Introduction to (Large) Language Models with Transformers

Selim Yaman

Machine Learning: Applications in Social Science Research
ICPSR 2023

July 5, 2023

Overview of Natural Language Processing (NLP)

- Widely used in machine translation, sentiment analysis, chatbots etc.
- In social sciences, we mostly use them for text classification, topic modeling, and sentiment analysis
- Three generations of NLP:
  - Keyword search
  - Bag of words (`stm`, `lda` etc)
  - Word-order aware ML models
- State-of-the-art models being developed and released rapidly.

Introduction to Transformers

- Transformers revolutionized NLP by "Attention is All You Need." (Vaswani 2017) (80k citations and counting)
- They effectively capture long-range dependencies in text
- Significant improvement from LSTM
- Architecture: an encoder and decoder
- Attention mechanisms weigh the significance of different input parts.

Introduction to Transfer Learning

- Universities/companies train large models from scratch
- We then use these pre-trained models to fine-tune for a different but related task: transfer learning
- Key advantage: Leverages existing neural networks to reduce the need for large amounts of training data
- Can significantly reduce computational resources and time

Why Transfer Learning in Social Science?

- Enables application of state-of-the-art models in NLP tasks, even with limited domain-specific data.
- Accelerates the process of model training and evaluation
- Facilitates understanding of complex patterns in text data, relevant to many social science questions
- Enhances the efficiency and effectiveness of ML models for social science research

How to Use Language Models

- Choose a relevant pre-trained model (e.g., BERT, RoBERTa, DistilBERT etc.)
- Explore models on Hugging Face Hub $>Link<$
- Fine-tune the model with your specific dataset
- Test the model and evaluate its performance
- Adjust model parameters for optimal results, if necessary

Utilizing Pre-Trained Models via APIs

- Zero-shot classification
- APIs (Application Programming Interfaces) provide a straightforward way to access pre-trained models.
- Examples: OpenAI's `GPT-4` or `ChatGPT` APIs.
- They bypass the need for model training and maintenance, and can handle requests in real-time.
- Caveat: API usage often comes with costs and usage restrictions
  - Chatgpt: \$0.002 per 1k tokens
  - GPT-4: \$0.06 per 1k tokens
- An interesting application example: Out of One, Many: Using Language Models to Simulate Human Samples

Limitations of Transfer Learning

- Still requires some computational resources and time, although less than training from scratch
- Data privacy issues can arise when using pre-trained models via APIs
- Potential bias in pre-trained models if they were trained on unrepresentative data