



Useful Tools in NLP

Models, Datasets, Evaluation, Logging, Optimisation



PyTorch Lightning



W&B

Ideology

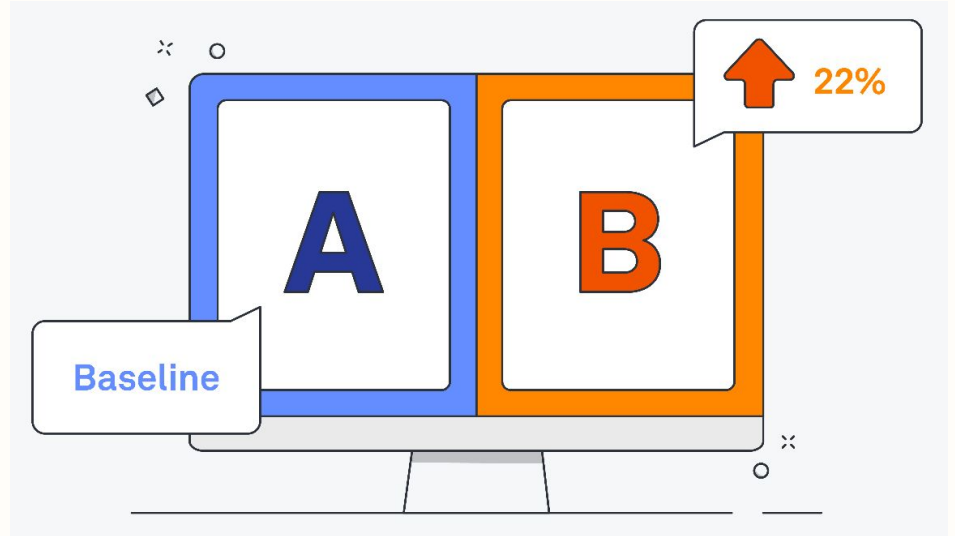
- Don't reinvent the wheel

"If I have seen
further than
others, it is by
standing upon the
shoulders of
giants."

— Sir Isaac Newton

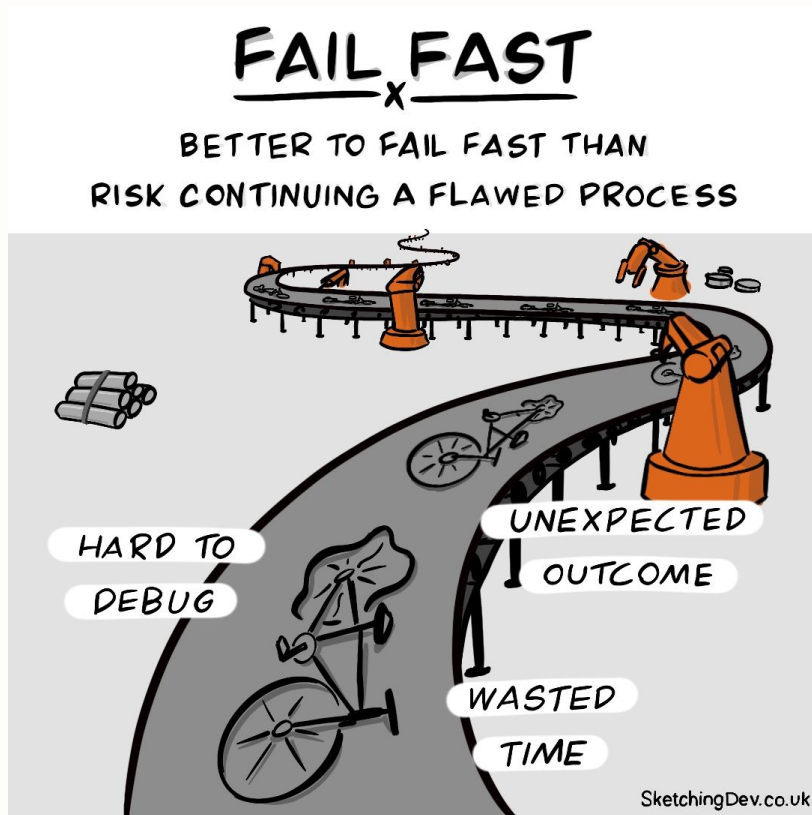
Ideology

- Don't reinvent the wheel
- Only add complexity for your hypothesis



Ideology

- Don't reinvent the wheel
- Only add complexity for your hypothesis
- Do as others do
- **Fail fast, early and often!**



Pipeline



Data



Models & Tokenizers



Evaluation



Logging & Optimisation



Simplification

Pipeline



Data



Models & Tokenizers



Evaluation



Logging & Optimisation



Simplification

Data



- Many datasets available!

