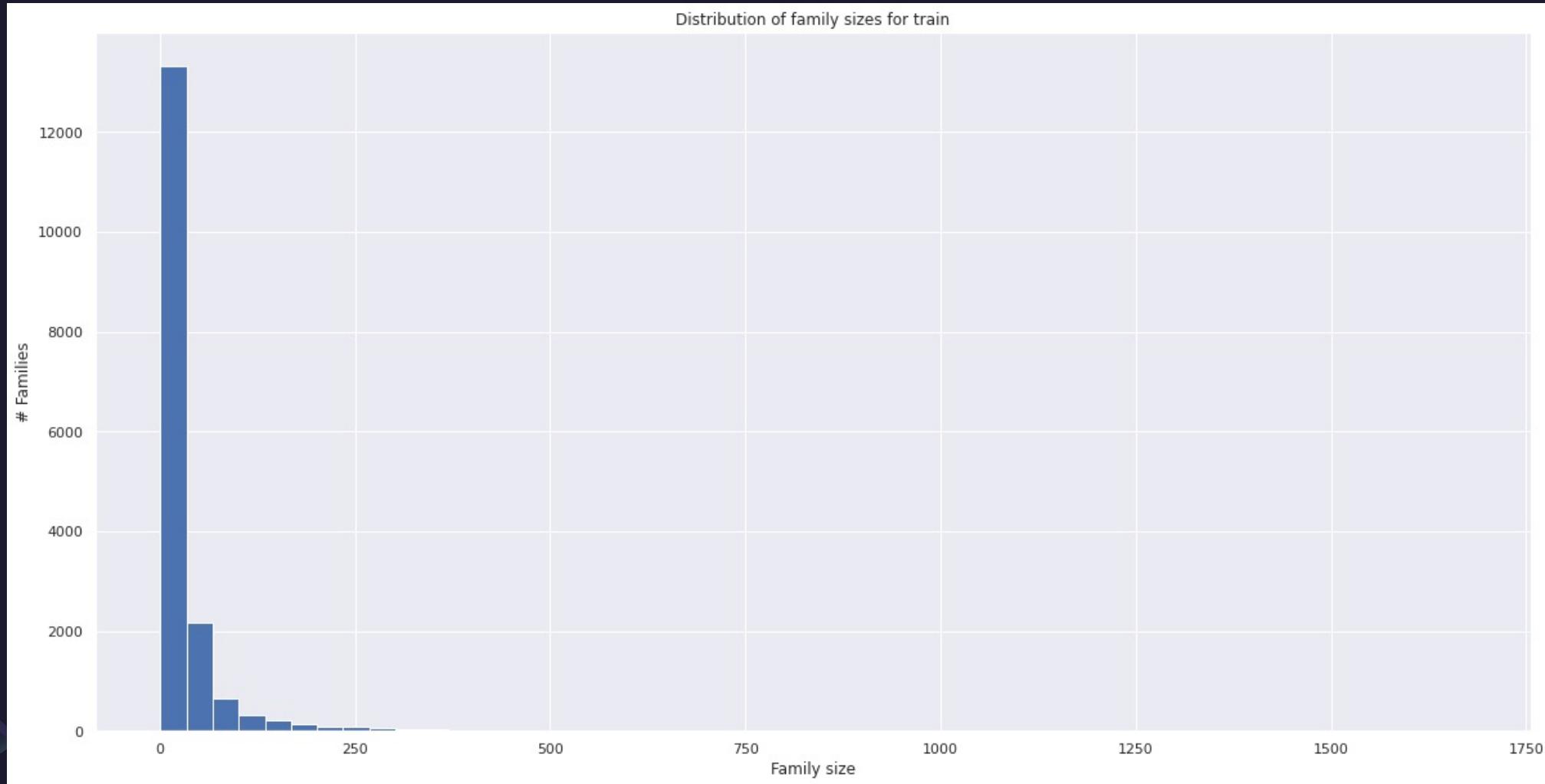
Here we first create a dictionary bases on vocabulary. Then encoding occurs with sequence padding along with vectorization

The model is trained on a **Quadro P5000** GPU for 50 epochs to learn the important characteristic features of each of the individual 500,000 sequences.

