



# Accelerating Transformers with Hugging Face Optimum and Infinity



**MLOps World - June 2022**

<https://github.com/huggingface/optimum>

<https://huggingface.co/infinity>

<https://huggingface.co/>

# About us



💻 **Philipp Schmid**

💼 Technical Lead

✉️ @\_philschmid

🔗 /philipp-schmid-a6a2bb196



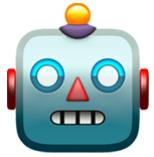
💻 **Lewis Tunstall**

💼 Machine Learning Engineer

✉️ @\_lewtun

🔗 /lewis-tunstall

# Plan of attack



Transformers



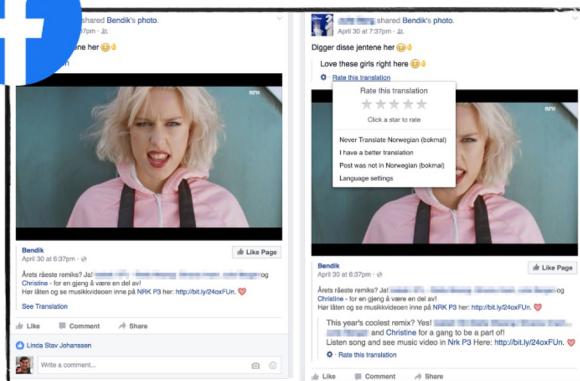
Hugging Face



Optimum



Infinity



Taco Tuesday

Jacqueline Bruzek x

Taco Tuesday

Hey Jacqueline,

Haven't seen you in a while and I hope you're doing well.



Microsoft Edge

what is critical race theory

ALL IMAGES VIDEOS MAPS NEWS SHOPPING  
in education in social work in sociology literature

Critical race theory  
3. Understanding racism requires understanding perceptions of those who have experienced it, because they often invisible to those who benefit from it  
• Reject essentiality of white experience  
• Reject essentiality of white experience  
• openly acknowledge that perceptions reflect the dominant culture's values and perspectives  
• the concept of racial hierarchy is a dominant narrative that provides the interests of the majority group

What is Critical Race Theory  
Examines relationships between race, law, and power.  
Includes economics, history, politics, law, and culture.  
Focuses on civil rights laws as fundamentally 'colorblind'.  
Analyze dominant ideas of society.  
Challenge dominant ideologies  
Rober Ogden, Ulrich/Green: An Introduction

Critical Race Theory: A concept  
Defining elements  
• focus on racism... "racism" not inherent nor race; discourses of race  
• critique of civil rights laws as fundamentally 'colorblind'  
• analyze dominant ideas of society.  
Analyze dominant ideas of society.  
Challenges dominant ideologies  
Rober Ogden, Ulrich/Green: An Introduction

Critical race theory (CRT), intellectual movement and loosely organized framework of legal analysis based on the premise that race is not a natural, biologically grounded feature of physically distinct subgroups of human beings but a socially constructed (culturally invented) category that is used to oppress and exploit people of colour.