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity
- Table to Text
- Multiple Choice
- Text Retrieval

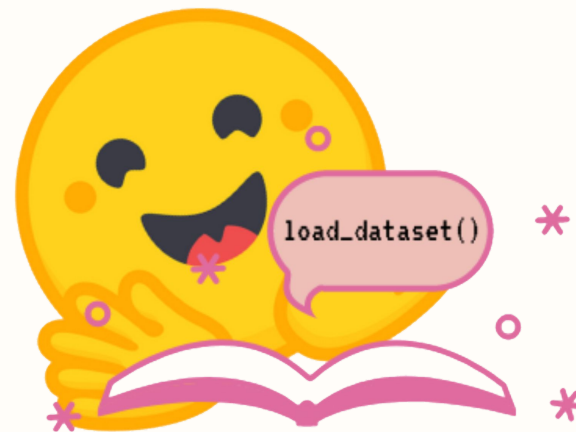
Audio

squad_v2 Viewer · Updated Apr 5 · ± 1.86M · ♥ 87	cais/mmlu Viewer · Updated 17 days ago · ± 1.59M · ♥ 85
glue Viewer · Updated Jun 1 · ± 1.39M · ♥ 241	super_glue Viewer · Updated Apr 5 · ± 717k · ♥ 115
poloclub/diffusiondb Viewer · Updated May 9 · ± 615k · ♥ 319	wikitext Viewer · Updated Jun 20 · ± 549k · ♥ 192
lighteval/mmlu Viewer · Updated Jun 9 · ± 546k · ♥ 6	HuggingFaceM4/COCO Viewer · Updated Dec 15, 2022 · ± 481k · ♥ 7
ai2_arc Viewer · Updated Apr 5 · ± 376k · ♥ 27	imdb Viewer · Updated Apr 5 · ± 302k · ♥ 122
hf-internal-testing/fixtures_image_utils Viewer · Updated Dec 7, 2021 · ± 283k	lavita/medical-qa-shared-task-v1-toy Viewer · Updated Jul 20 · ± 273k · ♥ 1
lavita/medical-qa-shared-task-v1-toy-eval Viewer · Updated Jul 27 · ± 267k	piqa Viewer · Updated Jan 25 · ± 261k · ♥ 45
trec Viewer · Updated Apr 5 · ± 258k · ♥ 30	winogrande Viewer · Updated Jun 5 · ± 248k · ♥ 24
squad Viewer · Updated Apr 5 · ± 219k · ♥ 140	wikiann Viewer · Updated Jun 1 · ± 219k · ♥ 57
openbookqa Viewer · Updated Apr 5 · ± 201k · ♥ 37	common_voice Viewer · Updated Jun 27 · ± 170k · ♥ 102

Data



- Many datasets available!
- **Easy access:**
 - `load_dataset("name")`



Data



- Many datasets available!
- Easy access:
 - `load_dataset("name")`
- **Dataset object functionality:**
 - Saving
 - Loading
 - Multiprocessing (map, reduce, etc)
 - Streaming data
 - Querying data



[HuggingFace - dataset object + functionality](#)

Pipeline



Data



Models & Tokenizers



Evaluation



Logging & Optimisation



Simplification

Models & Tokenizers

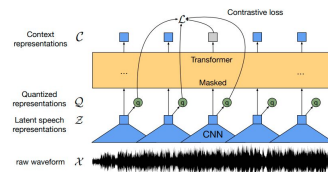
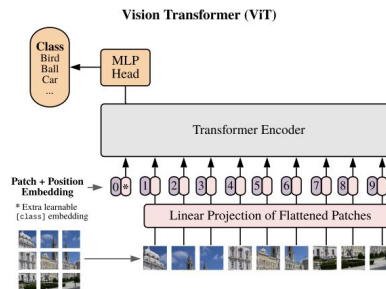
- Many models available!




