

26.10.22

Application of Transformer Models

GENERAL INTRODUCTION

- **Introduction Round**
- **Organizational Matters**
- **Structure of the Course**
- **The Role of NLP in AI**
- **Basic Introduction to Transformers**
- **How to Stay Up-To-Date**

INTRODUCTION ROUND

CHAT

Channels

wise22/23

Find channel

Threads

CHANNELS

00 - Announcements

01 - Questions

02 - Off-topic

Application of Transform...

Einführung in Data Scien...

Machine Learning with T...

Kursleitungen

DIRECT MESSAGES

Application of Transformer Models

21

Wednesdays, 10-11.45 am: Zoom; Course Handbook

Tip: Try **Ctrl F** to search this channel

Beginning of Application of Transformer Models

This is the start of the Application of Transformer Models channel, created by Ronja Tornow on August 31, 2022. Any member can join and read this channel.

Add members to this channel

Create a board

Set a Header

August 31

System

19:34

COURSE HANDBOOK



opencampus.sh Machine
Learning Program

EDU-Platform

Chat

Search...

opencampus.sh Machine Learning
Program

Course Kick-Off

How do I choose a course?

FAQ

COURSES

Einführung in Data Science und
maschinelles Lernen

Machine Learning with
TensorFlow

Deep Learning from Scratch

Deep Learning for Computer
Vision

Application of Transformer
Models

Requirements for a Certificate of
Achievement or ECTS

Application of Transformer Models

Here are the articles in this section:

Requirements for a Certificate of Achievement or ...

Week 1 - General Introduction

Week 2 - Self-Attention and Prompt Design

Week 3 - Introduction to Transformer Models

Week 4 - Fine-Tuning Pretrained Models

Week 5 - The Datasets Library

Week 6 - The Tokenizers Library

Week 7 - Main NLP Tasks

Week 8 - Joint Coding

Week 9 - Presentation of the Final Projects

← Previous
Week 11 - Final Presentation of the Projects


Next
Requirements for a Certificate of Achieveme... →

ORGANIZATIONAL MATTERS

- **Use your full names in the zoom meetings!**
- **Scan the QR-Code if you participate in presence.**
- **Complete your profile in the Mattermost chat with your full name and a photo.**
- **Please write me if you will not go on with the course!**


26. 10. 10: 00- 12: 00	GENERAL INTRODUCTION Remote via Zoom
02. 11. 10: 00- 12: 00	SELF-ATTENTION AND PROMPT DESIGN Remote via Zoom
09. 11. 10: 00- 12: 00	INTRODUCTION TO TRANSFORMERS Remote via Zoom
16. 11. 10: 00- 12: 00	FINE-TUNING PRETRAINED MODELS Remote via Zoom
23. 11. 10: 00- 12: 00	THE DATASETS LIBRARY Remote via Zoom
30. 11. 10: 00- 12: 00	THE TOKENIZERS LIBRARY Remote via Zoom
04. 01. 10: 00- 12: 00	TRAINING MODELS FROM SCRATCH Remote via Zoom
11. 01. 10: 00- 12: 00	PRESENTATION OF THE FINAL PROJECTS Remote via Zoom

HUGGING FACE TRANSFORMERS COURSE

 **Hugging Face**

Search models, datasets, users...

Models Datasets Spaces Docs Solutions Pricing



Course

Search documentation Ctrl+K

MAIN EN 400

0. SETUP

1. TRANSFORMER MODELS

Introduction

Natural Language Processing

Transformers, what can they do?

How do Transformers work?

Encoder models

Decoder models

Sequence-to-sequence models

Bias and limitations

Summary

End-of-chapter quiz

2. USING TRANSFORMERS


3. FINE-TUNING A PRETRAINED MODEL

4. SHARING MODELS AND TOKENIZERS

5. THE DATASETS LIBRARY

Introduction

Welcome to the Course!



This course will teach you about natural language processing (NLP) using libraries from the [Hugging Face](#) ecosystem — [Transformers](#), [Datasets](#), [Tokenizers](#), and [Accelerate](#) — as well as the [Hugging Face Hub](#). It's completely free and without ads.