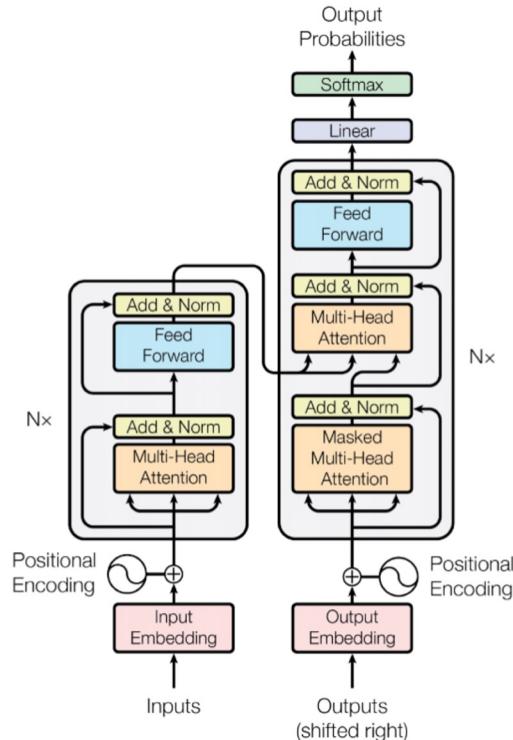
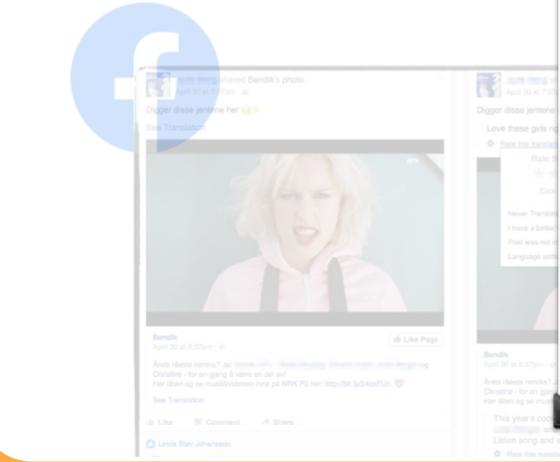
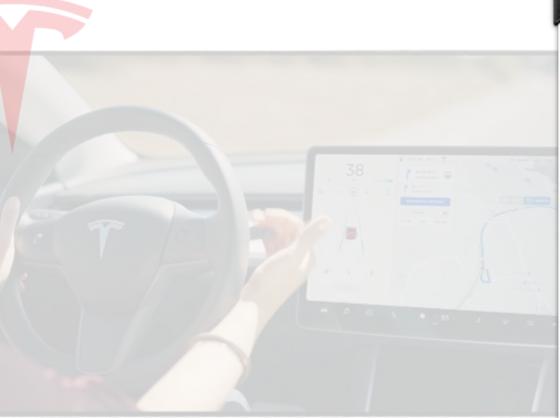


Figure 1: The Transformer - model architecture.

Critical Race Theory (CRT), intellectual movement and loosely organized framework of legal analysis based on the premise that race is not a natural



# Transformers for the rest of us?



Humble ML engineer



# The AI community building the future.

Build, train and deploy state of the art models powered by  
the reference open source in machine learning.



Star

59,315

More than 5,000 organizations are using Hugging Face

 **AI2 Allen Institute for AI**  
Non-Profit • 83 models

 **Facebook AI**  
Company • 248 models

 **Graphcore**  
Company • 15 models

 **Google AI**  
Company • 492 models

 **Amazon Web Services**  
Company • 1 model

 **SpeechBrain**  
Non-Profit • 36 models

 **Microsoft**  
Company • 140 models

 **Grammarly**  
Company



# Hugging Face Hub - where the magic happens



The screenshot shows the Hugging Face Hub homepage. On the left, there's a sidebar for the user 'victor' with options like Profile, Settings, Organizations, Resources, and a Sign Up button. The main area features a search bar at the top. Below it is a 'Victor M's Activity' feed with posts about liked models, updated datasets, and updated models. To the right is a 'Trending' section for the last 7 days, listing models like 'AnimeGANv2', 'kakaobrain/kogpt', 'bigscience/T0pp', 'oscar-corpus/OSCAR-2109', 'WhatTheFood', 'ImageCaptioning', and 'facebook/wmt21-dense-24-wide-en-x'. Each item has a thumbnail, name, description, and a heart count.

<https://huggingface.co/>

# Solving all kinds of problems



NLP



Speech



Vision



Domain X



Bio & CH



Time



RL

# Model repositories - optimised for ML collaboration



Hugging Face

Models Datasets Spaces Docs Solutions Pricing Log In Sign Up

**gpt2** like 113

Text Generation PyTorch TensorFlow JAX TF Lite Rust Transformers en mit gpt2 exbert

Model card Files and versions Community 1 Edit model card

Train Deploy Use in Transformers

**GPT-2**

Test the whole generation capabilities here: <https://transformer.huggingface.co/doc/gpt2-large>

Pretrained model on English language using a causal language modeling (CLM) objective. It was introduced in [this paper](#) and first released at [this page](#).

