Earlier we talked about NLP wrt to machine learning which was probability based approach. Now we will be learning about NLP using deep learning based approach

NLP

ML.DL pipeline

1. Data ingestion
2. EDA
3. Pre-Processing
4. Models
5. Evaluation

ML/DL → CV → image | video → collection of frames (FPS)

ML/DL → NLP → Natural Language Processing → text (sequence) → (audio)

In NLP we are going to deal with the dataset having sequence within it like text, audio, timeseries based dataset etc

Agenda:

1. What is NLP ?
2. Natural Language
3. NLP usecase | NLP tasks
4. NLP application | Domains.
5. Approaches to NLP → ~~RB~~, ML, DL
6. Challanges in NLP
7. NLP Process Pipeline
8. GenAI → LLM

Primary, LLM

OpenAI, Nvidia

usecase → application → full deployment

Practical

RNN, LSTM, GRU, Encoder, decoder

Attention, Transformer

Bert, GPT, ...

These all architectures we are going to cover under this NLP

text Processing, word embedding

What is a NLP ?

decoding this defination

Natural language processing (NLP) is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate human language. It involves processing natural language datasets, such as text corpora or speech corpora, using either rule-based or probabilistic (i.e. statistical and, most recently, neural network-based) machine learning approaches. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

Natural + Language + Processing

language is the medium of communication

medium of communication

NLP → Computer Science

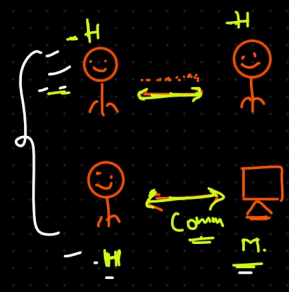
(text) (Corpora, Corpus)

Natural language communication b/w humans and machines was not possible 10-15 years back

10-15

Primary goal

So Machine can understand Human Language



The way 2 humans communicate with each other using natural language, in the same way main AIM OF NLP is to establish same intelligence when communicating with the machines

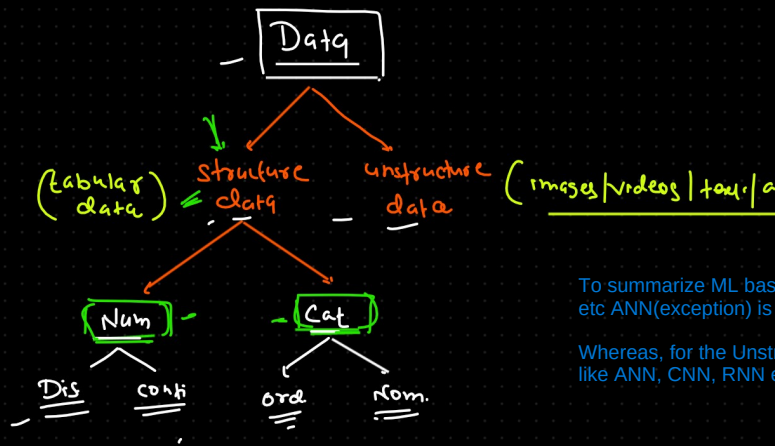
ATM

ATM example where we can talk to ATM machines for cash withdrawn using natural language

Data Science

ML / DL

Regression, classification



To summarize ML based approached liked Linear regression, SVM etc ANN(exception) is used on top of Structured data.

Whereas, for the Unstructured data we use DL based approached like ANN, CNN, RNN etc

Regression | classification

ML ⇒ Lin reg, ridge, Lasso, EN
SVM
DT
Log reg
Ensemble methods

N | C | N | C ← Domain | industry

↳ tabular data. Prediction

ANN

{Image | video} ⇒ CNN ⇒ ANN ⇒ image classification

DL ⇒ CNN ⇒ {Convolution, Neural Net}

Object detection
Object seg.
Object tracking
OCR
Usecase

RNN
CNN

⇒ Application ⇒ industry | Dep.

ML, ANN, CNN

NLP ⇒ field ⇒ text, Audio ⇒ Process ⇒ Sequence

Inp

Seq mapping.

"my name is Sunny and I am a DS
one Monday I am working on Genji
Project and after that I will explore
the RL."