Introduction

Welcome to the Course!

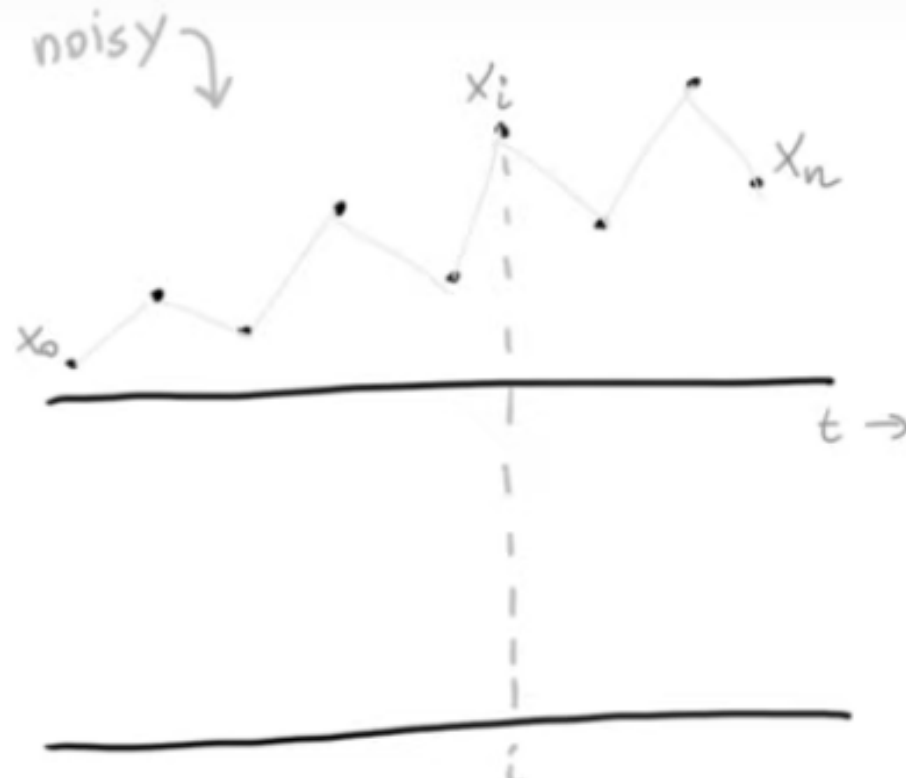
What to expect?

Who are we?

RASA ALGORITHM WHITEBOARD

☰ Rasa Algorithm Whiteboard - Transformers & Attention 1: Self Attention

ATTENTION MECHANISMS



noisy

timeseries data $x = [x_0 \dots x_n]$

re-weighting w_i

1:41 / 14:31

BREAKOUT

- **Which AI impressed you the most in the last few month?**
- **Do you know any transformer models?**

APPLICATION EXAMPLES

- **Whisper by OpenAI**
- **Stable Diffusion (Image Generation)**
- **Jukebox (Music Generation)**
- **Alpha Tensor**

APPLICATION EXAMPLES

- **Code Generation: Co-Pilot in VS-Code from GitHub and OpenAI (GPT model)**
- **Search: Google (Bert model)**
- **Prediction of Protein folding: Alpha Fold from Deepmind**
- **Image Generation: DALL-E 2 by OpenAI**

