Hugging Face

Search models, datasets, users...

google / **flan-t5-small**

like 200

Text2Text Generation | Transformers | PyTorch | TensorFlow | JAX | Safetensors

svakulenk0/qrecc | taskmaster2 | djaym7/wiki_dialog | deepmind/code_contests | lambada

gsm8k | aqua_rat | esnli | quasc | qed | 5 languages | t5 | Inference Endpoints

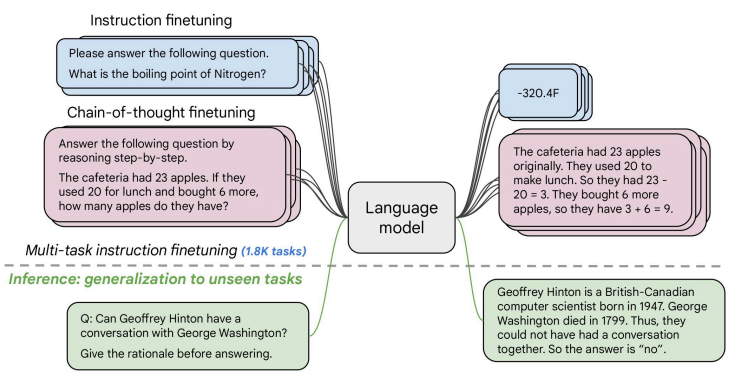
text-generation-inference | arxiv:2210.11416 | arxiv:1910.09700 | License: apache-2.0

Train | Deploy | Use this model

Model card | Files | Community 14

Edit model card

Model Card for FLAN-T5 small



Downloads last month
237,632

Safetensors

Model size	77M params
Tensor type	F32

Table of Contents

TL;DR

If you already know T5, FLAN-T5 is just better at everything. For the same number of parameters, these models have been fine-tuned on more than 1000 additional tasks covering also more languages. As mentioned in the first few lines of the abstract :

> *“ Flan-PaLM 540B achieves state-of-the-art performance on several benchmarks, such as 75.2% on five-shot MMLU. We also publicly release Flan-T5 checkpoints,1 which achieve strong few-*

Inference API

Text2Text Generation

Examples

The square root of x is the cube root of y. What is y to the power of 2, if x = 4?

Compute ctrl+Enter 0.0

This model can be loaded on Inference API (serverless).

JSON Output Maximize

Datasets used to train `google/flan…`

`gsm8k`
Viewer • Update… • 385k • 248

`deepmind/code_contests`
Viewer • Updated… • 4.12k • 86

`aqua_rat`
Viewer • Updated… • 3.24k • 19

Spaces using `google/flan-t5…` 100

> *shot performance even compared to much larger models, such as PaLM 62B. Overall, instruction finetuning is a general method for improving the performance and usability of pretrained language models."*

Disclaimer: Content from **this** model card has been written by the Hugging Face team, and parts of it were copy pasted from the T5 model card.

Model Details

Model Description

- **Model type:** Language model
- **Language(s) (NLP):** English, Spanish, Japanese, Persian, Hindi, French, Chinese, Bengali, Gujarati, German, Telugu, Italian, Arabic, Polish, Tamil, Marathi, Malayalam, Oriya, Panjabi, Portuguese, Urdu, Galician, Hebrew, Korean, Catalan, Thai, Dutch, Indonesian, Vietnamese, Bulgarian, Filipino, Central Khmer, Lao, Turkish, Russian, Croatian, Swedish, Yoruba, Kurdish, Burmese, Malay, Czech, Finnish, Somali, Tagalog, Swahili, Sinhala, Kannada, Zhuang,

facebook/MusicGen
Surn/UnlimitedMusicGen
Sharathhebbar24/One-stop-for-Ope…
GrandaddyShmax/AudioCraft_Plus
SpacesExamples/fastapi_t5
GrandaddyShmax/MusicGen_Plus
GrandaddyShmax/MusicGen_Plus_h…
unpairedelectron07/Text-to-Music-Ge…
Nick088/SuperPrompt-v1
Manjushri/MusicGen
radames/Candle-T5-Generation-W…
sunnyujjawal/AI-Music-Generator
patgpt4/MusicGen
AchyuthGamer/MusicGen
Gyufyjk/AudioCraft_Plus
awacke1/DockerGoFlanT5
Omnibus/MusicGen
ZeroTwo3/videoshop-backend
jbilcke-hf/ai-tube-model-musicgen-1
Fabrice-TIERCELIN/Text-to-Music
+ 80 Spaces