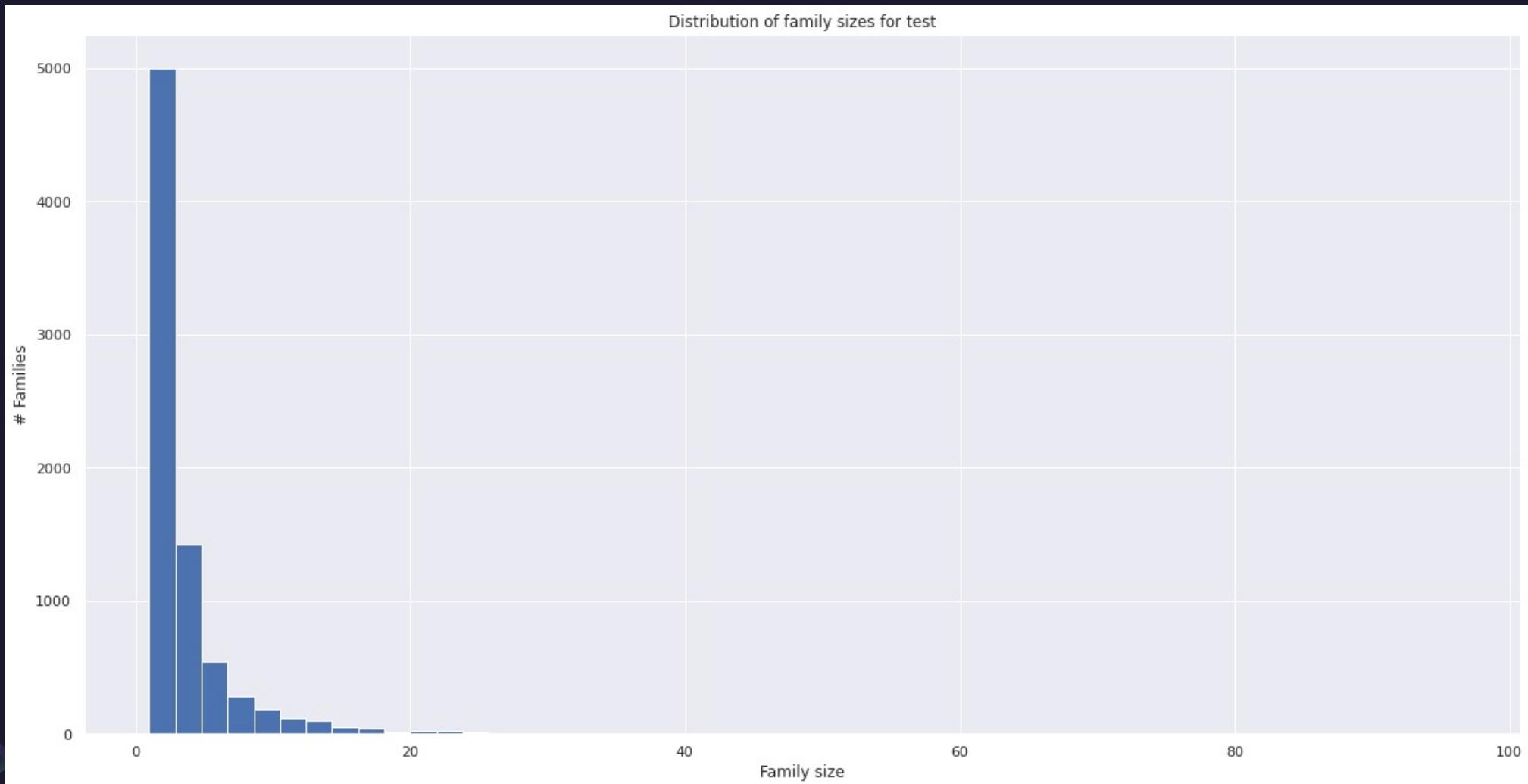


Exploratory Data Analysis

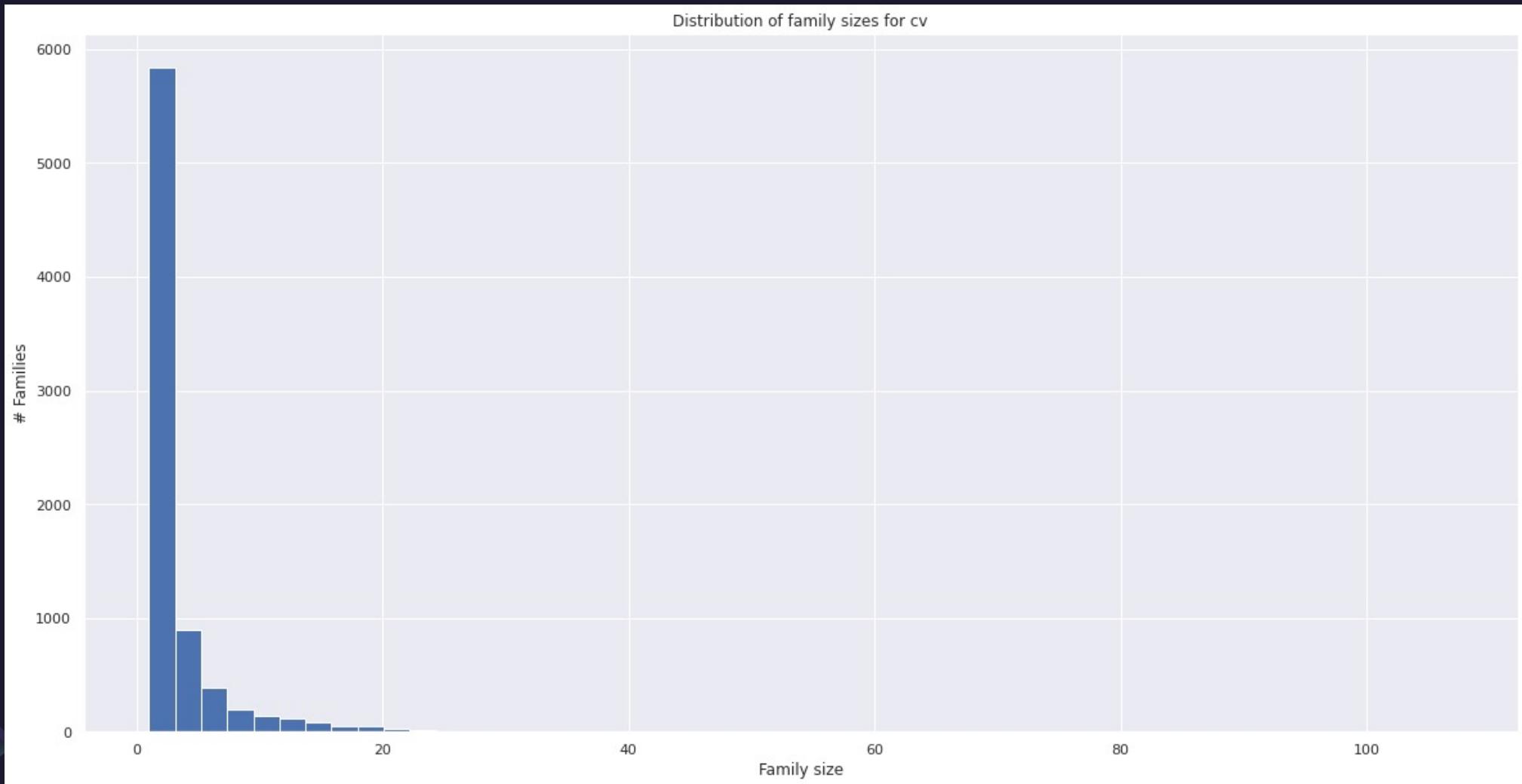
Distribution of Train Family Sizes



Distribution of Test Family Sizes

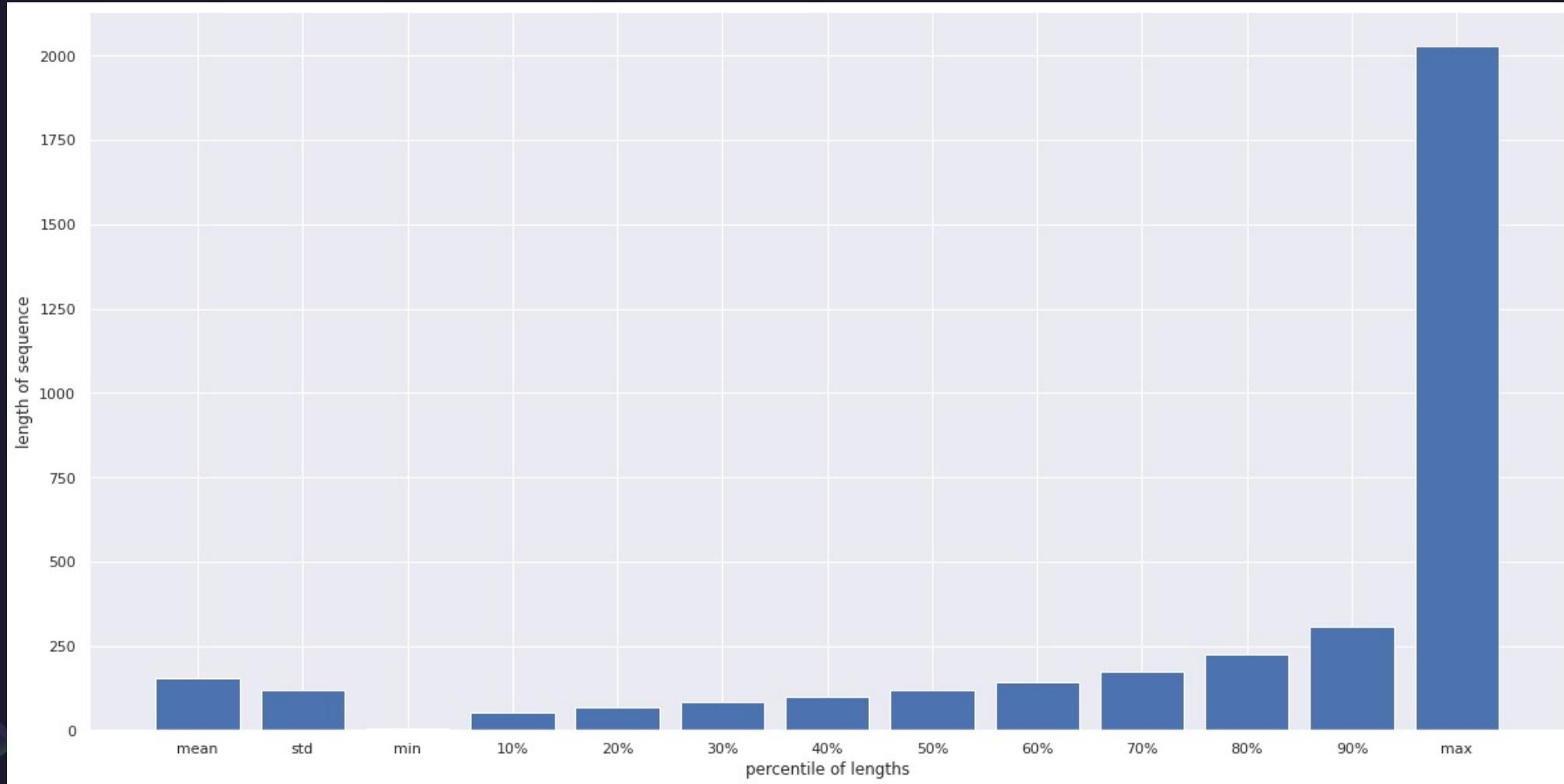


Distribution of Cross Validation (CV) Family Sizes

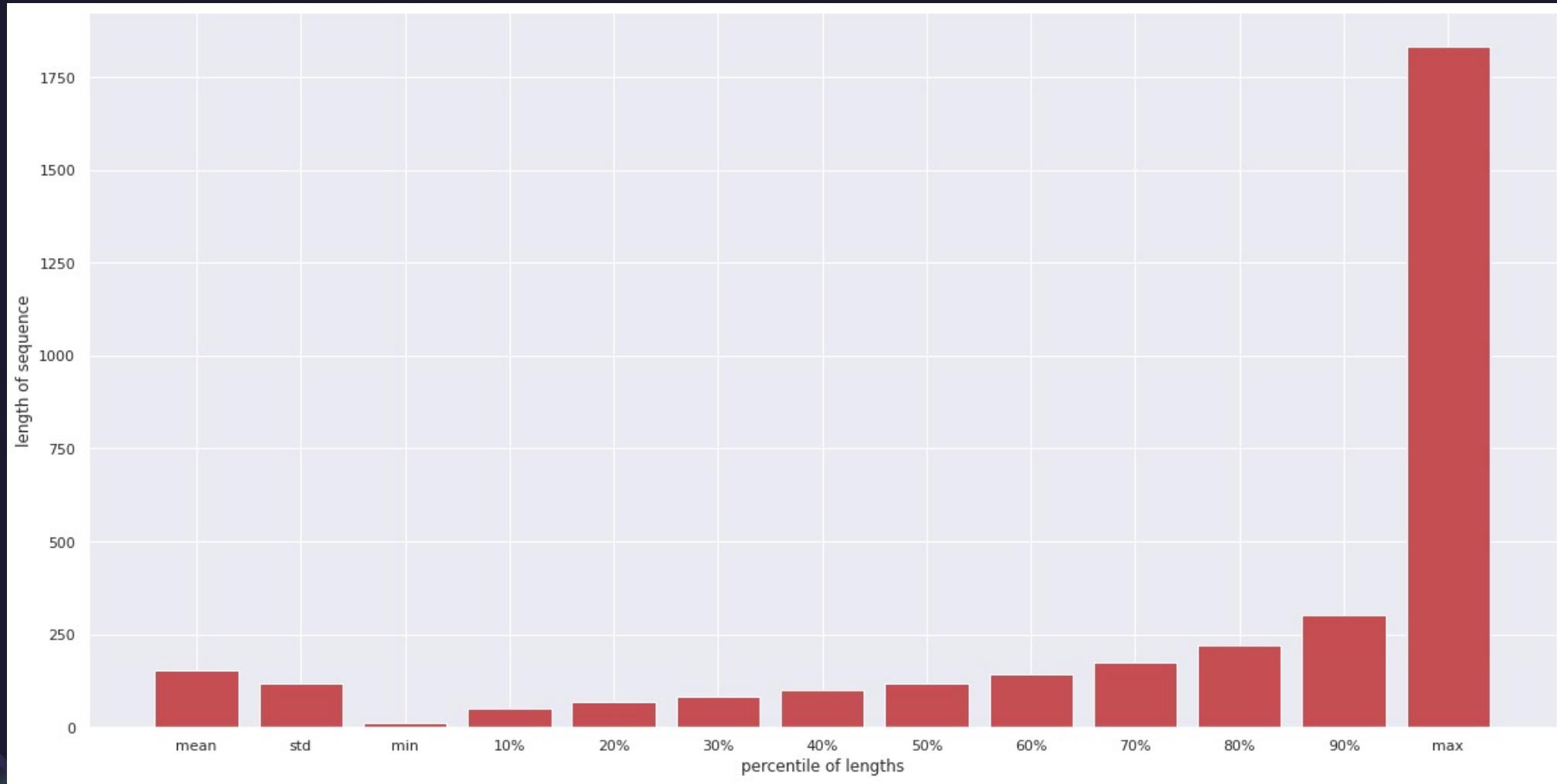


Percentile Distribution of Sequence Length

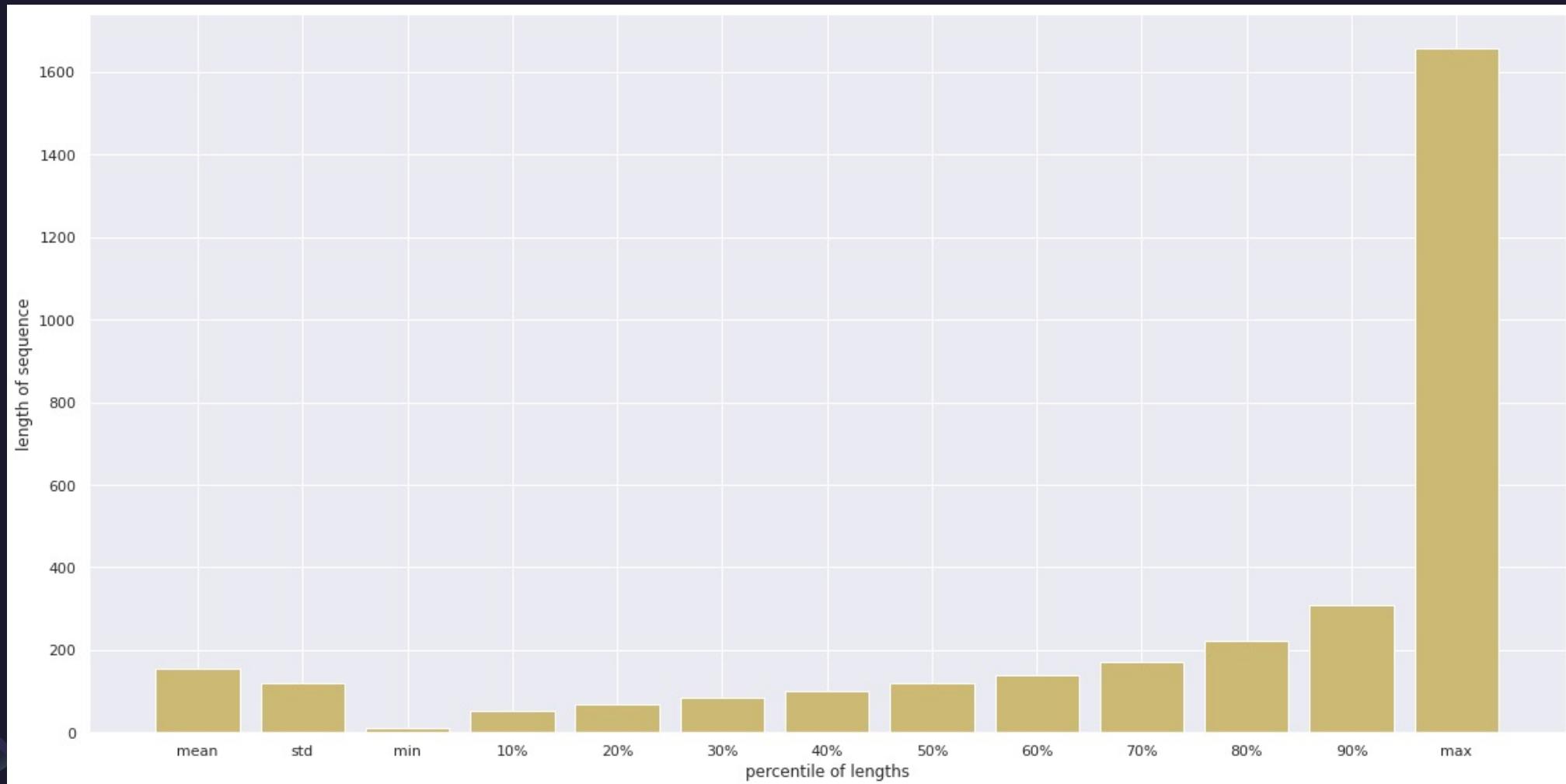
Train



Test

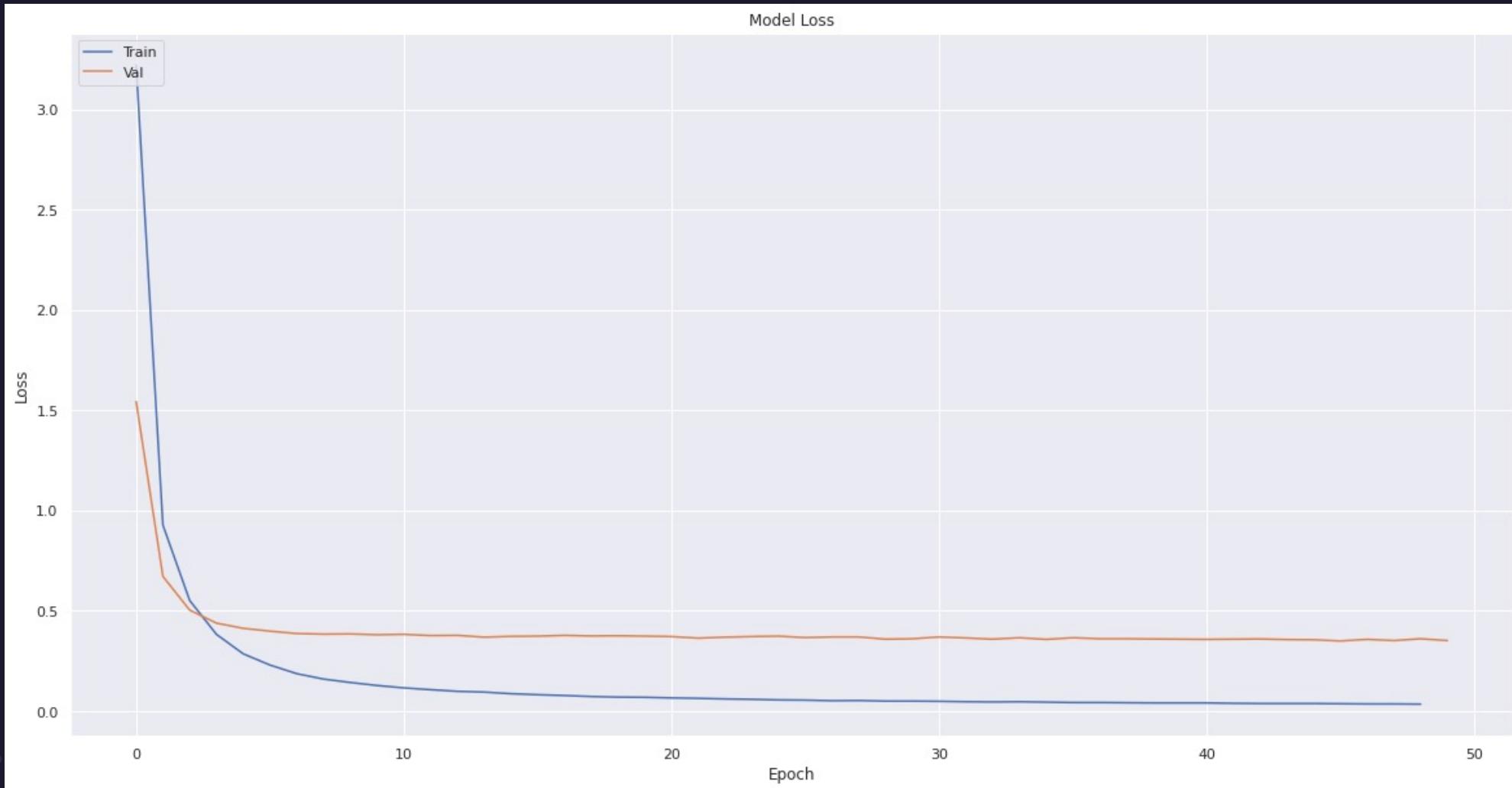


Cross Validation



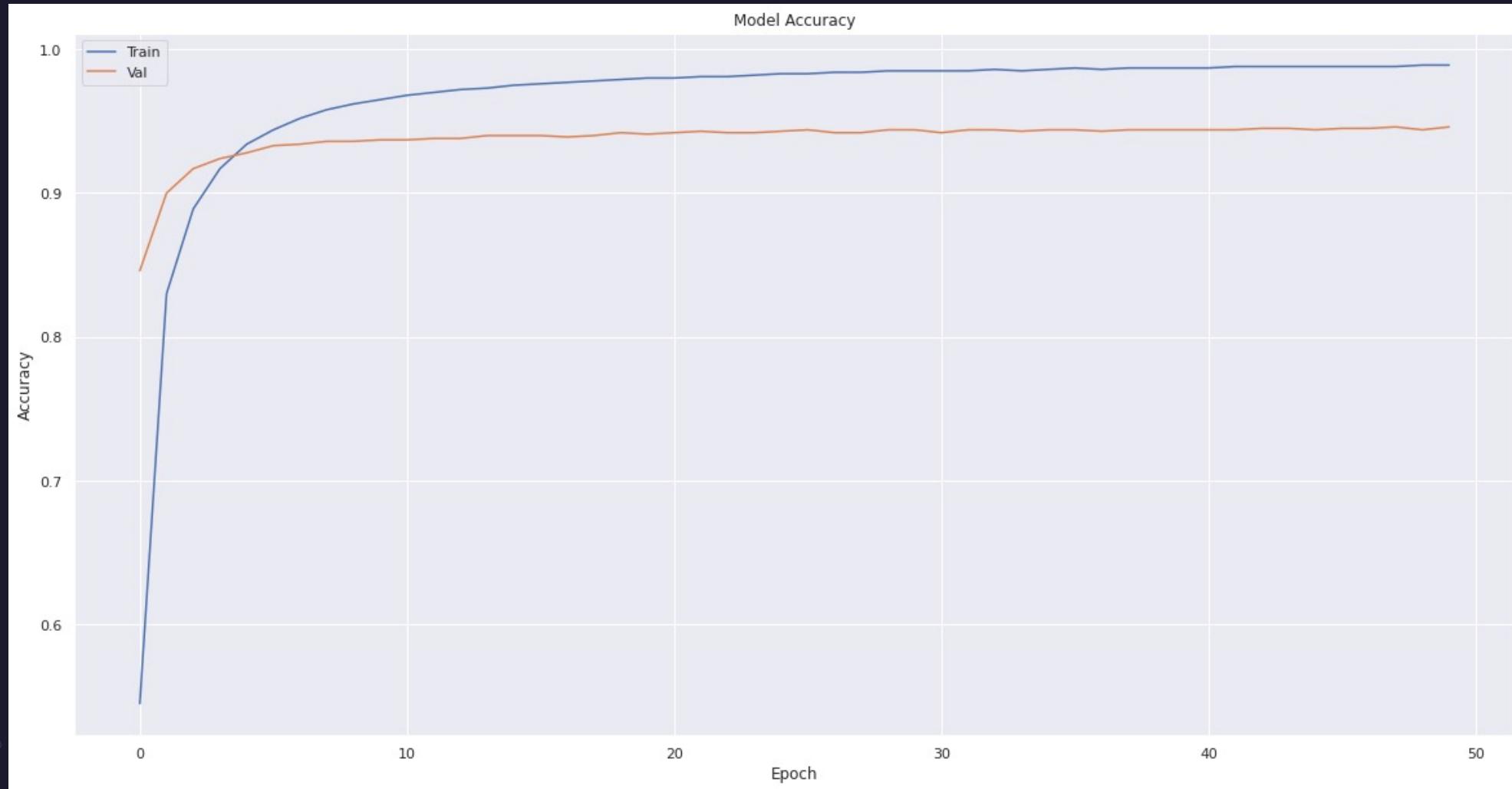
LOSS Curve

Model with LSTM and Dilated 1-D convolution



Accuracy Curve

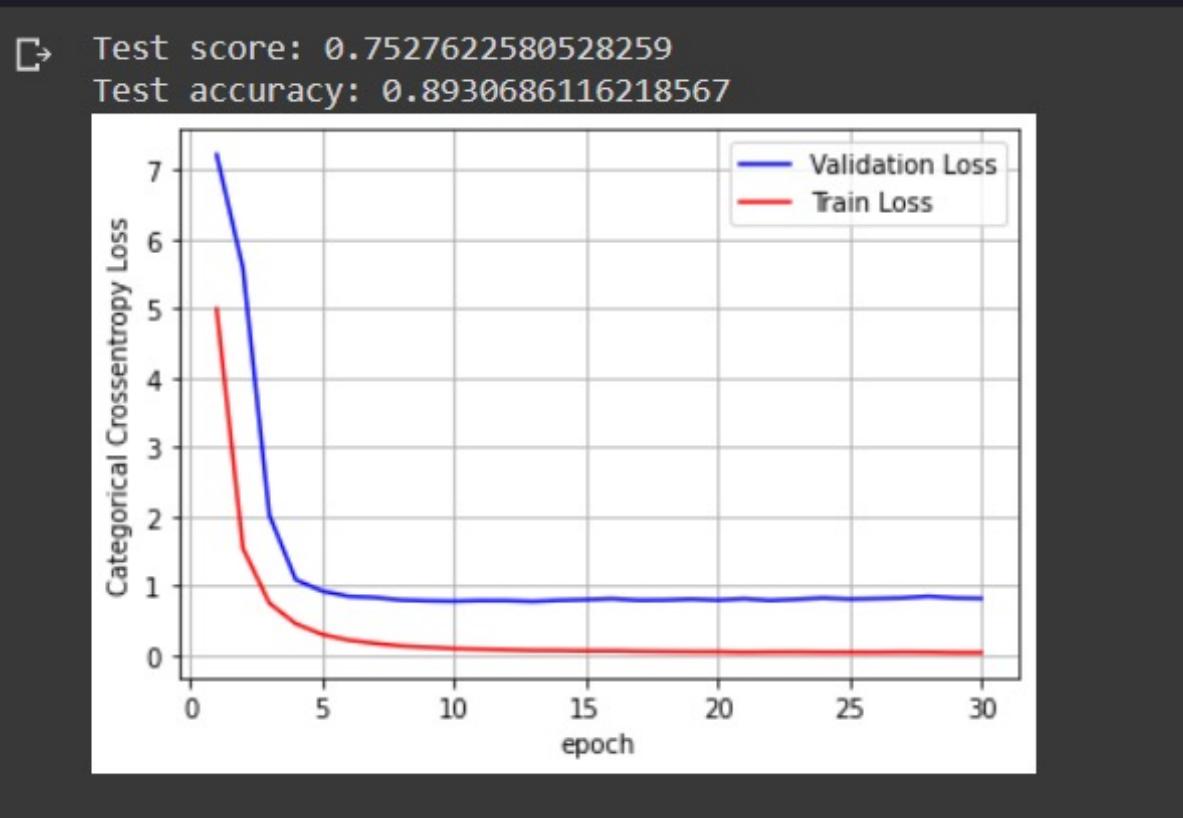