GPT-2



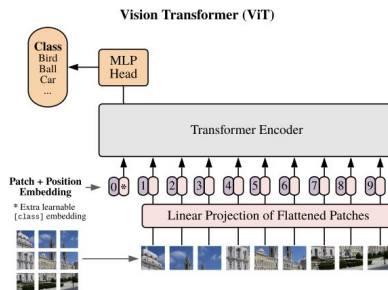
**Falcon model &
Stable diffusion**



[HuggingFace - Available models](#)

Models & Tokenizers

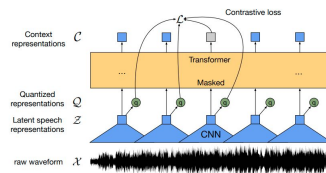
- Many models available!
- Easy access:
 - `AutoTokenizer.from_pretrained("name")`
 - `AutoModel.from_pretrained("name")`



GPT-2



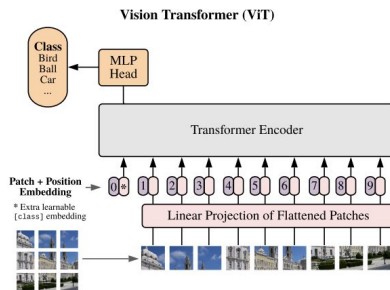
Falcon model & Stable diffusion



[HuggingFace - Models tutorial](#)
[HuggingFace - Examples](#)

Models & Tokenizers

- Many models available!
- Easy access:
 - `AutoTokenizer.from_pretrained("name")`
 - `AutoModel.from_pretrained("name")`
- Self contained:
 - `Modeling_“name”.py`
 - `Config_“name”.py`
 - `Tokenization_“name”.py`



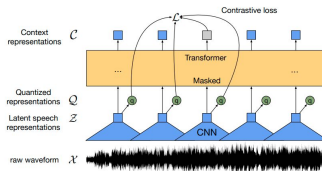
[Github - Transformers](https://github.com/huggingface/transformers)



GPT-2



Falcon model & Stable diffusion



Pipeline



Data



Models & Tokenizers



Evaluation



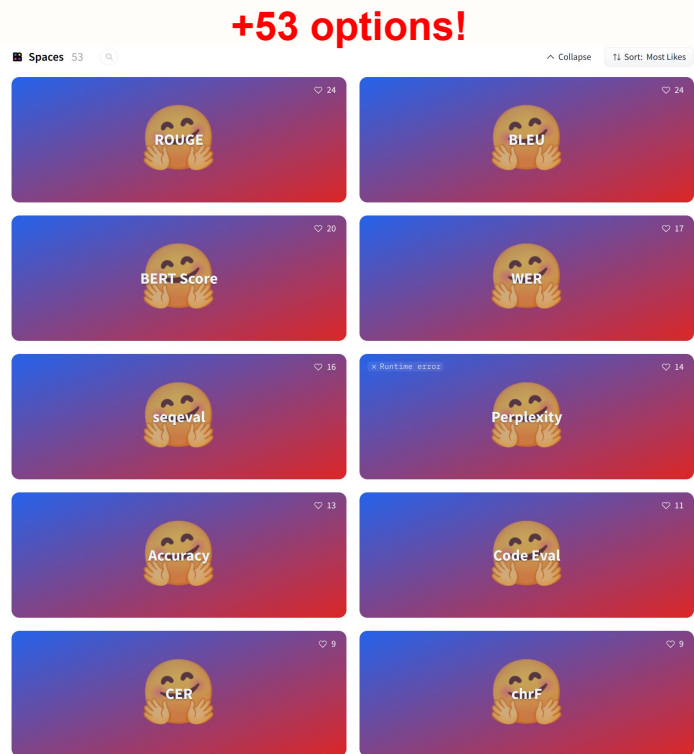
Logging & Optimisation



Simplification

Evaluation

- Many metrics available!



