Università degli Studi di Padova

Kolmogorov-Arnold Networks application to NLP

A New Machine-Learning Algorithm

M. Roccatello L. Milan A. Ozen, D. Marzola, M. Gugole

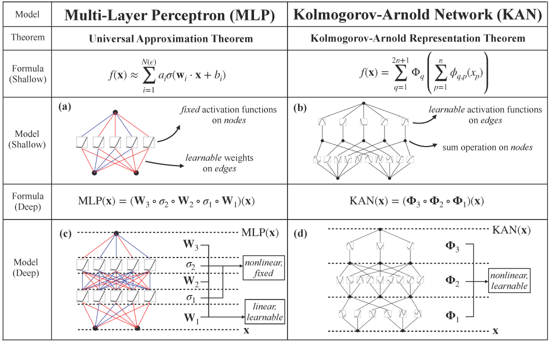
June 2, 2024

Introduction

- **Objective of our work:** Kolmogorov-Arnold Networks are a brand-new Neural Network model, not yet evaluated on NLP related tasks. Our aim is to experiment with KANs on Text Classification and Sentyment Analysis, via the analysis of two datasets

Kolmogorov-Arnold Networks

- Kolmogorov-Arnold Networks (KANs) as an alternative to Multi-Layer Perceptron Networks (MLPs)
- Key Difference: KANs use learnable activation functions on edges (weights), unlike fixed activation functions in MLPs.



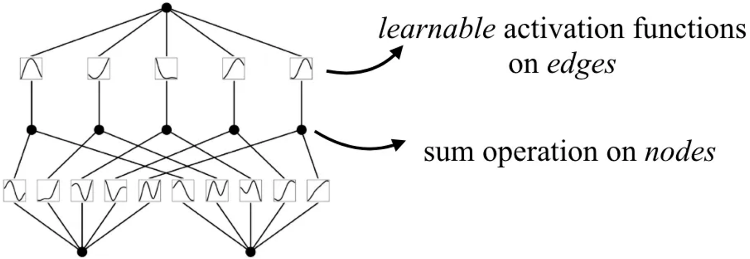
K.A. Representation Theorem

- Any multivariate continuous function on a bounded domain can be written as a finite composition of continuous univariate functions and addition

$$f(x) = f(x_1, \ldots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \phi_{q,p}(x_p) \right)$$

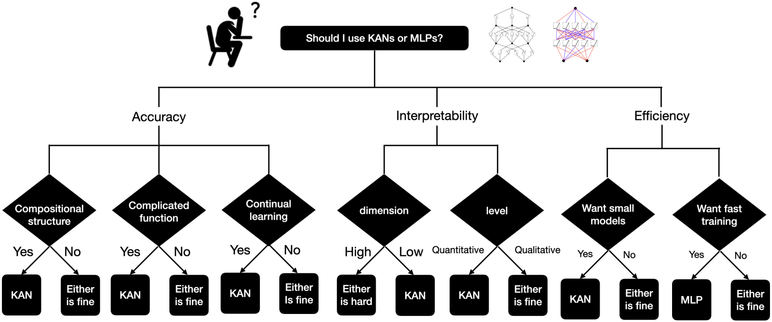
- Idea: break down a complicated task into smaller easier-parts, focus on each part individually and then put everything together to solve the bigger problem.

Università
degli Studi
di Padova



Training KANs

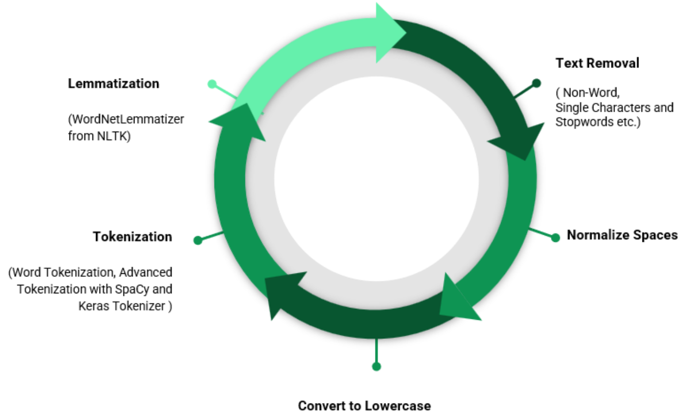
- Differentiable operations allow for backpropagation
- Techniques for stability and convergence:
  - Regularization: dropout, weight decay
  - Optimization: algorithms, learning rates
  - Normalization: batch, layer normalization
- Slow training speed (10 times slower than MLPs)



KANs for NLP Tasks

- Language modeling, semantic analysis, text generation
- Applications: sentiment analysis, language translation, document summarization

Introduction to text preprocessing



Text Cleaning: Both notebooks perform essential text cleaning operations such as removing special characters, converting text to lowercase, and eliminating extra spaces by using Regular Expressions.

