[![Hugging Face's logo](/front/assets/huggingface_logo-noborder.svg) Hugging

Face](/)

* [Models](/models)
* [Datasets](/datasets)
* [Spaces](/spaces)
* [Posts](/posts)
* [Docs](/docs)
* [Enterprise](/enterprise)
* [Pricing](/pricing)
* * * * *

* [Log In](/login)
* [Sign Up](/join)

#

[![](https://cdn-

avatars.huggingface.co/v1/production/uploads/1609621322398-5eff4688ff69163f6f59e66c.png)](/se

ntence-

transformers)

[sentence-transformers](/sentence-transformers)

/

[all-MiniLM-L6-v1](/sentence-transformers/all-MiniLM-L6-v1)

like 15

Follow

Sentence Transformers 1.26k

[Sentence Similarity](/models?pipeline_tag=sentence-similarity)[sentence-transformers](/models?library=sentence-transformers)[PyTorch](/models?library=pytorch)[ONNX](/models?library=onnx)[Safetensors](/models?library=safetensors)[OpenVINO](/models?library=openvino)[Transformers](/models?library=transformers)[English](/models?language=en)[bert](/models?other=bert)[feature-extraction](/models?other=feature-extraction)[text-embeddings-inference](/models?other=text-embeddings-inference)[Inference Endpoints](/models?other=endpoints_compatible)

arxiv: 5 papers

License: apache-2.0

[Model card](/sentence-transformers/all-MiniLM-L6-v1)[Files Files and versions](/sentence-transformers/all-MiniLM-L6-v1/tree/main)[Community 3](/sentence-transformers/all-MiniLM-L6-v1/discussions)

Train

Deploy

Use this model

A newer version of this model is available: [sentence-transformers/all-MiniLM-L6-v2](/sentence-transformers/all-MiniLM-L6-v2)

* all-MiniLM-L6-v1
  * Usage (Sentence-Transformers)
  * Usage (HuggingFace Transformers)
  * Evaluation Results
  * Background
  * Intended uses
  * Training procedure
    * Pre-training
    * Fine-tuning

all-MiniLM-L6-v1

This is a [sentence-transformers](https://www.SBERT.net) model: It maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.

Usage (Sentence-Transformers)

Using this model becomes easy when you have [sentence-

transformers](https://www.SBERT.net) installed:

```
pip install -U sentence-transformers
```

Then you can use the model like this:

```
from sentence_transformers import SentenceTransformer
sentences = ["This is an example sentence", "Each sentence is converted"]

model = SentenceTransformer('sentence-transformers/all-MiniLM-L6-v1')
embeddings = model.encode(sentences)
print(embeddings)
```

Usage (HuggingFace Transformers)

Without [sentence-transformers](https://www.SBERT.net), you can use the model like this: First, you pass your input through the transformer model, then you have to apply the right pooling-operation on-top of the contextualized word embeddings.

```
from transformers import AutoTokenizer, AutoModel
import torch
import torch.nn.functional as F

#Mean Pooling - Take attention mask into account for correct averaging
def mean_pooling(model_output, attention_mask):
    token_embeddings = model_output[0] #First element of model_output contains all token embeddings
    input_mask_expanded = attention_mask.unsqueeze(-1).expand(token_embeddings.size()).float()
    return torch.sum(token_embeddings * input_mask_expanded, 1) / torch.clamp(input_mask_expanded.sum(1), min=1e-9)


# Sentences we want sentence embeddings for
sentences = ['This is an example sentence', 'Each sentence is converted']

# Load model from HuggingFace Hub
tokenizer = AutoTokenizer.from_pretrained('sentence-transformers/all-MiniLM-L6-v1')
model = AutoModel.from_pretrained('sentence-transformers/all-MiniLM-L6-v1')

# Tokenize sentences
encoded_input = tokenizer(sentences, padding=True, truncation=True, return_tensors='pt')

# Compute token embeddings
with torch.no_grad():
```

```
        model_output = model(**encoded_input)

    # Perform pooling
    sentence_embeddings = mean_pooling(model_output, encoded_input['attention_mask'])

    # Normalize embeddings
    sentence_embeddings = F.normalize(sentence_embeddings, p=2, dim=1)

    print("Sentence embeddings:")
    print(sentence_embeddings)
```