Process flow ⇒

Machine

Approaches to NLP:

One important property of text/audio based data is the Sequence. In above example if we ask our model that who is data scientist then it won't be able to answer if it is ML, ANN or CNN based learning model because there is no way these models can retain the original sequence without loosing the contextual understanding of the information based. That why RNN were introduced where through sequence mapping machines are made capable of retaining the contextual understanding of the data

RNN ⇒ Stacked RNN, Deep RNN,
Bi-directional RNN
LSTM, GRU

NLP

Task | Usecase ⟶ Application ⟶ Industry | Domain

Different use cases of the NLP:

1. Text Summarization
2. Text Classification / Document Classification {Language Detection (language detection)} {Sentiment classification (sentimental classification)}
3. Text generation {Machine translation} (Sentence pred, Word pred)
4. Information / keyword extractor {POS (Part of speech tagging)} >>> POS tagging
5. Knowledge graph
6. topic modelling (Based on important words present in the sentence or paragraphs we are basically going to give a topic to that sentence)
7. Question Answering
   {Question → Answer}
8. Speech recognization
9. Chatbot {Conversational ai}

Both Question Answering and Chatbot(Conversational AI) may seem like similar use case but they are not. Question Answering is done on only some specific topic whereas, Chatbot does conversation over wide variety of topics

10. NER (Named Entity Recognition which is performed using BERT)

11. Sentence similarity

12. Fill in the blanks / Spell corrector

Application

NLP powered applications:

1. Spam filtering ⇒ mail → spam, ham
2. next word pred. ⇒ gmail (auto completion)
3. Inshorts (text summarization) ⇒ 30 words
4. Grammarly (spell corrector)
5. Paraphrasing (quillbot) — Quillbot
6. Google translator (text generation)
7. Google assistant (speech to text)
8. tweet classification / Sentiment Analysis
9. Contextual Adv. ⇒ Insta., Fb, twitter
   (text, audio) → Proces → Context → model

Contextual Advertisement: When we are searching for something over google then after then Insta, Fb, Twitter starts showing the advertisements similar to the searched keyword.

Puma Shoes
Nike shoes

Insta
Fb

10. Chatgpt (Chatbot)
    (Conversational ai)

11. Search engine (Sentence matching, Sentence retrieval, Sentence classification)
    (NLP)

12. Alexa (speech to speech)
    text
    NLP

NLP Pipeline

1. Data ingestion
2. text Preprocessing
   - Basic cleaning
   - Basic to advance Processing

(Optional)
(EDA ⇒ text)

3. Feature Eng
   - Manual approach (In machine learning)
   - DL approach (Automatic feature extraction)
4. Model building
5. Model evaluation
6. Deployment

1. Data ingestion

Sentiment Analysis

- xlxs
- csv

Sql | Nosql

3rd party Server

huge

Pipeline

ETL

kafka → Spark → mysql

If data is present with the 3rd party vendor like Azure block etc then data engineers help would be needed who will basically build a data pipeline using kafka and spark and can store the data in the sql database

API

Web Scrapping [flipkart review]

Image ⇒ text

Document ⇒ word, pdf

OCR

2. EDA ⇒ Charts

3. Text Preprocessing

Clean

4. Feature eng

NLP :- text ⇒ Numbers

In this step we are basically going to convert the text into embeddings(numbers)

1. manual (Machine learning based approaches)
   - BOW (Bag of word)
   - OHE (One hot encoding)
   - TFIDF (TFIDF)
2. DL ⇒ NN (Automatic features extraction using neural network based DL approach)
   - word2vec
   - Elmo
   - fasttext

5. Model building

1. Rule based approach (1950-1960)
   (Heuristic approach)
2. ML
   - Naive bayes (Cond. Poob) (Conditional Probability)

NLP ⇒ Seq mapping

Bigger Sentance ⇒ Fail

These two approaches were not able to map the sequences effectively. Meaning that it was getting failed for bigger sentence by not able to retain the contextual understanding effectively.

3 DL ⇒ RNN | LSTM, GRU

LLM ⇒ Base Archi. → {transformer}

Encoder, Decoder
Attention
apir
Transformer
chatgpt
↳ 2017

5 Model evaluation

1. Confusion mat (Acc, F-1score, Pre, Recall)
2. BLEU
3. Perplexity (gpt)
↳ Prob

Challanges

1. Ambiguity

I saw the man with the telescope.

This sentence can be interpreted in 2 ways:
1. The person saw the man who had a telescope.
2. The person used the telescope to see a man.

This sentence is ambiguous since there is no clear relationship b/w man and telescope

2. Contextual words

I ran to the store because we ran out of milk

3. Slang → Cherry on the cake is a slang that means perfectly designed or eloquent. But model can see it as we are placing a cherry on the cake

4. Synonyms → When using synonyms also models can mislead.

5. Sarcasm.