Disclaimer: The team releasing GPT-2 also wrote a [model card](#) for their model. Content from this model card has been written by the Hugging Face team to complete the information they provided and give specific examples of bias.

**Model description**

GPT-2 is a transformers model pretrained on a very large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was trained to guess the next word in sentences.

Downloads last month  
**12,949,594**

Hosted inference API

Text Generation Example 5

My name is Lewis and I like to

Compute

Computation time on cpu: 1.131 s

My name is Lewis and I like to call myself 'Nike', I am not an expert in tennis. I have actually studied tennis for three years and finally I decided to go and compete. The match did not feel interesting either, a bunch of

JSON Output Maximize

Spaces using gpt2



# Pull requests and discussions

Hugging Face

Models Datasets Spaces Docs Solutions Pricing Log In Sign Up

## distilgpt2

like 57

Text Generation PyTorch TensorFlow JAX TF Lite Rust Transformers openwebtext en arxiv:1910.01108 arxiv:2201.08542 arxiv:2203.12574 arxiv:1910.09700 arxiv:1503.02531

apache-2.0 gpt2 exbert Eval Results Carbon Emissions

Model card Files and versions Community 3 Train Deploy Use in Transformers

**Community Tab** Start discussions and open PR in the Community Tab.

New discussion New pull request

All Discussions Pull requests Show closed

**Limit use of collapsible sections; fix emissions info**  
#3 opened 10 days ago by Marissa □ 6

**Add emissions estimate to metadata**  
#2 opened 12 days ago by Marissa □ 5

**New model card for distilgpt2**  
#1 opened 12 days ago by Marissa □ 9

Resources

Announcement blog post PR & discussions documentation Hub documentation



# Built on top of open-source

 **transformers** Public

 Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX.

● Python ⭐ 64.5k 📂 15.1k

 **datasets** Public

 The largest hub of ready-to-use datasets for ML models with fast, easy-to-use and efficient data manipulation tools

● Python ⭐ 13.6k 📂 1.7k

 **tokenizers** Public

 Fast State-of-the-Art Tokenizers optimized for Research and Production

● Rust ⭐ 5.7k 📂 478

 **accelerate** Public

 A simple way to train and use PyTorch models with multi-GPU, TPU, mixed-precision

● Python ⭐ 2.5k 📂 176

 **optimum** Public

 Accelerate training and inference of  Transformers with easy to use hardware optimization tools

● Python ⭐ 546 📂 43

 **huggingface\_hub** Public

All the open source things related to the Hugging Face Hub.

● Python ⭐ 439 📂 107



# Hugging Face Optimum

<https://github.com/huggingface/optimum>

- An open-source library and extension of Hugging Face Transformers
- Provides a unified API of performance optimization tools to achieve maximum efficiency to train and run models on accelerated hardware
- Can be used for *accelerated training*, *quantization*, and *graph optimization*, with inference support for transformers pipelines.

**GRAPHCORE**

Train Transformers faster with IPUs

 **habana**<sup>®</sup>

Accelerate Transformers Training on Gaudi



**ONNX  
RUNTIME**

**intel**

Scale with Xeon



# Hugging Face Optimum

<https://github.com/huggingface/optimum>

Inference acceleration in 2 quick steps



Export your model  
into ONNX

`transformers`



Optimize your ONNX, including  
to your specific hardware

`optimum`

Today's workshop:

- Acceleration with *graph optimization* & *quantization*
- Running inference with **ONNX Runtime** in Optimum

<https://huggingface.co/>

# Intent classification as a use case



- **Goal:** classify customer queries into *intents* such as Purchase, Unsubscribe, Lost Card etc
- A common component in digital assistants → low latencies (sub millisecond) are critical!
- Many challenges around identifying in-scope vs out-of-scope queries

Today's workshop:

- Use the [BANKING77 dataset](#) (77 intents)
- Optimise a fine-tuned [DistilBERT](#) model