AGI RESEARCH



ANTHROPIC

THE TRADITIONAL ROLE OF NLP IN AI

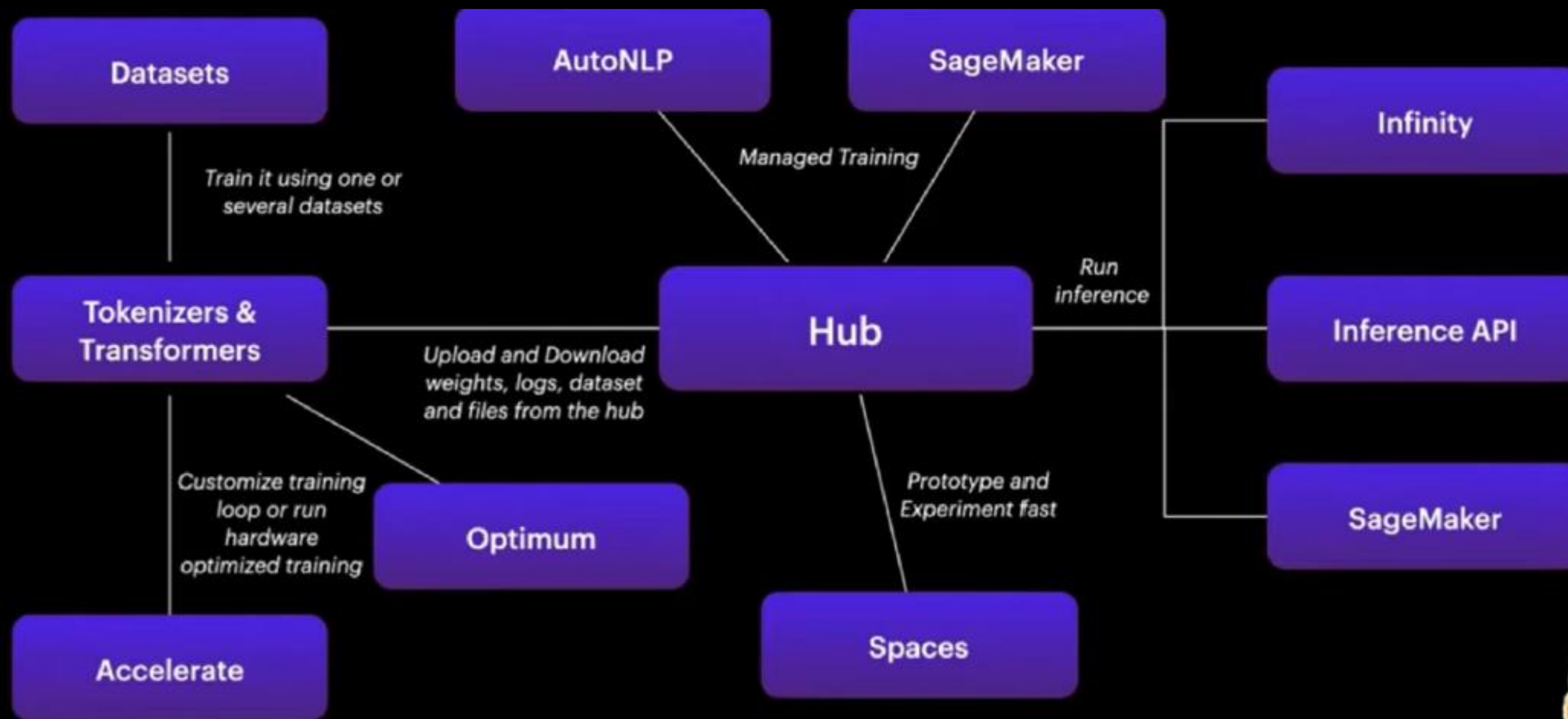
- **A subfield of linguistics, computer science, and artificial intelligence**
- **Dealing with the interaction between computers and human language**
- **The goal is a computer capable of "understanding" text contents, including contextual nuances**
- **The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.**

THE ROLE OF NLP FOR ARTIFICIAL GENERAL INTELLIGENCE (AGI)

Touring Test:

- **Test for intelligence in a computer, requiring that a human being should be unable to distinguish the machine from another human being by using the replies to questions put to both.**

HUGGING FACE ECO SYSTEM



WHAT IS A TRANSFORMER?