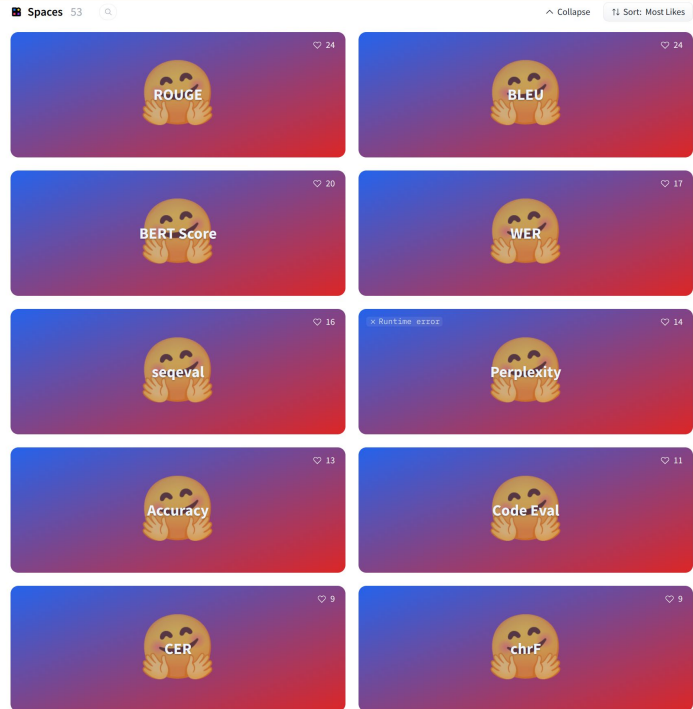
[HuggingFace - Evaluate](#)



Evaluation

- Many metrics available!
- Easy access:
 - `evaluate.load("name")`

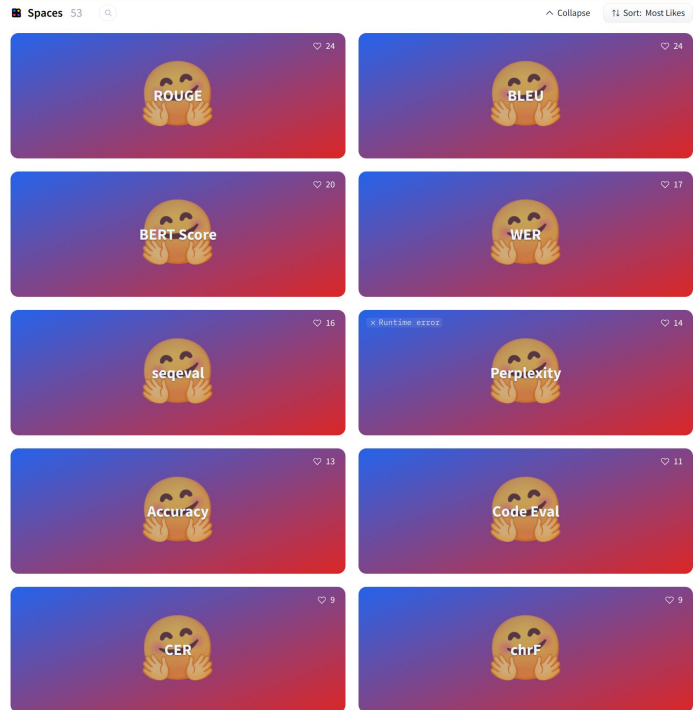
Don't reinvent the wheel



[HuggingFace - Tutorial](#)

Evaluation

- Many metrics available!
- Easy access:
 - `evaluate.load("name")`
- Choose the right metric
 - Classification?
 - Language modeling?
 - Speech modeling?
 - Machine Translation?
 - Summarisation?