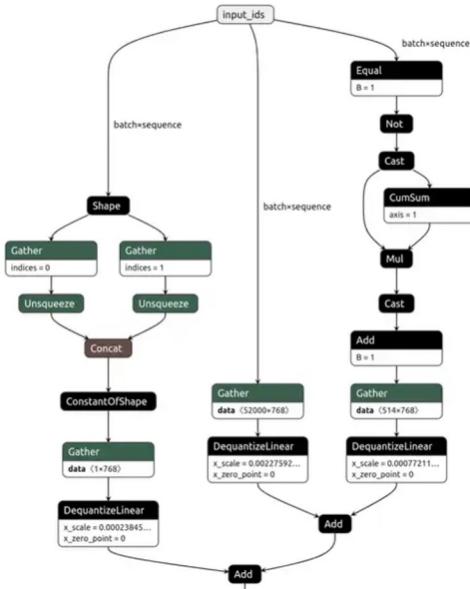
The image shows three interactions with a digital assistant, each consisting of a numbered circle, a question in a blue box, and a response in a grey box.

- 1** What is my balance?  
You have \$1,847.51 across your 3 accounts. ✓
- 2** How are my sports teams doing?  
Your last payday was on the 1st of November. ✗
- 3** Who has the best record in the NBA?  
Sorry, I can only answer questions about banking. ✓

# Exporting Transformers to ONNX



- ONNX is an open standard that defines a *common set of operators* and file format for deep learning models
- Operators used to construct a computational graph or *intermediate representation* (IR) that represents flow of data through the network
- Shines when coupled with a dedicated accelerator like ONNX Runtime or TensorRT



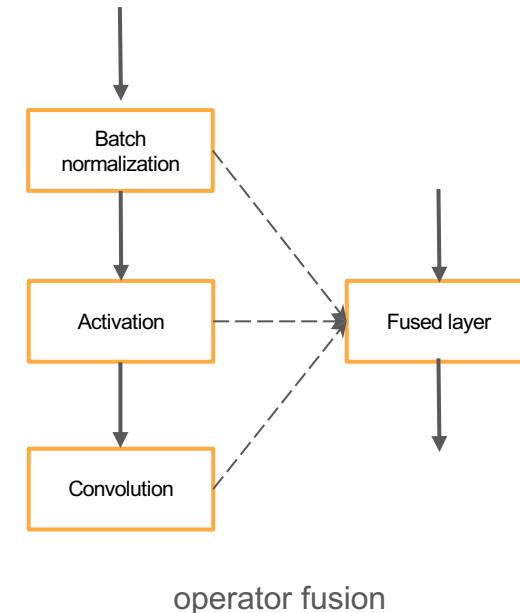
ONNX graph

# Making models faster with graph optimization



Basic idea:

- Apply **transformations** to the graph **nodes** and **layout**
- **Constant folding**: evaluate constant expressions at compile time instead of runtime
- **Redundant node elimination**: remove redundant nodes without changing graph structure
- **Operator fusion**: merge one node (i.e. operator) into another so they can be executed together



# Making models faster with **quantization**

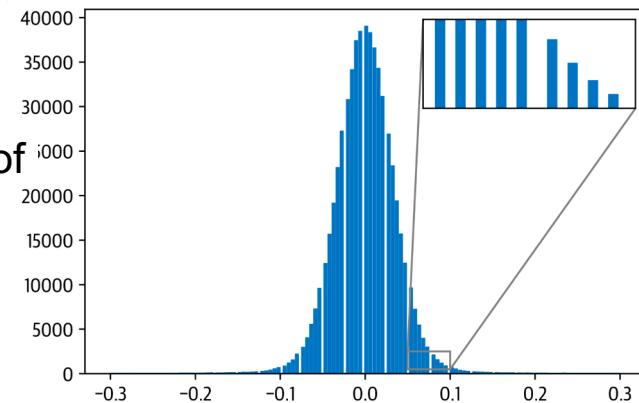


Basic idea:

- Represent weights and activations with **low-precision data types** like 8-bit integer instead of 32-bit floating point.
- Less memory storage & faster matmuls!