Evaluation Results

For an automated evaluation of this model, see the _Sentence Embeddings Benchmark_ : [https://seb.sbert.net](https://seb.sbert.net?model_name=sentence-transformers/all-MiniLM-L6-v1)

* * *

Background

The project aims to train sentence embedding models on very large sentence level datasets using a self-supervised contrastive learning objective. We used the pretrained [`nreimers/MiniLM-L6-H384-uncased`](https://huggingface.co/nreimers/MiniLM-L6-H384-uncased)

model and fine-tuned in on a 1B sentence pairs dataset. We use a contrastive learning objective: given a sentence from the pair, the model should predict which out of a set of randomly sampled other sentences, was actually paired with it in our dataset.

We developped this model during the [Community week using JAX/Flax for NLP & CV](https://discuss.huggingface.co/t/open-to-the-community-community-week-using-jax-flax-for-nlp-cv/7104), organized by Hugging Face. We developped this model as part of the project: [Train the Best Sentence Embedding Model Ever with 1B Training Pairs](https://discuss.huggingface.co/t/train-the-best-sentence-embedding-model-ever-with-1b-training-pairs/7354). We benefited from efficient hardware infrastructure to run the project: 7 TPUs v3-8, as well as intervention from Googles Flax, JAX, and Cloud team member about efficient deep learning frameworks.

Intended uses

Our model is intented to be used as a sentence and short paragraph encoder. Given an input text, it ouptuts a vector which captures the semantic information. The sentence vector may be used for information retrieval, clustering or sentence similarity tasks.

By default, input text longer than 128 word pieces is truncated.

Training procedure

Pre-training

We use the pretrained [`nreimers/MiniLM-L6-H384-uncased`](https://huggingface.co/nreimers/MiniLM-L6-H384-uncased) model. Please refer to the model card for more detailed information about the pre-training procedure.

Fine-tuning

We fine-tune the model using a contrastive objective. Formally, we compute the cosine similarity from each possible sentence pairs from the batch. We then apply the cross entropy loss by comparing with true pairs.

Hyper parameters

We trained ou model on a TPU v3-8. We train the model during 100k steps using a batch size of 1024 (128 per TPU core). We use a learning rate warm up of 500. The sequence length was limited to 128 tokens. We used the AdamW optimizer with a 2e-5 learning rate. The full training script is accessible in this current repository: `train_script.py`.

Training data

We use the concatenation from multiple datasets to fine-tune our model. The total number of sentence pairs is above 1 billion sentences. We sampled each dataset given a weighted probability which configuration is detailed in the `data_config.json` file.

Dataset | Paper | Number of training tuples

---|---|---

[Reddit comments (2015-2018)](https://github.com/PolyAI-LDN/conversational-datasets/tree/master/reddit) | [paper](https://arxiv.org/abs/1904.06472) | 726,484,430

[S2ORC](https://github.com/allenai/s2orc) Citation pairs (Abstracts) | [paper](https://aclanthology.org/2020.acl-main.447/) | 116,288,806

[WikiAnswers](https://github.com/afader/oqa#wikianswers-corpus) Duplicate question pairs | [paper](https://doi.org/10.1145/2623330.2623677) | 77,427,422

[PAQ](https://github.com/facebookresearch/PAQ) (Question, Answer) pairs | [paper](https://arxiv.org/abs/2102.07033) | 64,371,441

[S2ORC](https://github.com/allenai/s2orc) Citation pairs (Titles) | [paper](https://aclanthology.org/2020.acl-main.447/) | 52,603,982

[S2ORC](https://github.com/allenai/s2orc) (Title, Abstract) | [paper](https://aclanthology.org/2020.acl-main.447/) | 41,769,185

[Stack Exchange](https://huggingface.co/datasets/flax-sentence-embeddings/stackexchange_xml) (Title, Body) pairs | - | 25,316,456

[MS MARCO](https://microsoft.github.io/msmarco/) triplets | [paper](https://doi.org/10.1145/3404835.3462804) | 9,144,553