Tokenization: Both use tokenization as a fundamental step to split text into individual words.

SpaCy Tokenization with Movies Dataset

SpaCy's more sophisticated tokenization, which includes sentence segmentation and detailed token-level processing.

Keras Tokenizer with Clinical Dataset

The Keras Tokenizer method involves fitting the tokenizer on the text data to build a vocabulary of unique tokens. It then converts the text into sequences of integers based on this vocabulary.

Removing Stop Words: Both approaches involve removing common stop words to reduce noise and focus on meaningful words.

NLTK provides a comprehensive list of block words for various languages. We used English in our work. This list is organized to include the most common words that are generally uninformative in most contexts.

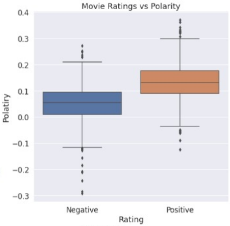
Lemmatization : The work related to Movies Dataset uses lemmatization to group different inflected forms of a word into a single item using WordNetLemmatizer from NLTK.

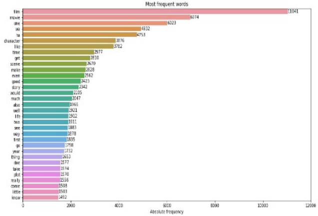
Università degli Studi di Padova

Exploratory Data Analysis (EDA)

- **01 Common Techniques**: Removing NaN Values, Removing Duplicates, Unique Values and Label Encoding.
- **02 Polarity Calculation**: TextBlob used for sentiment analysis (scores range from -1 to 1).
- **03 Visualize Frequent Non-Stop Words**: Visualize the text by removing stop words, tokenizing the words, creating corpus and counting word frequencies.
- **04 N-gram Analysis**: Identify common word combinations (bigrams, trigrams) using CountVectorizer.

Related Parts with Movies Dataset





Feature Extraction

TF-IDF Vectorization

Measure word importance in documents relative to the corpus.

It transforms text data into numerical features, where each word has a weight representing its importance.The shapes of the resulting matrices are number of documents by number of features.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Dimensionality Reduction with SVD

Truncated Singular Value Decomposition (SVD) used to reduce feature space dimensions, enhancing model performance and reducing complexity.

Embedding : In the Clinical Dataset, the Embedding method used is a trainable embedding layer provided by the Keras library.

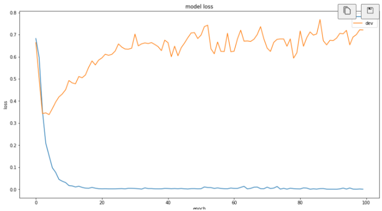
This is not a pre-trained embedding like Word2Vec or GloVe but rather an embedding layer that learns the representations during the training process of our specific model by converting words into dense vector representations.

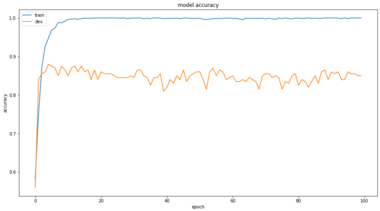
IMDB dataset

- **The dataset:** collection of movies reviews with a score representing the judgement (0: negative, 1 positive)
- **E.D.A. and data preparation**: data cleaning, tokenization, lemmatization, study of correlation between sentences polarity and labels. Vectorization of text into numerical vector. SVD dimensionality reduction.

IMDB dataset: MLP results

- **The model:** MLP with 500 input neurons, 256 and 192 neurons in the hidden layers, 2 output layers.

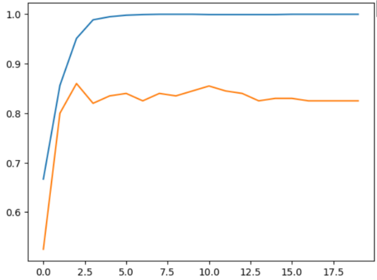


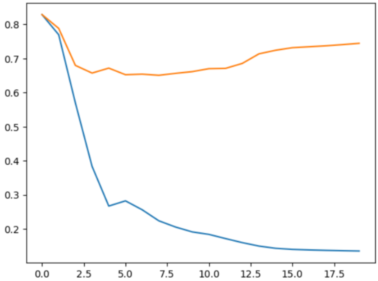


IMDB dataset: KAN results

- Grid search to find the best parameters

```
model = KAN(width=[500, n, 2], grid = grid, k = 3, seed
= 2024)
results = model.train(dataset, opt="LBFGS", steps=20, metr
=(train_acc, val_acc), loss_fn=torch.nn.CrossEntropyLoss(),
lamb = lamb)
```



