

10.11.21

Natural Language Processing with Transformers

INTRODUCTION TO TRANSFORMER MODELS

- **Fragen zu Prompt Design**
- **Quiz**
- **Breakout Diskussion**
- **Projektideen**

QUIZ



<https://forms.office.com/r/DdRmqN9YgD>












BREAKOUT DISKUSSION

Angenommen, Du willst eine Textklassifizierung durchführen:

- **Ist es besser, ein Encoder- oder ein Decoder-Modell zu verwenden?**
- **Welche Faktoren beeinflussen das Ergebnis?**








Tasks

-  Fill-Mask
-  Question Answering
-  Summarization
-  Table Question Answering
-  Text Classification
-  Text Generation
-  Text2Text Generation
-  Token Classification
-  Translation
-  Zero-Shot Classification
-  Sentence Similarity
- + 12

Libraries

-  PyTorch
-  TensorFlow
-  JAX
- + 21

Datasets

-  wikipedia
-  common_voice
-  bookcorpus
-  dcep europarl jrc-acquis
-  glue
-  conll2003
-  squad
-  oscar
- + 602

Languages

- en
- es
- fr
- de
- sv
- zh
- fi
- ja
- + 164

Licenses

- apache-2.0
- mit
- cc-by-4.0
- + 15

Models 19,973

[↑↓ Sort: Most Downloads](#)

bert-base-uncased

 Fill-Mask • Updated May 18 • 23.8M • ❤️ 52

roberta-large

 Fill-Mask • Updated May 21 • 13.4M • ❤️ 18

distilbert-base-uncased

 Fill-Mask • Updated Aug 29 • 5.64M • ❤️ 26

bert-base-cased

 Fill-Mask • Updated Sep 6 • 3.85M • ❤️ 4

gpt2

 Text Generation • Updated May 19 • 3.68M • ❤️ 22

distilbert-base-uncased-finetuned-sst-2-english

 Text Classification • Updated Feb 9 • 3.19M • ❤️ 16

roberta-base

 Fill-Mask • Updated Jul 6 • 3.12M • ❤️ 6

bert-base-multilingual-cased

 Fill-Mask • Updated May 18 • 2.61M • ❤️ 2

PROJEKTIDEEN

- **Vorhersage der Beantwortungsschwierigkeit von Aufgaben (Karo)**
- **Klassifikation von Antwort-Mails hinsichtlich Höflichkeit und ggf. hinsichtlich von fachlichen Kriterien (Chris und Sabrina)**
- **Die richtigen Worte für das perfekte Produkt (Leon)**
- **Sentiment-Analyse & Themen-Tagging von Nachrichtenartikeln (Leon)**
- **Sentiment-Analyse zur Vorhersage der Volatilität von Aktienkursen (Jule)**
- **Generieren wissenschaftlicher/philosophischer Texte (Martin)**
- **SHU-T: Generierung von Antworten auf Hass-Artikel (Martin)**
- **Creative Code: Generierung von Code für die Video-Bearbeitung (Martin)**
- **Paraphrasing Texts (Peyman)**
- **Automatische Erkennung/Ergänzung fehlender Produktattribute (Laura)**
- **Vorhersage von Produktkategorien anhand des Produkttextes (Laura)**

TODOS BIS ZUR NÄCHSTEN WOCHEN

- **Chapter 2 des Hugging Face Kurses absolvieren:**
<https://huggingface.co/course/chapter2>
(Selektiert dabei die Variante, die Ihr favorisiert: TensorFlow bzw. Pytorch)
- **Projektideen konkretisieren!**
- **Nächstes Mal sollten dann alle ein Team mit einem konkreten Projekt haben.**

Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets

Irene Solaiman*
OpenAI
irene@openai.com

Christy Dennison*
OpenAI
christy@openai.com

Abstract

Language models can generate harmful and biased outputs and exhibit undesirable behavior. We propose a Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets, an iterative process to significantly change model behavior by crafting and fine-tuning on a dataset that reflects a predetermined set of target values. We evaluate our process using three metrics: quantitative metrics with human evaluations that score output adherence to a target value, and toxicity scoring on outputs; and qualitative metrics analyzing the most common word associated with a given social category. Through each iteration, we add additional training dataset examples based on observed shortcomings from evaluations. PALMS performs significantly better on all metrics compared to baseline and control models for a broad range of GPT-3 language model sizes without compromising capability integrity. We find that the effectiveness of PALMS increases with model size. We show that significantly adjusting language model behavior is feasible with a small, hand-curated dataset.

1 Introduction

Progress in scaling up generative language models has enabled impressive results on a wide range of tasks, leading to novel research and industry applications. As language models increase in size and impact, increasing attention is being given to the social impact and cultural context of language models across research and industry organizations. The risks and potential harms of language models are difficult to identify, measure, and mitigate, as