Main ingredients



Attention
mechanisms



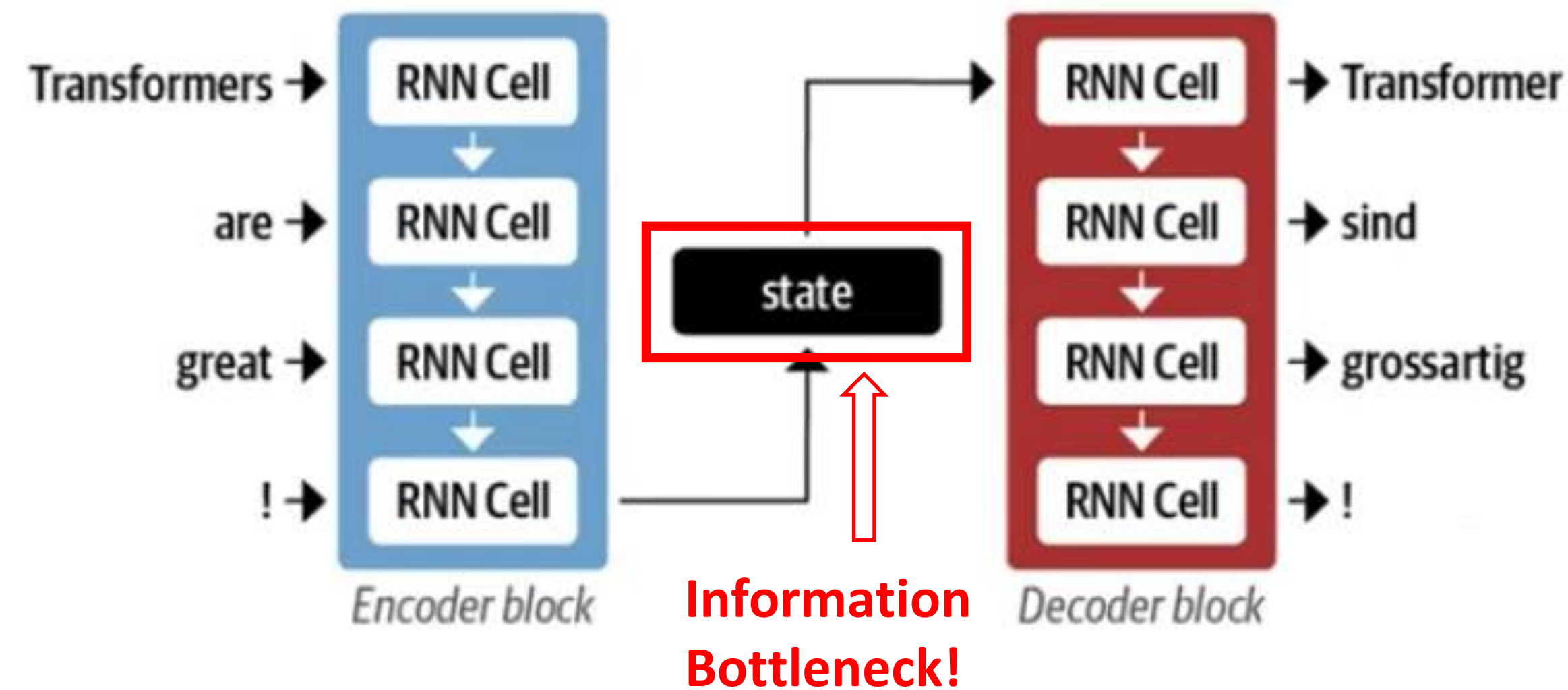
Self-supervised learning
(Pretraining)



Transfer learning
(Fine-tuning)