In practice:

- Map range of floating-point values to **smaller range**



$$\text{real\_value} = (\text{int8\_value} - \text{zero\_point}) \times \text{scale}$$



# Making models faster with **quantization**

Three main ways to quantize:

- **Dynamic quantization:** quantize weights & activations on-the-fly. Simplest to start with.
- **Static quantization:** precompute quantization scheme by observing activation patterns on sample of data. Generally gives better latency, but more complex to calibrate.
- **Quantization aware training:** simulate quantization during training with "fake" quantization of FP32 values.



# Hugging Face Infinity



## Plug and Predict

Infinity comes as a single-container and can be deployed in any production environment. It can easily be scaled to thousands of requests every second using orchestration services like kubernetes.



## Unmatched Performance

Infinity achieves unmatched performance for state-of-the-art transformer models. Infinity achieves 1ms latency for BERT-like models on GPU, and 4ms on CPU.



## Enterprise Ready

Infinity meets the highest security requirements and can be integrated anywhere from public clouds to air gapped environments. You control your models, your data, and the traffic.

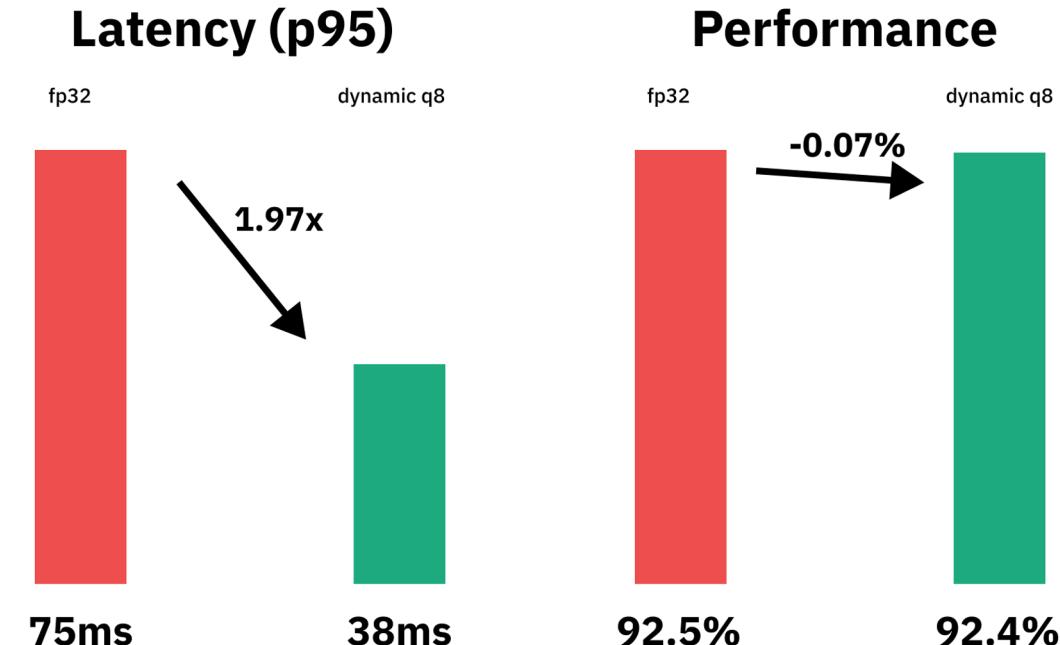


# Free Infinity on Hugging Face Inference Endpoint soon...



# Session 1: Dynamic Quantization & Optimization with Optimum

\*run on c6i.xlarge with sequence length of 128





# Session 2: Post-Training Static Quantization with Optimum

\*run on c6i.xlarge with sequence length of 128

## Latency (p95)

fp32



q8



2.83x

75ms

26ms

## Performance

fp32



q8



-0.28%

92.5%

92.3%

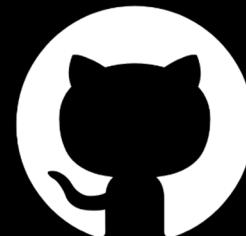
# Where to get help



<https://hf.co/join/discord>



[https://  
discuss.huggingface.co/](https://discuss.huggingface.co/)



[https://github.com/  
huggingface/optimum](https://github.com/huggingface/optimum)