[GOOAQ: Open Question Answering with Diverse Answer Types](https://github.com/allenai/gooaq) | [paper](https://arxiv.org/pdf/2104.08727.pdf) | 3,012,496

[Yahoo Answers](https://www.kaggle.com/soumikrakshit/yahoo-answers-dataset) (Title, Answer) | [paper](https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html) | 1,198,260

[Code Search](https://huggingface.co/datasets/code_search_net) | - | 1,151,414

[COCO](https://cocodataset.org/#home) Image captions | [paper](https://link.springer.com/chapter/10.1007%2F978-3-319-10602-1_48) | 828,395

[SPECTER](https://github.com/allenai/specter) citation triplets | [paper](https://doi.org/10.18653/v1/2020.acl-main.207) | 684,100

[Yahoo Answers](https://www.kaggle.com/soumikrakshit/yahoo-answers-dataset) (Question, Answer) | [paper](https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html) | 681,164

[Yahoo Answers](https://www.kaggle.com/soumikrakshit/yahoo-answers-dataset) (Title, Question) | [paper](https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html) | 659,896

[SearchQA](https://huggingface.co/datasets/search_qa) | [paper](https://arxiv.org/abs/1704.05179) | 582,261

[Eli5](https://huggingface.co/datasets/eli5) | [paper](https://doi.org/10.18653/v1/p19-1346) | 325,475

[Flickr 30k](https://shannon.cs.illinois.edu/DenotationGraph/) | [paper](https://transacl.org/ojs/index.php/tacl/article/view/229/33) | 317,695

[Stack Exchange](https://huggingface.co/datasets/flax-sentence-embeddings/stackexchange_xml) Duplicate questions (titles) | | 304,525

AllNLI ([SNLI](https://nlp.stanford.edu/projects/snli/) and [MultiNLI](https://cims.nyu.edu/~sbowman/multinli/) | [paper SNLI](https://doi.org/10.18653/v1/d15-1075), [paper MultiNLI](https://doi.org/10.18653/v1/n18-1101) | 277,230

[Stack Exchange](https://huggingface.co/datasets/flax-sentence-embeddings/stackexchange_xml) Duplicate questions (bodies) | | 250,519

[Stack Exchange](https://huggingface.co/datasets/flax-sentence-embeddings/stackexchange_xml) Duplicate questions (titles+bodies) | | 250,460

[Sentence Compression](https://github.com/google-research-datasets/sentence-compression) | [paper](https://www.aclweb.org/anthology/D13-1155/) | 180,000

[Wikihow](https://github.com/pvl/wikihow_pairs_dataset) | [paper](https://arxiv.org/abs/1810.09305) | 128,542

[Altlex](https://github.com/chridey/altlex/) | [paper](https://aclanthology.org/P16-1135.pdf) | 112,696

[Quora Question Triplets](https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs) | - | 103,663

[Simple Wikipedia](https://cs.pomona.edu/~dkauchak/simplification/) | [paper](https://www.aclweb.org/anthology/P11-2117/) | 102,225

[Natural Questions (NQ)](https://ai.google.com/research/NaturalQuestions) | [paper](https://transacl.org/ojs/index.php/tacl/article/view/1455) | 100,231

[SQuAD2.0](https://rajpurkar.github.io/SQuAD-explorer/) | [paper](https://aclanthology.org/P18-2124.pdf) | 87,599

[TriviaQA](https://huggingface.co/datasets/trivia_qa) | - | 73,346

Total | | **1,124,818,467**

Downloads last month

208,722

Safetensors[](https://huggingface.co/docs/safetensors)

Model size

22.7M params

Tensor type

I64

.

F32

.

Inference Providers [NEW](https://huggingface.co/blog/inference-providers)

[Sentence Similarity](/tasks/sentence-similarity "Learn more about sentence-similarity")

This model is not currently available via any of the supported third-party Inference Providers, and the model is not deployed on the HF Inference API.

System theme

Company

[TOS](/terms-of-service) [Privacy](/privacy) [About](/huggingface) [Jobs](https://apply.workable.com/huggingface/) [](/)

Website

[Models](/models) [Datasets](/datasets) [Spaces](/spaces) [Pricing](/pricing) [Docs](/docs)