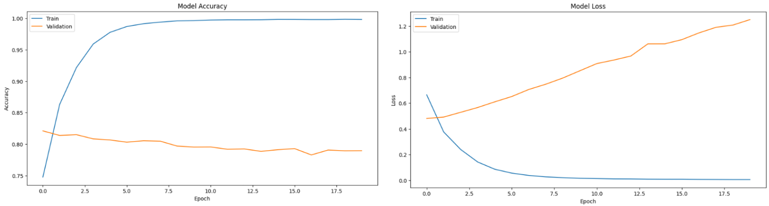


ClinicalReviews dataset

- **The dataset:** collection of reviews in a clinical to address to the correct environment
- **EDA and data preparation**: data cleaning, tokenization, lemmatization, study of distribution of labels and reducing the 5 most present label to spees up the exeution time. Vectorization of text into numerical vector.

ClinicalReviews: MLP

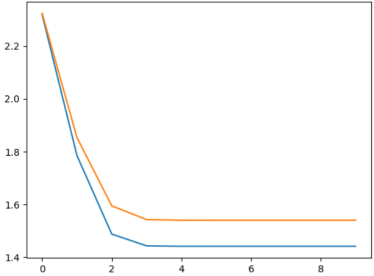
- **The model:** An embedding layer to 50 features, an hidden layer of 21 neurons and 5 neurons in the output layer

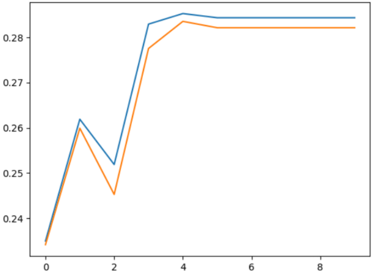


ClinicalReviews: KAN

Università degli Studi di Padova

```
model = KAN(width=[500, 10, 5], grid = 3, k=3, seed
= 2024)
results = model.train(dataset, opt = "LBFGS",
steps = 10, metrics =(train_acc, val_acc),
loss_fn=torch.nn.CrossEntropyLoss(), lamb = 0.01)
```





	train set	test set	train set	test set
accuracy	1.0	0.86	1.0	0.81
precision	1.0	0.88	1.0	0.75
recall	1.0	0.87	1.0	0.88
f1 score	1.0	0.87	1.0	0.81
AUC	1.0	0.94	1.0	0.82

	train set	test set	train set	test set
accuracy	1.0	0.78	0.28	0.28
precision	1.0	0.77	0.26	0.28
recall	1.0	0.77	0.24	0.24
f1 score	1.0	0.77	0.19	0.19
AUC	0.94	0.98	0.50	0.50

Table: Comparison between MLP (left) and KAN (right): results obtained on the IMDB (top) and Clinical Reviews (bottom) datasets.

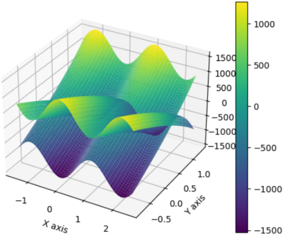
Università degli Studi di Padova

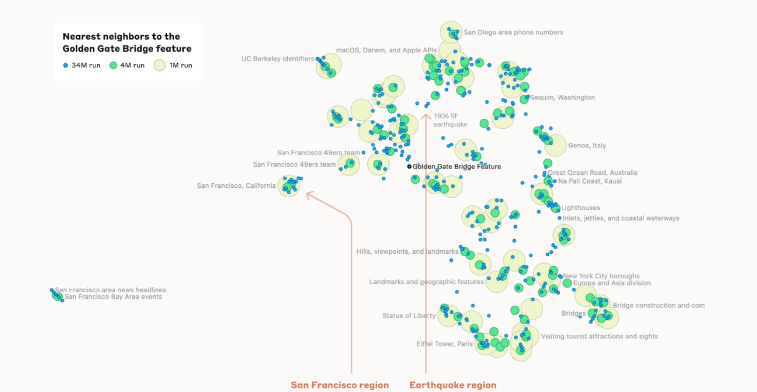
- **An example of the output**:

```python
formula1, formula2 = model.symbolic_formula()[0]
formula1
```

Python

$$-3425.46\sqrt{1-0.49x_2} - 450.6\sin(2.96x_1 + 7.83) + 3197.56$$





Feature Map of Claude 3 Sonnet by Anthropic.

NLP-KAN Team

Università degli Studi di Padova

Mattia Gugole

Mattia Roccatello

Lisa Milan

Aysenur Ozen

Devis Marzola