CSE472:DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING

L:2 T:0 P:2 Credits:3

Course Outcomes: Through this course students should be able to

CO1 :: Understand the origin, evolution, and linguistic foundations of Natural Language Processing

CO2 :: Understand word embeddings, semantic relationships, and embedding visualization techniques

CO3 :: Apply deep learning sequence models for NLP to text and sentiment classification tasks

CO4 :: Apply seq2seq models with attention for translation and summarization tasks

CO5 :: Use transformer architectures and pretrained language models for fine-tuned NLP tasks

CO6 :: Apply generative NLP and large language models for text generation and evaluation tasks, considering explainability and hallucination

Unit I

Foundations of NLP : origin of NLP, language and grammar, linguistic essentials, morphology, syntax, semantics, challenges of NLP, applications of NLP

Text Processing : tokenization, stemming, lemmatization, stop-word removal, punctuation handling, handling out-of-vocabulary words, normalization, Bag-of-Words, n-grams, TF-IDF

Unit II

Word Embeddings and Vector Representations : vector space models, dense word embeddings, Word2Vec, CBOW, Skip-Gram models, GloVe embeddings, capturing semantic similarity, analogy relationships, visualizing embedding spaces using PCA or t-SNE

Unit III

Deep Learning Sequence Models for NLP : sequential text data, recurrent neural networks, long short term memory networks, gated recurrent units, bidirectional RNNs, sequence modeling applications, sentiment classification, text classification, sequence training techniques, teacher forcing, truncated backpropagation through time, evaluation metrics for sequence tasks

Unit IV

Sequence-to-Sequence Models and Attention Mechanisms : encoder–decoder architectures for NLP, sequence-to-sequence models for machine translation and summarization, attention in deep NLP, soft attention, alignment mechanisms, Bahdanau and Luong attention, integrating attention into encoder–decoder networks, evaluation techniques, BLEU and ROUGE scores, limitations of classical seq2seq models

Unit V

Transformers : transformer architecture, self-attention, multi-head attention, positional encoding, transformer encoder and decoder blocks, tokenization methods, Byte-Pair Encoding and WordPiece

Pretrained Language Models : pretrained transformer models, BERT, GPT, T5, masked language modeling, next sentence prediction, causal language modeling, transfer learning for NLP tasks, fine-tuning for text classification, named entity recognition, question answering, HuggingFace Transformers

Unit VI

Generative NLP and LLMs : generative NLP models, text generation strategies, greedy search, beam search, top-k, nucleus sampling, instruction-tuned, large language models, model behaviors in summarization, dialogue generation, reasoning tasks, evaluation metrics, perplexity and human judgment measures, explainability and hallucination in LLMs

List of Practicals / Experiments:

List of Practicals

- Given a set of text documents, implement a preprocessing pipeline that performs tokenization, lowercasing, noise removal, stopword removal, stemming/lemmatization, and converts the processed text into numerical representations using Bag-of-Words and TF-IDF
- Load pretrained Word2Vec or GloVe embeddings, select a set of words, compute their embeddings, and visualize relationships in 2-D space using PCA or t-SNE

- Train CBOW and Skip-Gram Word2Vec models on a sample text corpus, compute word similarities and analogies, and compare the performance of both models
- Load pretrained GloVe or Word2Vec embeddings, select word groups belonging to different semantic categories, reduce dimensionality using PCA or t-SNE, and visualize the embeddings to interpret clusters and relationships
- Build a sentiment analysis model using an LSTM network on a dataset such as the IMDB movie reviews or SMS Spam dataset. Compare its performance with a baseline model that uses Bag-of-Words or TF-IDF features
- Implement a Bidirectional LSTM model for Named Entity Recognition using the CoNLL-2003 dataset or a smaller annotated dataset. Train the model to label words as Person, Location, Organization, or Other
- Implement a sequence-to-sequence model with Bahdanau or Luong attention to translate short English sentences into another language using a small parallel dataset. Compare translation quality with and without attention
- Build an attention-based sequence-to-sequence model to generate short summaries from news headlines or paragraph datasets. Train briefly using a subset of the dataset and evaluate generated summaries using ROUGE scores
- Fine-tune a pretrained BERT model for sentiment classification on a dataset such as the IMDB movie reviews or the SMS Spam dataset. Evaluate the model using accuracy and F1-score, and analyze how BERT's contextual embeddings improve classification performance
- Use a pretrained GPT-2 model to generate text based on a given prompt. Apply different multiple decoding strategies and compare the fluency, creativity, and coherence of the generated outputs

Text Books:

1. GETTING STARTED WITH DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING by SUNIL PATEL, BPB Publication, New Delhi

References:

1. SPEECH AND LANGUAGE PROCESSING: AN INTRODUCTION TO NATURAL LANGUAGE PROCESSING, COMPUTATIONAL LINGUISTICS AND SPEECH RECOGNITION by DANIEL JURAFSKY, Pearson Education India