Collection including `google/flan-t…`

Igbo, Xhosa, Romanian, Haitian, Estonian, Slovak, Lithuanian, Greek, Nepali, Assamese, Norwegian

- **License:** Apache 2.0
- **Related Models:** All FLAN-T5 Checkpoints
- **Original Checkpoints:** All Original FLAN-T5 Checkpoints
- **Resources for more information:**
  - Research paper
  - GitHub Repo
  - Hugging Face FLAN-T5 Docs (Similar to T5)

Usage

Find below some example scripts on how to use the model in `transformers`:

Using the Pytorch model

Running the model on a CPU

▶ Click to expand

Running the model on a GPU

Flan-T5 release Collection
The Flan-T5 cov… • 7 items • U.. • △ 14

▶ Click to expand

Running the model on a GPU using different precisions

FP16

▶ Click to expand

INT8

▶ Click to expand

Uses

Direct Use and Downstream Use

The authors write in the original paper's model card that:

> *"The primary use is research on language models, including: research on zero-shot NLP tasks and in-context few-shot learning NLP tasks, such as reasoning, and question answering; advancing fairness and safety research, and understanding limitations of current large language models"*

See the research paper for further details.

Out-of-Scope Use

More information needed.

Bias, Risks, and Limitations

The information below in this section are copied from the model's official model card:

> *“Language models, including Flan-T5, can potentially be used for language generation in a harmful way, according to Rae et al. (2021). Flan-T5 should not be used directly in any application, without a prior assessment of safety and fairness concerns specific to the application.”*

Ethical considerations and risks

> *“Flan-T5 is fine-tuned on a large corpus of text data that was not filtered for explicit content or assessed for existing biases. As a result the model itself is potentially vulnerable to generating equivalently inappropriate content or replicating inherent biases in the underlying data.”*

Known Limitations

> *"Flan-T5 has not been tested in real world applications."*

Sensitive Use:

> *"Flan-T5 should not be applied for any unacceptable use cases, e.g., generation of abusive speech."*

Training Details

Training Data

The model was trained on a mixture of tasks, that includes the tasks described in the table below (from the original paper, figure 2):

table.png

Training Procedure

According to the model card from the original paper:

> *"These models are based on pretrained T5 (Raffel et al., 2020) and fine-tuned with instructions for*

> *better zero-shot and few-shot performance. There is one fine-tuned Flan model per T5 model size."*

The model has been trained on TPU v3 or TPU v4 pods, using `t5x` codebase together with `jax`.

Evaluation

Testing Data, Factors & Metrics

The authors evaluated the model on various tasks covering several languages (1836 in total). See the table below for some quantitative evaluation:

image.png

For full details, please check the research paper.

Results

For full results for FLAN-T5-Small, see the research paper, Table 3.

Environmental Impact

Carbon emissions can be estimated using the Machine Learning Impact calculator presented in Lacoste et al. (2019).

- **Hardware Type:** Google Cloud TPU Pods - TPU v3 or TPU v4 | Number of chips ≥ 4.
- **Hours used:** More information needed
- **Cloud Provider:** GCP
- **Compute Region:** More information needed
- **Carbon Emitted:** More information needed

Citation

BibTeX:

```
@misc{https://doi.org/10.48550/arxiv.2210.1
  doi = {10.48550/ARXIV.2210.11416},

  url = {https://arxiv.org/abs/2210.11416}

  author = {Chung, Hyung Won and Hou, Le a

  keywords = {Machine Learning (cs.LG), Co

  title = {Scaling Instruction-Finetuned L
```

```
  publisher = {arXiv},

  year = {2022},

  copyright = {Creative Commons Attributio
}
```