Model with LSTM and Dilated 1-D convolution



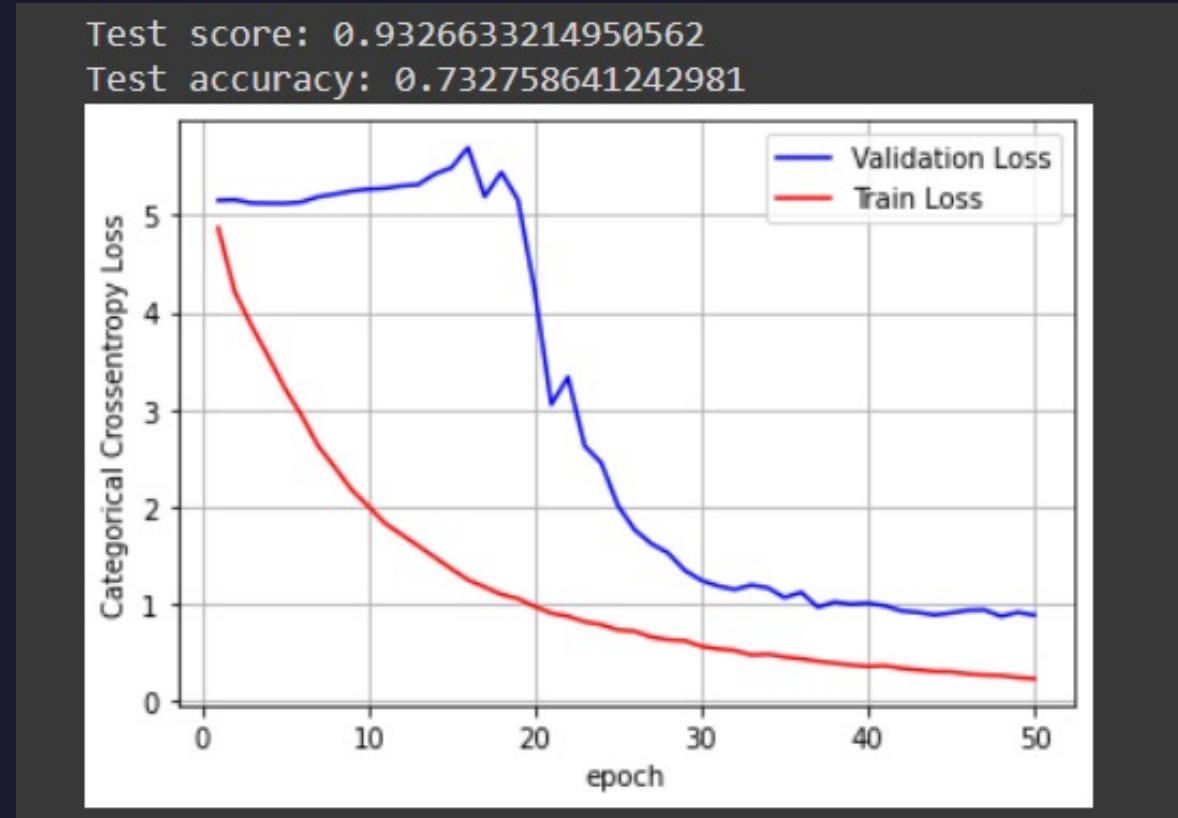
Old Model

Model without LSTM and normal 1-D convolution

With Data Augmentation



Without Data Augmentation



Performance Metrics

Model with LSTM and Dilated 1-D convolution



695/695 [=====] - 3s 5ms/step - loss: 0.3454 - accuracy: 0.9473				
Test loss: 0.3454049229621887				
Test accuracy: 0.9473400115966797				
+-----+-----+-----+-----+				
Model epochs test loss test accuracy				
+-----+-----+-----+-----+				
Deep CNN-LSTM 50 0.3454 0.9473				
+-----+-----+-----+-----+				
Classification Report				
	precision	recall	f1-score	support
5809	0.00	0.00	0.00	1
5810	0.00	0.00	0.00	1
5811	1.00	1.00	1.00	1
5812	0.00	0.00	0.00	1
5813	1.00	1.00	1.00	1
5814	1.00	1.00	1.00	1
5815	0.00	0.00	0.00	1
5816	1.00	1.00	1.00	1
5817	1.00	1.00	1.00	1
5818	0.00	0.00	0.00	1
5819	1.00	1.00	1.00	1
5820	0.00	0.00	0.00	1
5821	1.00	1.00	1.00	1
5822	1.00	1.00	1.00	1
5823	0.00	0.00	0.00	1
5824	1.00	1.00	1.00	1
5825	0.00	0.00	0.00	1
5826	0.00	0.00	0.00	1
5827	0.00	0.00	0.00	1
5828	1.00	1.00	1.00	1
5829	1.00	1.00	1.00	1
5830	0.00	0.00	0.00	1
5831	0.00	0.00	0.00	1
accuracy			0.95	22218
macro avg	0.92	0.91	0.91	22218
weighted avg	0.95	0.95	0.94	22218