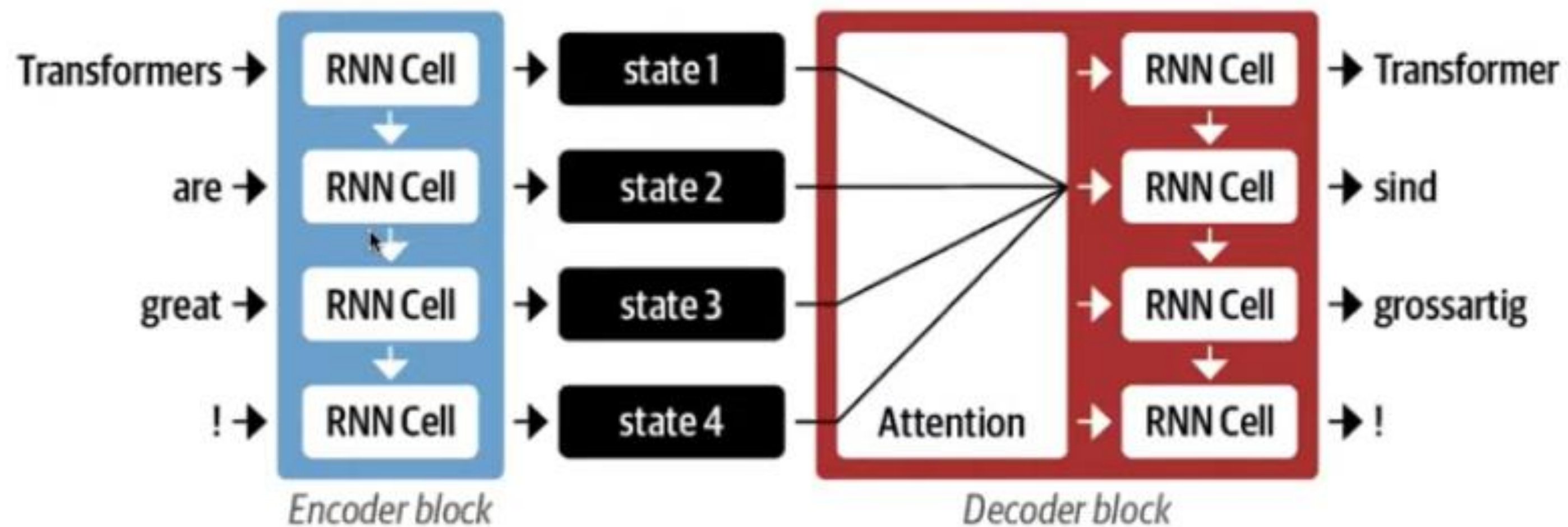
CLASSICAL SEQUENCE TO SEQUENCE APPROACH



Originally developed for recurrent neural networks



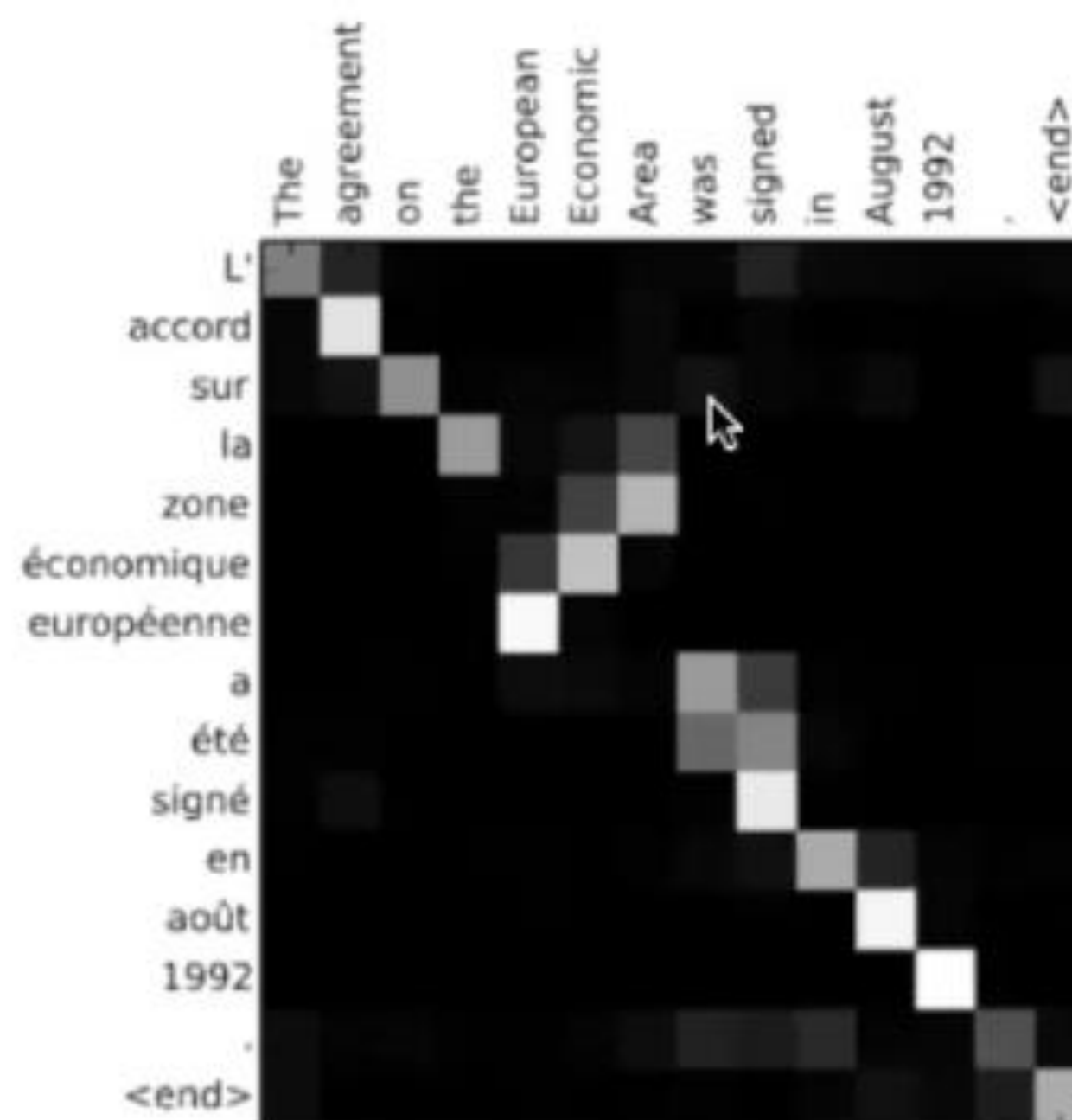
ATTENTION MECHANISM



Assign a weight or "pay attention" to specific states



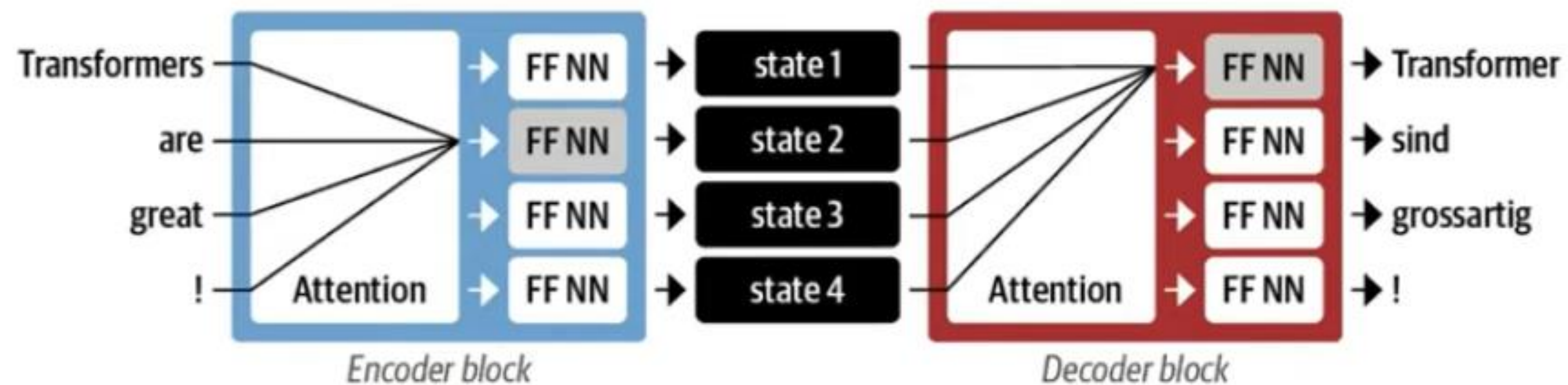
INTERPRETATION



Attention gives better modelling of word order



ATTENTION IS ALL YOU NEED



Transformers much easier to scale with compute & data



BENEFITS

- **Solving the bottleneck problem in sequence-to-sequence tasks**
- **Provides some interpretability**
- **No Vanishing Gradient Problem**
- **Multimodality**