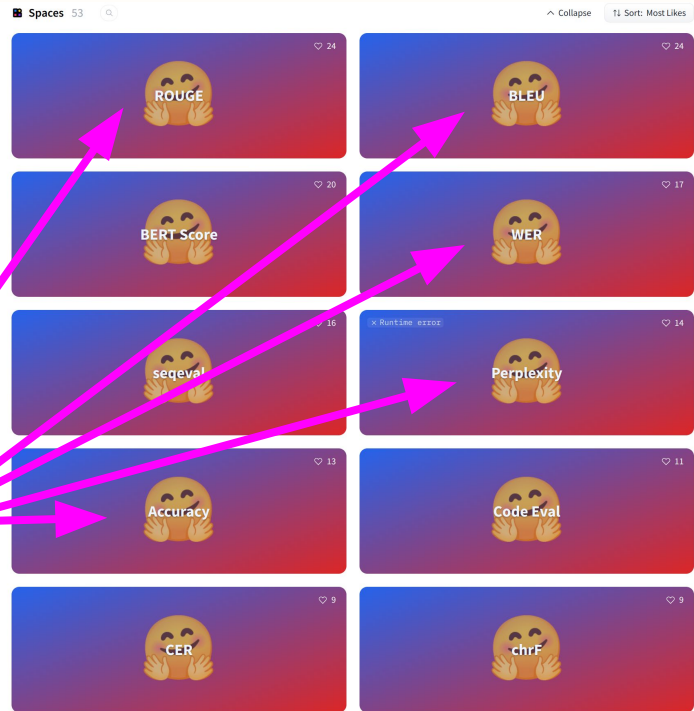


[HuggingFace - choose the right metric](#)

Evaluation



- Many metrics available!
- Easy Access:
 - `evaluate.load("name")`
- Choose the right metric
 - Classification?
 - Language modeling?
 - Speech modeling?
 - Machine Translation?
 - Summarisation?



[HuggingFace - choose the right metric](#)

Pipeline



Data



Models & Tokenizers



Evaluation



Logging & Optimisation



Simplification



Weights & Biases

- Logging visualisations

Fail fast, early and often!

Weights & Biases



- Logging visualisations
- **Easy access:**
 - `wandb.init()`
 - `wandb.log()` - metrics, images, text, plotly
 - Integration 🔥 😊 ⚡

Fabio... show them something.



Weights & Biases

- Logging visualisations
- Easy access:
 - `wandb.init()`
 - `wandb.log()` - metrics, images, text, plotly
 - Integration 🔥 😊 🔮
- Versioning + Debugging (system)

**Fabio... show them
something, again.**



Weights & Biases

- Logging visualisations
- Easy Access:
 - `wandb.init()`
 - `wandb.log()` - metrics, images, text, plotly
 - Integration 🔥 😊 ⚡
- Versioning + Debugging (system)
- Sweeps for hyperparameters

**Fabio... now show them
something cool!**

[WandB - Guides](#)

Pipeline



Data



Models & Tokenizers



Evaluation



Logging & Optimisation



Simplification

Pytorch Lightning

- **One boilerplate to rule them all!**
 - Reusable boilerplate

Do as others do

I need an ML
boilerplate...

It should be simple,
reusable, readable...



Too many boilerplates...

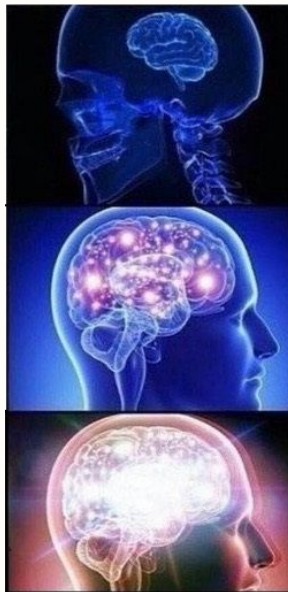
Pytorch Lightning

- One boilerplate to rule them all!
 - Reusable boilerplate
- **Encapsulation vs Flexibility trade-off**

I need an ML boilerplate...

It should be simple, reusable, readable...

Flexibility without complications...



Only add complexity for your hypothesis

[Introduction to PyTorch Lightning](#)

Pytorch Lightning

- One boilerplate to rule them all!
 - Reusable boilerplate
- Encapsulation vs Flexibility trade-off
- **Easy integration**
 - **Logging, checkpointing**
 - **Mixed-precision, gradient clipping, scheduling**
 - **Debugging (overfit, less data, etc)**
 - **Data parallel, multi-GPUs, Deepspeed**

I need an ML
boilerplate...

It should be simple,
reusable, readable...

Flexibility without
complications...

PyTorch Lightning!

[Introduction to PyTorch Lightning](#)



Recap



Data



Models & Tokenizers



Evaluation



Logging & Optimisation



Simplification

Recap



Data



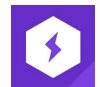
Models & Tokenizers



Evaluation



Logging & Optimisation



Simplification

PhD without NLP tools...



PhD with NLP tools!