Predictions

```
● def pred(test_seq):
    print("Input Test Sequence: ",test_seq)
    test_seq=[test_seq]
    df2=pd.DataFrame(test_seq)
    df2.columns=['sequence']
    test2 = (df2.sequence).apply(space_in_sequence)
    test2 = right_padding_with_index(test2[:], max_seq_length)
    ohe_test2 = one_hot(Test2[:])
    y_pred2 = model.predict(ohe_test2)
    c=np.argmax(y_pred2,axis=1)
    print()
    print("Predicted Protein Family for the inputted sequence: ",keys[c])
```

```
[158] pred("HNLQMRDSMNTYNNMVNRCAFATCIRSFQEKKVNAEEMDCTKRCVTKFVGYSQRVALRFAE")
```

Input Test Sequence: HNLQMRDSMNTYNNMVNRCAFATCIRSFQEKKVNAEEMDCTKRCVTKFVGYSQRVALRFAE
100% [██████] 1/1 [00:00:00:00, 25.50it/s]

```
Predicted Protein Family for the inputted sequence: ['PF02953.15']
```

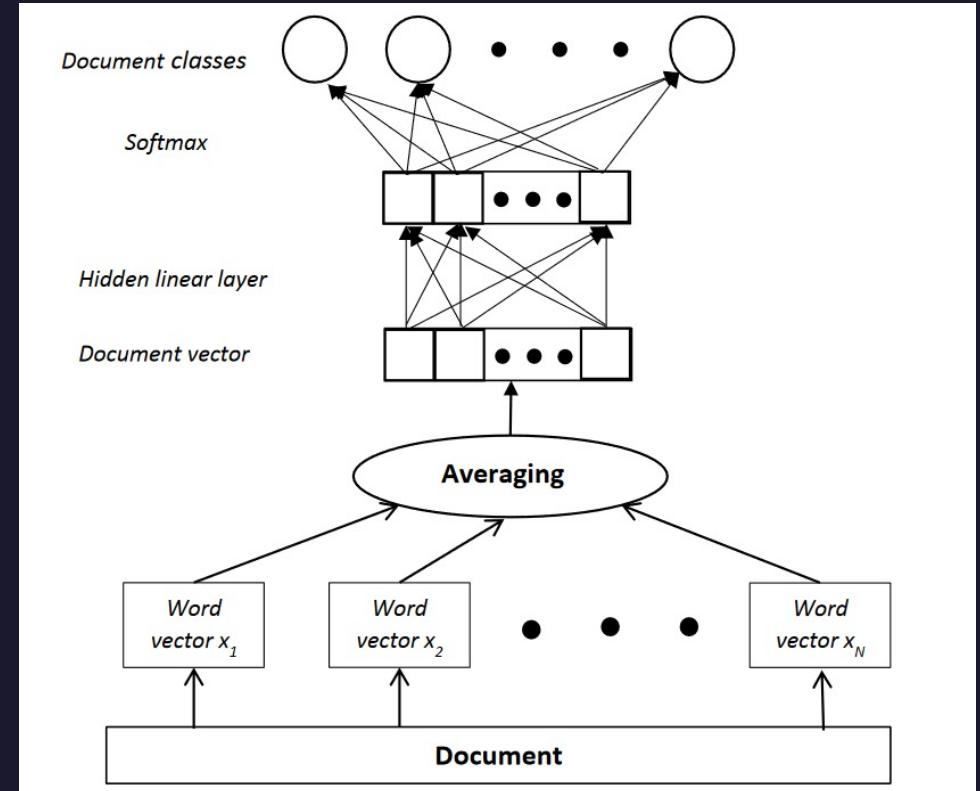
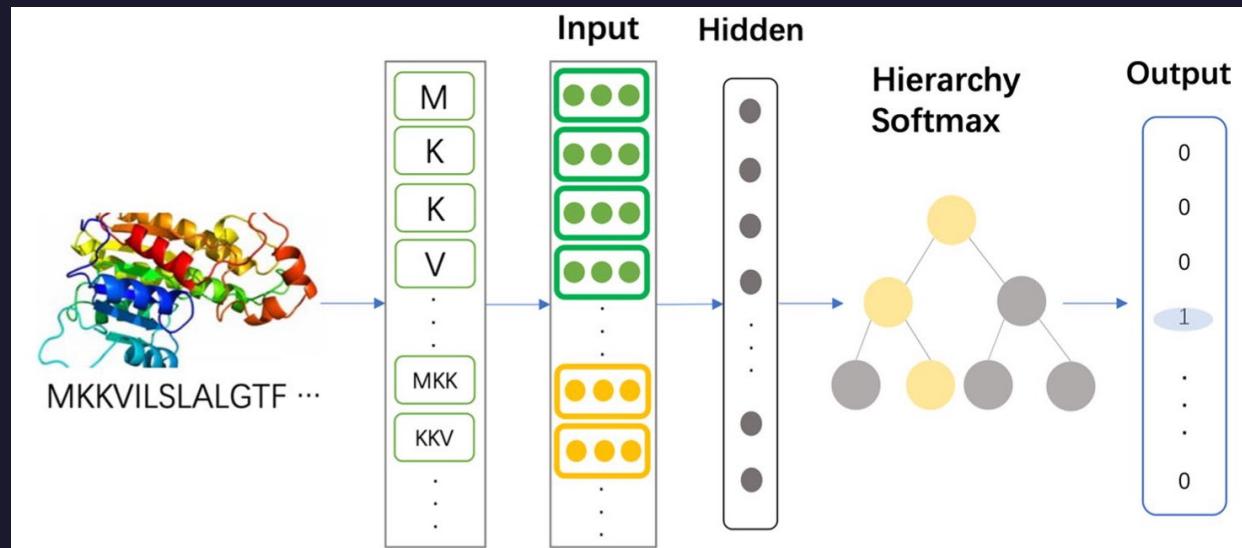
```
[145] def pred(test_seq):
    print("Input Test Sequence: ",test_seq)
    test_seq=[test_seq]
    df2=pd.DataFrame(test_seq)
    df2.columns=['sequence']
    test2 = (df2.sequence).apply(space_in_sequence)
    test2 = right_padding_with_index(test2[:], max_seq_length)
    ohe_test2 = one_hot(test2[:])
    y_pred2 = model.predict(ohe_test2)
    c=np.argmax(y_pred2,axis=1)
    print()
    print("Predicted Protein Family for the inputted sequence: ",keys[c])
```

```
[156] pred("GAVRVDVSGGLTDAMVSSYLNTDKSLTVTIVNADNQDRDISLAISSGGQPAGAVSVYETSAEHDLPVRNAGADGRLAVKKQSIVTI")
```

Input Test Sequence: GAVRVDVSGGLTDAMVSSYLNTDKSLTVTIVNADNQDRDISLAISSGGQPAGAVSVYETSAEHDLPVRNAGADGRLAVKKQSIVTI
100% [██████] 1/1 [00:00:00:00, 19.72it/s]

```
Predicted Protein Family for the inputted sequence: ['PF13620.6']
```