HUGGING FACE MODEL HUB

- <https://huggingface.co/models>
- **Check out models for one or two tasks, you are not that familiar with yet.**

STAYING UP-TO-DATE

STAYING UP-TO-DATE

- **Newsletter**
 - **Medium**
 - **Reddit**
 - **The Batch**
 - **...**
- **Twitter**
 - **Thomas Wolf**
 - **Nvidia**
 - **...**
- **LinkedIn**
 - **Steve Nouri**
 - **Hugging Face**
 - **...**
- **ArXive / ArXive Sanity**

TODOS FOR NEXT WEEK

- Watch [video 1](#) (Self-Attention) and [video 2](#) (Keys, Values, Queries) of the Rasa Series on Transformers & Attention.
- Note at least one question on the each of the videos!
- Do [chapter 1](#) of the Hugging Face course.



Huggingface Datasets

- <https://huggingface.co/datasets>

Hugging Face

Search models, datasets, users...

Models

Datasets

Pricing

Resources

Log In

Sign Up

Task Category

conditional-text-generation text-classification

structure-prediction sequence-modeling

question-answering text-scoring + 3

Task

machine-translation language-modeling

named-entity-recognition sentiment-classification

dialogue-modeling extractive-qa + 128

Language

en es fr de ru ar + 104

Multilinguality

monolingual multilingual translation

other language-learner

Size

10K<=n<100K 1K<=n<10K n<1K 100K<=n<1M

n>1M 1k<10K + 18

License

mit cc-by-4.0 cc-by-sa-4.0 cc-by-sa-3.0

apache-2.0 cc-by-nc-4.0 + 56

Datasets 638

Search Datasets

Sort: Alphabetical

acronym_identification

Acronym identification training and development sets for the acronym identification task at SOU@AAAI-21.

annotations_creators: expert-generated language_creators: found languages: en licenses: mit multilinguality: monolingual size_categories: 10K<=n<100K source_datasets: original task_categories: structure-prediction task_ids: structure-prediction-other-acronym-identification

ade_corpus_v2

ADE-Corpus-V2 Dataset: Adverse Drug Reaction Data. This is a dataset for Classification if a sentence is ADE-related (True) or not (False) and Relation Extraction between Adverse Drug Event and Drug. DRUG-AE.rel provides relations between drugs and adverse effects. DRUG-DOSE.rel provides relations between drugs and dosages. ADE-NEG.txt pro...

annotations_creators: expert-generated language_creators: found languages: en licenses: unknown multilinguality: monolingual size_categories: 10K<=n<100K size_categories: 1K<=n<10K size_categories: n<1K source_datasets: original task_categories: text-classification task_categories: structure-prediction task_categories: structure-prediction task_ids: fact-checking task_ids: conference-resolution task_ids: conference-resolution

adversarial_qa

AdversarialQA is a Reading Comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles using an adversarial model-in-the-loop. We use three different models; BIOAD (Seo et al., 2016), BERT-Large (Devlin et al., 2018), and RoBERTa-Large (Liu et al., 2019) in the annotation loop and construct three datasets...

annotations_creators: crowdsourced language_creators: found languages: en licenses: cc-by-sa-4.0 multilinguality: monolingual size_categories: 10K<=n<100K source_datasets: original task_categories: question-answering task_ids: extractive-qa task_ids: open-domain-qa



Paperswithcode Datasets

- <https://www.paperswithcode.com/datasets?mod=texts&page=1>



835 dataset results for **Texts** x



Penn Treebank

The English Penn Treebank corpus, and in particular the section of the corpus corresponding to the articles of Wall Street Journal (WSJ), is one of the most known and used corpus for L...
1,545 PAPERS • 10 BENCHMARKS



SQuAD (Stanford Question Answering Dataset)

The Stanford Question Answering Dataset (SQuAD) is a collection of question-answer pairs derived from Wikipedia articles. In SQuAD, the correct answers of questions can be any se-...
1,254 PAPERS • 7 BENCHMARKS



