
Exploring transfer learning
with transformers.

Agenda

- Overview of Transfer Learning.
- Why Transfer Learning is becoming an active research area, especially in NLP.
- Transformers!
- A unified approach to transfer learning in NLP

Transfer Learning

- Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task.
- Training a model on a data-rich task and fine-tuning it on a related downstream task

Overview of Transfer Learning

- Traditionally, in transfer learning, pre-training is done using supervised learning on large labeled datasets.
- In Modern techniques, pre-training is often done via unsupervised learning on the unlabeled datasets.

Transformers

- The Transformer in NLP is a novel architecture that aims to solve sequence-to-sequence tasks while handling long-range dependencies with ease.
- Earlier, Recurrent Neural Networks were leveraged in transfer learning for NLP.
- Example: Translation, Summarization.

Why Transformers for Transfer Learning

- Due to en masses availability of unlabeled text data (Thanks to the Internet), transfer learning in NLP has become an active research area.
- Major NLP tasks like question answering, machine translation, summarization can be treated as a related task.

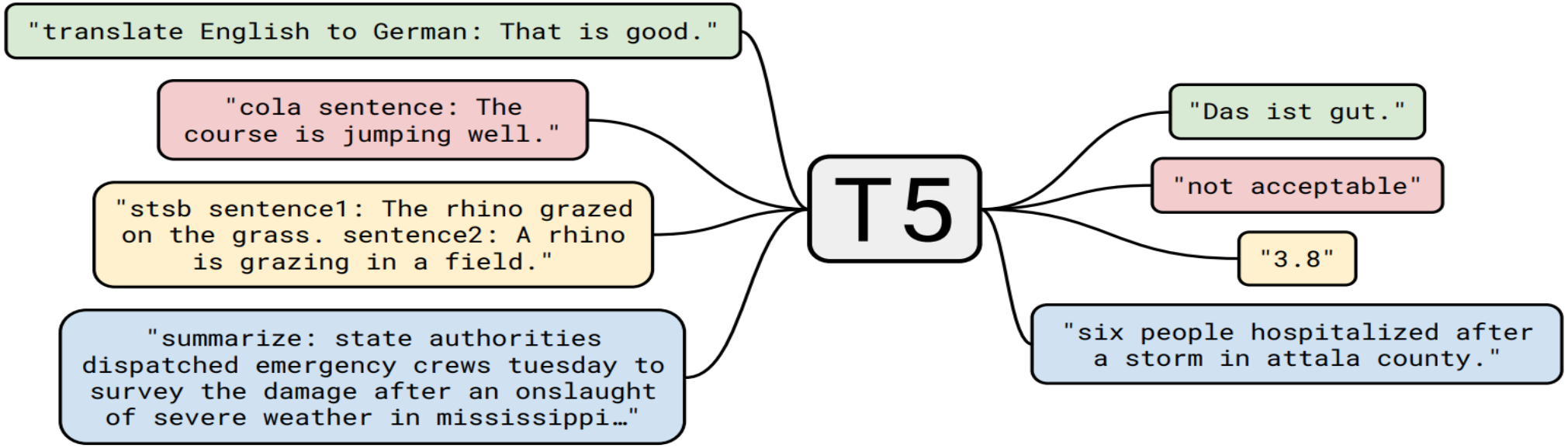
Unified Approach

- **Key – For related NLP tasks, treat every text-based/ text processing task as a “text-to-text” task.**
- **Using a text-to-text framework, one can apply the same model, objective, training procedure, and decoding process to every task one considers.**

Input Output Format

- To specify which task the model should perform, a task-specific (text) prefix is added to the original input sequence before feeding it to the model.
- For instance, to ask the model to translate the sentence “That is good.” from English to German, the model would be fed the sequence “translate English to German: That is good.” and would be trained to output “Das ist gut.”

T5 model - Text-to-Text Transfer Transformer



Source - https://arxiv.org/pdf/1910.10683v3.pdf

Dataset

- Colossal Clean Crawled Corpus (C4)
- Publicly-available web archive provides “web extracted text” by removing markup and other non-text content from the scraped HTML files.

Model Structures

- A major distinguishing factor for different architectures is the “mask” used by different attention mechanisms in the model.
- The self-attention operation in a Transformer takes a sequence as input and outputs a new sequence of the same length.

Model Structure

- **Fully Visible**
- **Casual**
- **Casual with prefix**

Model Structure

- An encoder-decoder Transformer, consists of two-layer stacks:
- The encoder, which is fed an input sequence, and the decoder, which produces a new output sequence.
- The encoder uses a “fully-visible” attention mask.
- This form of masking is appropriate when attending over a “prefix”, i.e. some context provided to the model that is later used when making predictions.

Model Structure

- The self-attention operations in the Transformer's decoder use a "causal" masking pattern.
- This is used during training so that the model can't "see into the future" as it produces its output.

Thank You