FastText – Architecture Diagram



FastText – Input Format



```
fast_train.txt
2 __label__PF1103.8 RDSIYYQIFKRPALIFELVDNRPPQAQNRFESVEVKETAFRIDGVFLPEDATSRVIFFAEVQFQKDEGLYHRRFTESLMLNRRNQSQYDDWYCVVIFPSRSLEPNKRTHRIFLNSDQVQRIYLDELGTSDTLPIGINLMQL
3 TTASSETMAEQAKOLIQRVKLEETGILPKTEIIIEIITTAIVKFSSLREEVEAML
4 __label__PF13005.7 TCCPDCGGELRLVGEDASEILDAMIAQMVKIEVARLKKSCRCE
5 __label__PF02261.16 MLRMMMSKIHARATVTEADLNVYGSITIDEDILDAVGMLPNEKVHVNNGNARFETYIIAGERGSGVICVNGAAARLVQRGDIVIIIISYVVDNAEAKDHKPTVAIMGEINTK
6 __label__PF00630.19 TACPKQCTARGLGLKAAPVTQPTRFVVLNDCHGQPLGRSEGELEVDIVTDDNRRNVNDVIVDRDGTYEVSYTPMQRGTVSINVVRMGEAIGGCDFV
7 DHEPVAWAAATAAGFVIAAVLRRSRAFPVLLTALLAGFATASLKAARIAHPVLAAPVFAELRGFVEIREERERTDRFLVRLVTQMQAARAPTLERVLRSVKGTAFTVGSFVTLKARLQPPLQPLRPGSYDFARDLYFQGIAASGFVLGAITT
8 __label__PF0094.12 LFVFGTROGESNHFLADSQCLGHFETPPHYALYDLGTYPAVIEGHTILGEVYILDETDLARVLDKLEDVPVERREQIETPFGEAWIYLQDGSMLDTIISSGDWQ
9 __label__PF0635.19 FTQDFDVFRIEGLDORMEQLKSTVRPKLEVGEYFSPVLSVTLEEIFYHVAKHARRTVPPKDTWAFSPGKRGYKMLPHFQIGLWEDELFITVAVNECPOKASIGKLEKKIDPILENIPGHYIWSDDHTKPGIRTDRM
NPHTLNFFRMQVVKAAEMLCGLEIPREEAAAGMSPDELAKTIEDFVHVLPY
10 __label__PF06580.13 SEIKLLHAQVNPFLNALNTLSAIVRDRPEKACHLVLNLSFFRKNLKRSEEALSEEVAAYLEIEQARFLDKL
11 __label__PF01379.20 LRIATRKSPLALWQAEYDASRLLRAAHPPDRLVELVGMTTRGKPLDPLAKVGGKGLFVKELEQGLLEGRADIVHSMDVPEFPEGHLAAILEREDPRDALVSHYRFAELPADARIGTSSLRQCQIKCRLPGCSLYDLR
GNVNTRLAKLDAFEGDAIVLASAGLKRLGFQERIAETLPEQCLPAIGOAGVCECRSTDTRNA
12 __label__PF01484.17 FAAFGCSAFAITACLVVISSAYMVVKEIGDEVLEIDIQMFRSEIDSASEI
13 __label__PF09148.10 KIKTNQVLEDGQLQTIEAFYRGERVDKNGDIFVSDFDKVKSITTTISGDIVTIKFGQVNTKMFKEGFEQRTPYRTPYGVFDMRLYTYKLSKKIDDRQIKMGLDYKMEVTDLMKANNRDI
14 __label__PF09351.10 LYHASIPMMVKYGLNLKVILIAKEHCACKNLNPEEMKFRLLIEDMRSLDYQVQSVSNTAKFVATRLAQDTYFPDTETFPPELQSRVDATISILSRIDPSSMDGKENADILMETKSMGIFRFTGYSYIVQYACPNFHFLSSAYCIFRHLGVPLTAFDYL
15 __label__PF01346.18 KTRLSYAAGSTLGQDITMIAERQEWGPVDKIALLAGVVDVSQHQLPQEQLTQLVAKADATANAARDKRHQASQSLQDDVYFLARFKKQKGATQSPSGFWYRI
16 __label__PF07862.11 MSLDQARSFLARMQDDOALRRAEVLAAATADDVQIAQIAGRGDFDSDEL
17 __label__PF100351.9 ILSGMMVYIYLSTYHERFWFSRPELTFVQGDSAIYFKELIKAPSFKRGYQLTHDNTLSMKTNMVVRQMTLYPELIAVLYQASGSEDVIEPVYFYIGIVFGLQGYTYVAALFVTSWLMGSTWLAGMLTVAFIINR
ADTTRIDYSTPLRNWAMPYFAQVVAALTGYLKNLNNSSAERFCYLLVSASTYTFFMMWEYSHYLLFIQIAISLFLDLLLTQTEKVHEVYKIYFLSFLGYMLQFNPALVSPLLSLLVVAALLAKYLOMNTKGTLLSMLKVVYFYLVFTMTVTLNLVAKT
FISHKEHEHLMKVLEVKGFLNTTKNFVWFLCQESFQTPSQDLFLRLTQSSLPFYILVLIICLVSQVQAIKFMRMGGQPVSETTLKVGRIGERPEVYVHLHSISLGLAVTMEGMKYVWTPVCMIAAFGVCSPELWTTLFKWLRLKIVHPVLLAVLSMA
VPTIIGFSLWKEFPPRIMGELTELQEFYDPPDTVGLM
18 __label__PF06160.12 IALVAILILSIVMMVMRKQTQIKLEQLEQRYNTLKGIPALFKLNKAVALSRVNEAMSTVVEECKTDDEVQEKLLKTCVSDAACDDLIYGHKVKARRRNELSDDLACETDVNKIHTVLDDILEQENEQRVHINALKETFR
KVKKTIHENRTATQSSEYELETEIJAIEKMSFKFEEWMFASEFNKAADQOKEIKEISITRNEIVEALPSLYERAKGILPRAIDEVGYNYARAKNGVLLHELEVSKNLDVISMOKNDLNRLHSGTPENVKEDLDDLEVRIAQLEQIRLEEAFDEVNDGLTA
LFDSIREVNCEFDQDIIKSLYARVFENWTQRLODTQTRLDVLNDMQRRLDKIVLWDQPYTTLIAYKELAQNAGFSKEVLMKEKLNACSDERAKQQLKQLIINEIRVKMLKHLRPNVSQAQYEDLCKGEDMMHDVRDILEHSPLDVTQLNAKLR
EAIDFIYLTLYNNVNLVGMAIMVENTVFGNRYRSSCEIDSELTRCELCFRDGQYTAKLKGIOCIEKMPHPGAYEKLN
19 __label__PF01934.17 FEHQALFDSQTDWQSEIGELALQRIGHTLIECILDTGNDMIDGFMIRDPGSYDDIMDILVDEKVTKEGDELKKLIAYRKLTVQQYLADSGELYRLLIAKHTALQDFPKRI
20 __label__PF08544.13 AYWLLOREFTPELYKLIRSKSODLDRGLVCPEDMRKVSKMRKLGIAEKITGAGCGGHLTVVKKGQOIPRGPVSVIDHOG
21 __label__PF07950.11 VFLAAGVGNVRVPLQVEGDSNIGLAFVAHGFARHPTLSSLAYAGLGVCGHMWGAALKWLYGAPTMAGWSGSGSKAIDKKTQRRPMWwGLHGATFCAFVwAAGGLGIVARGGL
22 __label__PF14333.6 REGLLHRLFYMLIAILLFASTVLTAITVQFIVMLVSKGEPENERLADFGTDLGIWMAKAARYQAAASEVKWPWPS
23 __label__PF02580.16 KFVIQRVNHASVCVDTGTVGAIEKGLLVLIGVGKQDTKEIADKYLQKLIGLRIFEDMEGKTNLSLKQVEGELLLISQFTLYANCKGNRPSFIEAGSPDMANELEYIISQAKKEIPVVTGIFGADMKVALENDGPFTIVLD
24 __label__PF06971.13 KIPRPTIKRLAIYVRCLEKQLLQEKNSSISSKEIGELLGKASQVRKDLS
25 __label__PF11221.8 SDILTQLQTCYDQLLQFFSTISYLSQRHPLVAPEDPDPNDFTPPPSNAVTHTQSQTQPGASQTASHQPQIQPGPDETDRAPYPLRPPVPPQVFANAQREADELVQKGQIELLSRRLPGIGASEQQAAEIRVLAEKVRDME
EKRAKAKRKEQYVRLRDGVILG
26 __label__PF00688.18 DNIDQSSFIQRRLSKSOERREMOIREILSILGLPQRPRPLLHERHTAAPMMLDLYNAILEDGRDRGLVSYSEPAYTPGPPLVTQQDSRFLSDADMMSFANTVDEPEEDLQLYHQRREFRFLSRIPPGETVTAAEFRYKDF
VRERYENETHVSVFQVLQDQHLSRVAEEGWLVDLTTVSNHVINPGQNLGLQLLVETKVGMSGNGPQYGSQVQFV
27 __label__PF02675.15 LDISGCDPEKIRSKKEIITOFADLCEYKIMKRFGDPIVVRFGADPKVQGYSQLAQLIETSMISGHFAEDTNKAFIDVFSCKEYPPKAACFKCMDFYFGTVEVDYSVIFR
28 __label__PF01384.20 GANSNSPPFAPAIGANAVSTMRAAFLIGLAAALGALTQGGSISETVGAGLIDGVAITSLAATTGLLTATAFMAGIYSGYPVPAAFATTGAMVGVLGSLGGPAVDTYRQIALFWVLPPVSGGLAYLTAKLLRHDDIPETGV
PLLAAVGGTVANIQLGVIPSPAGEQNSLAGFIADLVPTEAVGVELTVLVSITAAAVSFQTRLRIQESVETGIRTFLLVGSVVAFFSSGSQVGLATGPLENLYTAE
LGLPGIVLLILGATGILLAGAMGAPRLQJATLGVRSSIAQLAIGLGPISFNNIIISVGIGLGGAGGSAGVKRKGTVFLLTVTSVII
29 __label__PF00920.21 KPMIGVINTWTTVTPCNMHLDLAAPVRAEVAREAGGHPVDFTNTVSDGISMGTGMRASLISREVITDSTIELTRGHSLDGVVILVGCCKTIPAAAMALARMDVPGCILYGGTIMPGLGDQALSIQDVFEEVGAHAGTLDD
AGLDKVEKAACPGAGACGGQFTANTMAMILTMLGLSPMGVNDIPAPHDKPEAACRGRLAVERAKGSGTPRPRFITEASLRNAVVGASAGSGSTNAVLHVAIAAEAGIPFDIAEFDRISSIEAPVTDLKPGGRFLAHMFAGGSRLFGQRLIEGGLADPT
VSGKSLHEECAASEESLNQRVIOSVANPVPKPDGFRVLTGDAPEGAVLKLSGHARSEFSGPARFECEEDAFAAVEANSVKAGDIIIIRNEGPKGGPMREMGLVTAALVGQGLADVALITDGRFSGASKFGVIGHVSPREADGGPIGRVNGDRVRIDVA
RIDVADLSARPQSSGRAPTGVFAKYAAVUSSASRGA
30 __label__PF06252.12 VSNDYRAALESRFGVTTCKDLTQAOKSFIDELOELAKTDQERYSRERAARARAEEAGTPKRFDELDNRPGMASAQLRKIEAMWTDISDVPDPAAARALRFLRLLIAKVSDLRFLLDQMGWVINALNVMKH
31 __label__PF02397.16 KRTFDLIGSLLLLLSPLLLTLSSLAKLSSRGPFYRSTRPGIGGLPFDCLKFRMTDDAGVSDEEALNEADGALKIRDDPRITPVGFLRRFLSDELPLQNVNVRGEMSLVGPRLPLRDFEKELEWHKRYLVLPGT
GLWQVSGRSELDFDVLVRDFLYLERWVALDVLVILLKTPVAFVT
32 __label__PF12416.8 EFRSHKPEVLLMLAIKKSSLLHTKDFKHLTAFNSSEKESRTPPLQSPGHSITANMLQSQANVYQSLVQLGLLQVGNPLVDCDIIEVLLQFKQLKLNKFKVSLYQEKADEVVLLMFDVGNVTNIELKLNNTDAYTLDVGL
```