Visual Genome

Visual Genome contains Visual Question Answering data in a multi-choice setting. It consists of 101,174 images from MSCOCO with 1.7 million QA pairs, 17 questions per image on aver-...
903 PAPERS • 11 BENCHMARKS



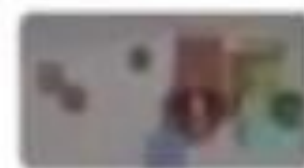
GLUE (General Language Understanding Evaluation benchmark)

General Language Understanding Evaluation (GLUE) benchmark is a collection of nine natural language understanding tasks, including single-sentence tasks CoLA and SST-2, similarity...
847 PAPERS • 14 BENCHMARKS



SNLI (Stanford Natural Language Inference)

The SNLI dataset (Stanford Natural Language Inference) consists of 570k sentence-pairs manually labeled as entailment, contradiction, and neutral. Premises are image captions fro...
743 PAPERS • 1 BENCHMARK



CLEVR (Compositional Language and Elementary Visual Reasoning)

CLEVR (Compositional Language and Elementary Visual Reasoning) is a synthetic Visual Question Answering dataset. It contains images of 3D-rendered objects; each image comes...
528 PAPERS • 1 BENCHMARK



Visual Question Answering (VQA)

Visual Question Answering (VQA) is a dataset containing open-ended questions about im-ages. These questions require an understanding of vision, language and commonsense...
435 PAPERS • 2 BENCHMARKS



Billion Word Benchmark

The One Billion Word dataset is a dataset for language modeling. The training/hold-out data was produced from the WMT 2011 News Crawl data using a combination of Bash shell and...
417 PAPERS • 1 BENCHMARK

Stanford

Linguistic Data Consortium



- <https://catalog.ldc.upenn.edu/>
- Stanford licenses data; you can get access by signing up at: <https://linguistics.stanford.edu/resources/resources-corpora>
- Treebanks, named entities, coreference data, lots of clean newswire text, lots of speech with transcription, parallel MT data, etc.
 - Look at their catalog
 - Don't use for non-Stanford purposes!

The screenshot shows the LDC website interface. On the left is a navigation menu with links: ABOUT, MEMBERS, COMMUNICATIONS, LANGUAGE RESOURCES (selected), Data, Obtaining Data, Catalog, By Year, Top Ten Corpora, Projects, Search, Memberships, Data Scholarships, Tools, and Papers. The main content area is titled 'Top Ten LDC Corpora' and lists the following:

Corpus ID	Corpus Name
LDC93S1	TIMIT Acoustic-Phonetic Continuous Speech Corpus
LDC2013T19	OntoNotes Release 5.0
LDC2006T13	Web 1T 5-gram Version 1
LDC96L14	CELEX2
LDC99T42	Treebank-3
LDC2008T19	The New York Times Annotated Corpus
LDC93S10	TIDIGITS
LDC97S62	Switchboard-1 Release 2
LDC2011T07	English Gigaword Fifth Edition
LDC93T3A	TIPSTER Complete

The Stanford logo is visible in the bottom right corner of the screenshot.

Dependency parsing: Universal Dependencies



- <https://universaldependencies.org>

Universal Dependencies

Universal Dependencies (UD) is a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing more than 100 treebanks in over 70 languages.

- [Short introduction to UD](#)
- [UD annotation guidelines](#)
- More information on UD:
 - [How to contribute to UD](#)
 - [Tools for working with UD](#)
 - [Discussion on UD](#)
 - [UD-related events](#)
- Query UD treebanks online:
 - [SETS treebank search](#) maintained by the University of Turku
 - [PML Tree Query](#) maintained by the Charles University in Prague
 - [Kontext](#) maintained by the Charles University in Prague
 - [Grew-match](#) maintained by Inria in Nancy
- [Download UD treebanks](#)

If you want to receive news about Universal Dependencies, you can subscribe to the [UD mailing list](#). If you want to discuss individual annotation questions, use the [Github issue tracker](#).

Stanford