Line 2, Column 1

Tab Size: 4 Plain Text

FastText Model – Training Logs



```
(base) gokul@Gokuls-MacBook-Pro fastText-0.9.2 % ./fasttext supervised -input ..  
/fasttext_datasets/fast_sampled_train.txt -output ../fasttext_datasets/model2 -l  
r 0.5 -epoch 50 -wordNgrams 1 -bucket 200000 -dim 50 -loss hs  
Read 20M words  
Number of words: 460686  
Number of labels: 17180  
Progress: 0.1% words/sec/thread: 698614 lr: 0.499589 avg.loss: 1.855945 ETA  
Progress: 0.2% words/sec/thread: 698492 lr: 0.499164 avg.loss: 1.758754 ETA  
Progress: 0.3% words/sec/thread: 695538 lr: 0.498748 avg.loss: 1.701415 ETA  
Progress: 0.3% words/sec/thread: 693780 lr: 0.498329 avg.loss: 1.637730 ETA  
Progress: 0.4% words/sec/thread: 692566 lr: 0.497927 avg.loss: 1.634475 ETA  
Progress: 0.5% words/sec/thread: 692376 lr: 0.497504 avg.loss: 1.600382 ETA  
Progress: 0.6% words/sec/thread: 690900 lr: 0.497100 avg.loss: 1.591604 ETA  
Progress: 0.7% words/sec/thread: 690439 lr: 0.496693 avg.loss: 1.566882 ETA  
Progress: 0.7% words/sec/thread: 690478 lr: 0.496276 avg.loss: 1.575260 ETA  
Progress: 0.8% words/sec/thread: 690411 lr: 0.495874 avg.loss: 1.576262 ETA  
Progress: 0.9% words/sec/thread: 689048 lr: 0.495479 avg.loss: 1.572719 ETA  
Progress: 1.0% words/sec/thread: 688032 lr: 0.495081 avg.loss: 1.571430 ETA  
Progress: 1.1% words/sec/thread: 687721 lr: 0.494679 avg.loss: 1.566565 ETA  
Progress: 1.1% words/sec/thread: 686160 lr: 0.494292 avg.loss: 1.563820 ETA  
Progress: 1.2% words/sec/thread: 684910 lr: 0.493883 avg.loss: 1.576079 ETA  
Progress: 1.3% words/sec/thread: 684731 lr: 0.493467 avg.loss: 1.598632 ETA  
Progress: 1.4% words/sec/thread: 685098 lr: 0.493045 avg.loss: 1.589179 ETA  
Progress: 1.5% words/sec/thread: 685101 lr: 0.492630 avg.loss: 1.582792 ETA  
Progress: 1.6% words/sec/thread: 684781 lr: 0.492231 avg.loss: 1.571949 ETA  
Progress: 1.6% words/sec/thread: 684302 lr: 0.491817 avg.loss: 1.556768 ETA  
Progress: 1.7% words/sec/thread: 683972 lr: 0.491418 avg.loss: 1.550285 ETA  
Progress: 1.8% words/sec/thread: 682987 lr: 0.491029 avg.loss: 1.541101 ETA  
Progress: 1.9% words/sec/thread: 682768 lr: 0.490613 avg.loss: 1.539524 ETA  
Progress: 2.0% words/sec/thread: 682719 lr: 0.490205 avg.loss: 1.545949 ETA  
Progress: 2.0% words/sec/thread: 683120 lr: 0.489793 avg.loss: 1.551547 ETA  
Progress: 2.1% words/sec/thread: 682916 lr: 0.489379 avg.loss: 1.545877 ETA  
Progress: 2.2% words/sec/thread: 682882 lr: 0.488964 avg.loss: 1.521212 ETA  
Progress: 2.3% words/sec/thread: 682542 lr: 0.488553 avg.loss: 1.500235 ETA  
ress: 2.4% words/sec/thread: 682640 lr: 0.488139 avg.loss: 1.477337 ETAProg  
ress: 2.5% words/sec/thread: 682714 lr: 0.487724 avg.loss: 1.451820 ETAProg  
ress: 2.5% words/sec/thread: 682463 lr: 0.487327 avg.loss: 1.428536 ETAProg  
ress: 2.6% words/sec/thread: 682364 lr: 0.486931 avg.loss: 1.403524 ETAProg  
ress: 2.7% words/sec/thread: 681854 lr: 0.486525 avg.loss: 1.379286 ETAProg  
ress: 2.8% words/sec/thread: 681570 lr: 0.486121 avg.loss: 1.361208 ETAProg  
ress: 2.9% words/sec/thread: 681453 lr: 0.485725 avg.loss: 1.348291 ETAProg  
ress: 2.9% words/sec/thread: 681154 lr: 0.485334 avg.loss: 1.328238 ETAProg  
ress: 3.0% words/sec/thread: 680703 lr: 0.484941 avg.loss: 1.313698 ETAProg  
ress: 3.1% words/sec/thread: 680751 lr: 0.484524 avg.loss: 1.301641 ETAProg  
ress: 3.2% words/sec/thread: 680451 lr: 0.484115 avg.loss: 1.285479 ETAProg  
ress: 3.3% words/sec/thread: 680192 lr: 0.483712 avg.loss: 1.273163 ETAProg  
ress: 3.3% words/sec/thread: 679706 lr: 0.483319 avg.loss: 1.259390 ETAProg  
ress: 3.4% words/sec/thread: 679679 lr: 0.482909 avg.loss: 1.246104 ETAProg  
ress: 3.5% words/sec/thread: 679729 lr: 0.482498 avg.loss: 1.233005 ETAProg  
ress: 3.6% words/sec/thread: 679666 lr: 0.482092 avg.loss: 1.217732 ETAProg  
ress: 3.7% words/sec/thread: 679522 lr: 0.481680 avg.loss: 1.203961 ETAProg  
ress: 3.7% words/sec/thread: 679312 lr: 0.481277 avg.loss: 1.196166 ETAProg
```

```
ress: 97.5% words/sec/thread: 540287 lr: 0.012288 avg.loss: 0.616387 ETAProg  
ress: 97.6% words/sec/thread: 540243 lr: 0.012013 avg.loss: 0.616077 ETAProg  
ress: 97.7% words/sec/thread: 540202 lr: 0.011733 avg.loss: 0.615829 ETAProg  
ress: 97.7% words/sec/thread: 540166 lr: 0.011450 avg.loss: 0.615522 ETAProg  
ress: 97.8% words/sec/thread: 540123 lr: 0.011174 avg.loss: 0.615239 ETAProg  
ress: 97.8% words/sec/thread: 540076 lr: 0.010902 avg.loss: 0.614903 ETAProg  
ress: 97.9% words/sec/thread: 540033 lr: 0.010627 avg.loss: 0.614602 ETAProg  
ress: 97.9% words/sec/thread: 539987 lr: 0.010353 avg.loss: 0.614286 ETAProg  
ress: 98.0% words/sec/thread: 539952 lr: 0.010070 avg.loss: 0.613930 ETAProg  
ress: 98.0% words/sec/thread: 539914 lr: 0.009790 avg.loss: 0.613610 ETAProg  
ress: 98.1% words/sec/thread: 539879 lr: 0.009507 avg.loss: 0.613275 ETAProg  
ress: 98.2% words/sec/thread: 539842 lr: 0.009226 avg.loss: 0.612986 ETAProg  
ress: 98.2% words/sec/thread: 539804 lr: 0.008942 avg.loss: 0.612685 ETAProg  
ress: 98.3% words/sec/thread: 539765 lr: 0.008652 avg.loss: 0.612374 ETAProg  
ress: 98.3% words/sec/thread: 539724 lr: 0.008369 avg.loss: 0.612053 ETAProg  
ress: 98.4% words/sec/thread: 539687 lr: 0.008073 avg.loss: 0.611700 ETAProg  
ress: 98.4% words/sec/thread: 539643 lr: 0.007799 avg.loss: 0.611380 ETAProg  
ress: 98.5% words/sec/thread: 539601 lr: 0.007519 avg.loss: 0.611058 ETAProg  
ress: 98.6% words/sec/thread: 539564 lr: 0.007223 avg.loss: 0.610713 ETAProg  
ress: 98.6% words/sec/thread: 539520 lr: 0.006940 avg.loss: 0.610394 ETAProg  
ress: 98.7% words/sec/thread: 539481 lr: 0.006654 avg.loss: 0.610061 ETAProg  
ress: 98.7% words/sec/thread: 539442 lr: 0.006376 avg.loss: 0.609736 ETAProg  
ress: 98.8% words/sec/thread: 539402 lr: 0.006097 avg.loss: 0.609445 ETAProg  
ress: 98.8% words/sec/thread: 539363 lr: 0.005803 avg.loss: 0.609112 ETAProg  
ress: 98.9% words/sec/thread: 539327 lr: 0.005521 avg.loss: 0.608804 ETAProg  
ress: 99.0% words/sec/thread: 539279 lr: 0.005247 avg.loss: 0.608477 ETAProg  
ress: 99.0% words/sec/thread: 539242 lr: 0.004961 avg.loss: 0.608177 ETAProg  
ress: 99.1% words/sec/thread: 539210 lr: 0.004676 avg.loss: 0.607934 ETAProg  
ress: 99.1% words/sec/thread: 539174 lr: 0.004394 avg.loss: 0.607596 ETAProg  
ress: 99.2% words/sec/thread: 539132 lr: 0.004103 avg.loss: 0.607276 ETAProg  
ress: 99.2% words/sec/thread: 539093 lr: 0.003825 avg.loss: 0.606978 ETAProg  
ress: 99.3% words/sec/thread: 539051 lr: 0.003545 avg.loss: 0.606665 ETAProg  
ress: 99.3% words/sec/thread: 539013 lr: 0.003263 avg.loss: 0.606367 ETAProg  
ress: 99.4% words/sec/thread: 538973 lr: 0.002986 avg.loss: 0.606053 ETAProg  
ress: 99.5% words/sec/thread: 538934 lr: 0.002708 avg.loss: 0.605779 ETAProg  
ress: 99.5% words/sec/thread: 538901 lr: 0.002424 avg.loss: 0.605444 ETAProg  
ress: 99.6% words/sec/thread: 538866 lr: 0.002142 avg.loss: 0.605126 ETAProg  
ress: 99.6% words/sec/thread: 538826 lr: 0.001865 avg.loss: 0.604832 ETAProg  
ress: 99.7% words/sec/thread: 538788 lr: 0.001586 avg.loss: 0.604549 ETAProg  
ress: 99.7% words/sec/thread: 538752 lr: 0.001302 avg.loss: 0.604256 ETAProg  
ress: 99.8% words/sec/thread: 538718 lr: 0.001020 avg.loss: 0.604055 ETAProg  
ress: 99.9% words/sec/thread: 538688 lr: 0.000734 avg.loss: 0.603803 ETAProg  
ress: 99.9% words/sec/thread: 538652 lr: 0.000454 avg.loss: 0.603496 ETAProg  
ress: 100.0% words/sec/thread: 538616 lr: 0.000173 avg.loss: 0.604090 ETAProg  
ress: 100.0% words/sec/thread: 538447 lr: 0.000000 avg.loss: 0.605840 ETAProg  
ress: 100.0% words/sec/thread: 538446 lr: 0.000000 avg.loss: 0.605840 ETA:  
(base) gokul@Gokuls-MacBook-Pro fastText-0.9.2 % ./fasttext test ../fasttext_dat  
assets/model2.bin ../fasttext_datasets/fast_sampled_train.txt  
N 6872001  
P@1 0.908  
R@1 0.908
```

FastText Model

Hyper-Parameters

- FastText Embedding Model = 50
- Loss = Hierarchical Loss
- Bucket Size = 200,000
- Learning Rate = 0.5
- Epochs = 50
- Word N-Grams = 1

Predictions

```
N      6872001
P@1    0.908
R@1    0.908
(base) gokul@Gokuls-MacBook-Pro fastText-0.9.2 % ./fasttext predict-prob ../fasttext_datasets/model2.bin - -1 0.5
STDRVFAAFTGASLVGVARCTRFPEGSLVDGVYVLEEYRGFAQRIMRRLIEECGRDGALYLYAKPEHLDFYREMGFEP
__label__PF11992.8 1.00014
VKYQNLVSQLNPHFLFNSLATLESLIYSDRHLAVKFLSQLTRVYRYVLTRNAKLVSLGEELSFIKDYTDLLQTRFGKGL
__label__PF05315.11 1.00014
KLAGFVQIDDAYLGGERNGGKAGRGSENKQSFLIAVQTDDTFTAPRFVVIEPVRSFDNPSLQDWIARRLAGCEVYTDGL
ACFRRLEDAGHAHTTLDTSGGRAATEATGARWVNVLGNLKRAISGVYHAIAQGKYAKRYLAEAAYRFN
__label__PF12762.7 1.00015
PEGYLCHRCHVGGFIQHCPT
__label__PF13696.6 1.00015
FKVILYGSSIYVVGHVLLSLGAVPFLSPIRSSLDFSGLFVIAFATGCIKPCVSAFAADQFTEDQKDLSQFFSFFYFAI
NGGSLFAIIITPILRGRVQCFGNAHCFPLAFGVPGVLMLLALILFLMGWSMYKKHPPSKENVGSKVVAIVYTSLRKMVG
ASRDKPVTHWLDHAAPEHSQKMIDSTRGLNVAVIFCPLIFFWALFDQQGSTWVLQARRLDGRVGHFSILPEQIHAINPV
CVLILVPIFEGWVYPALRKITRVTPLRKMAVGGLLTAFSFAIAGVLQLKVNETMEFPPSLGRIYLQRVGNESLISDFRYK
SDGRLIGDGMLPKGRTELADAGIYTFTNTGLKNESQEIDISTPNKGYVMAVFRL
__label__PF00854.21 1.00014
^C
(base) gokul@Gokuls-MacBook-Pro fastText-0.9.2 %
```



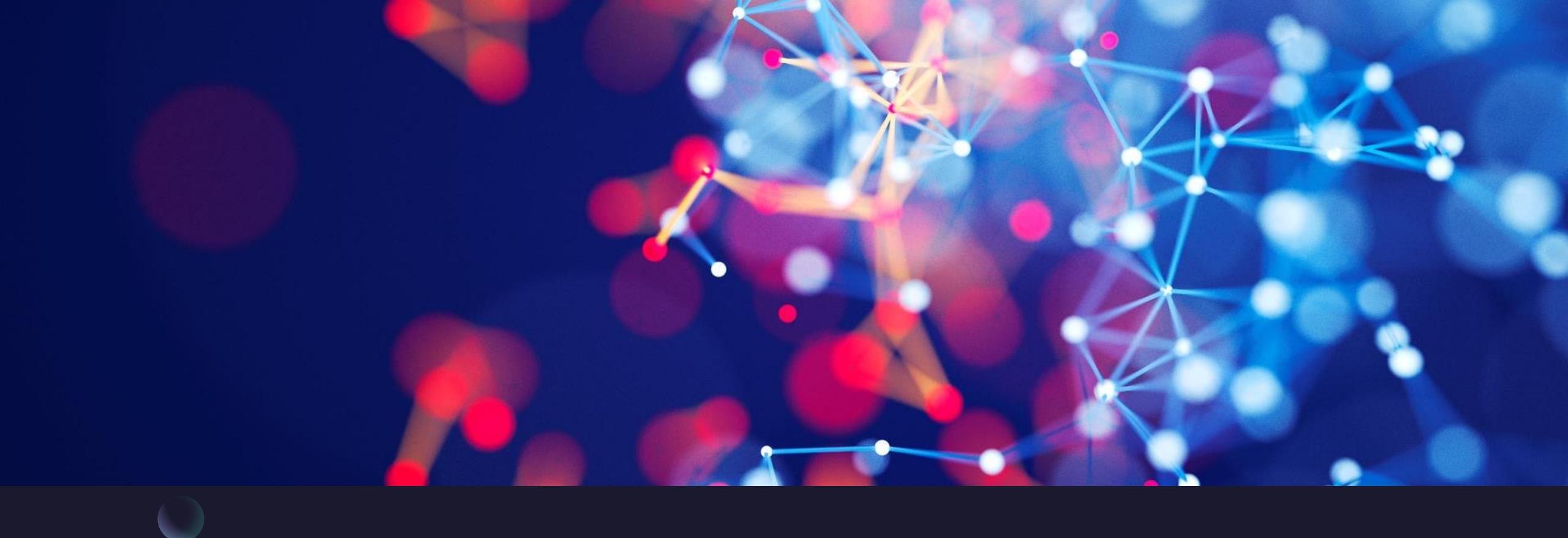
Inferences

CNN-LSTM Model

- Using a Dilated CNN instead of a normal one proves to be more efficient as the field of reception is spread. This results in a more wider field of view with the same number of kernel parameters and same computational complexity.
- In order to combat the problem of a highly imbalanced dataset, we first over-sample the minority classes using RandomUpsampler and then under-sample the majority classes with an appropriate sampling rate using a RandomUndersampler.
- Using an LSTM layer post 4 sequences of batch convolutions helped the model learn a contextual representation of the amino acid sequence, which is extremely important when it comes to sequence classification problems.
- This improved our test accuracy from 0.8936 to 0.9473 (5.37% improvement).
- Usage of batch normalization sped up training by normalization the hidden layer activation and handling internal covariate shift.

FastText Model

- Under-sampling the majority classes to 400 samples and up-sampling the minority classes to the same size has helped the FastText model to learn a good representation of the PFAM protein database by achieving a P@1 and R@1 score of 0.908.

A complex network of interconnected nodes and edges, primarily in shades of blue, red, and yellow, against a dark background. A single teal sphere is visible on the left side.

Summary

Identification of a family of a protein sequence is crucial as it give scientists and researchers a reference point, to which based on the family they can attribute it to known properties. This can prove very useful when identifying the sequence of an unknown disease, to identify its family and to create a potential vaccine based on it's family properties.