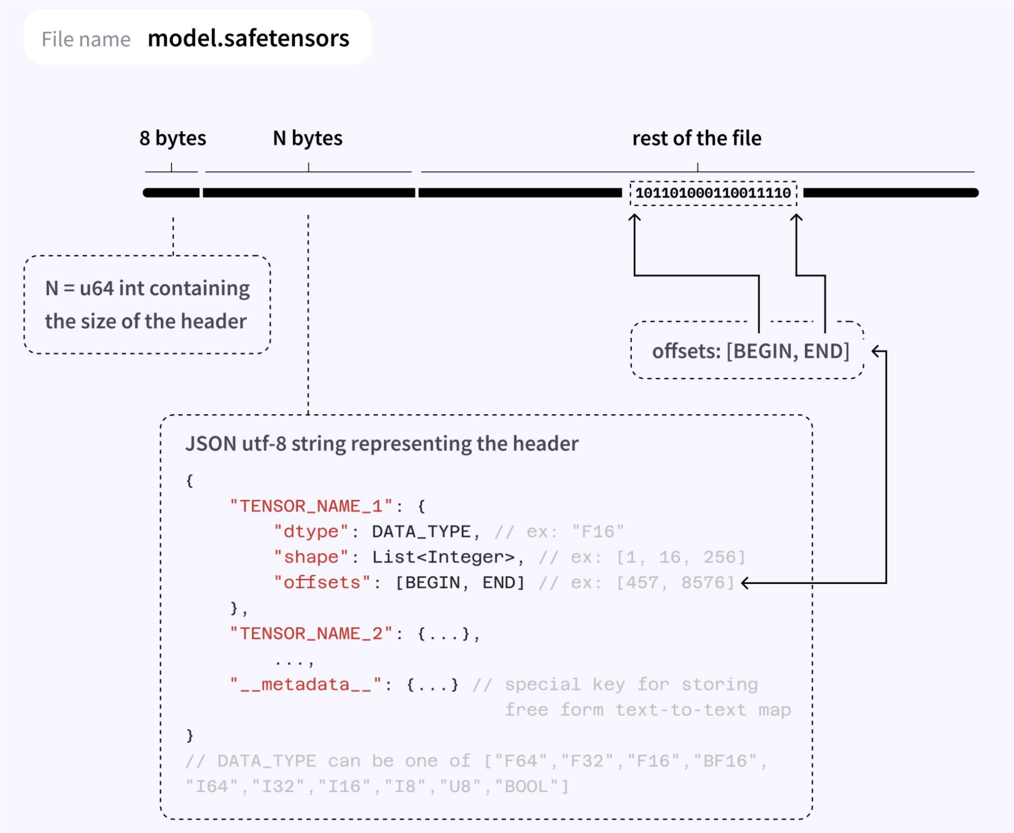
Safetensors

simple format for storing tensors safely (as opposed to pickle)
and that is still fast (zero-copy)

https://huggingface.co/docs/safetensors/index

How it works



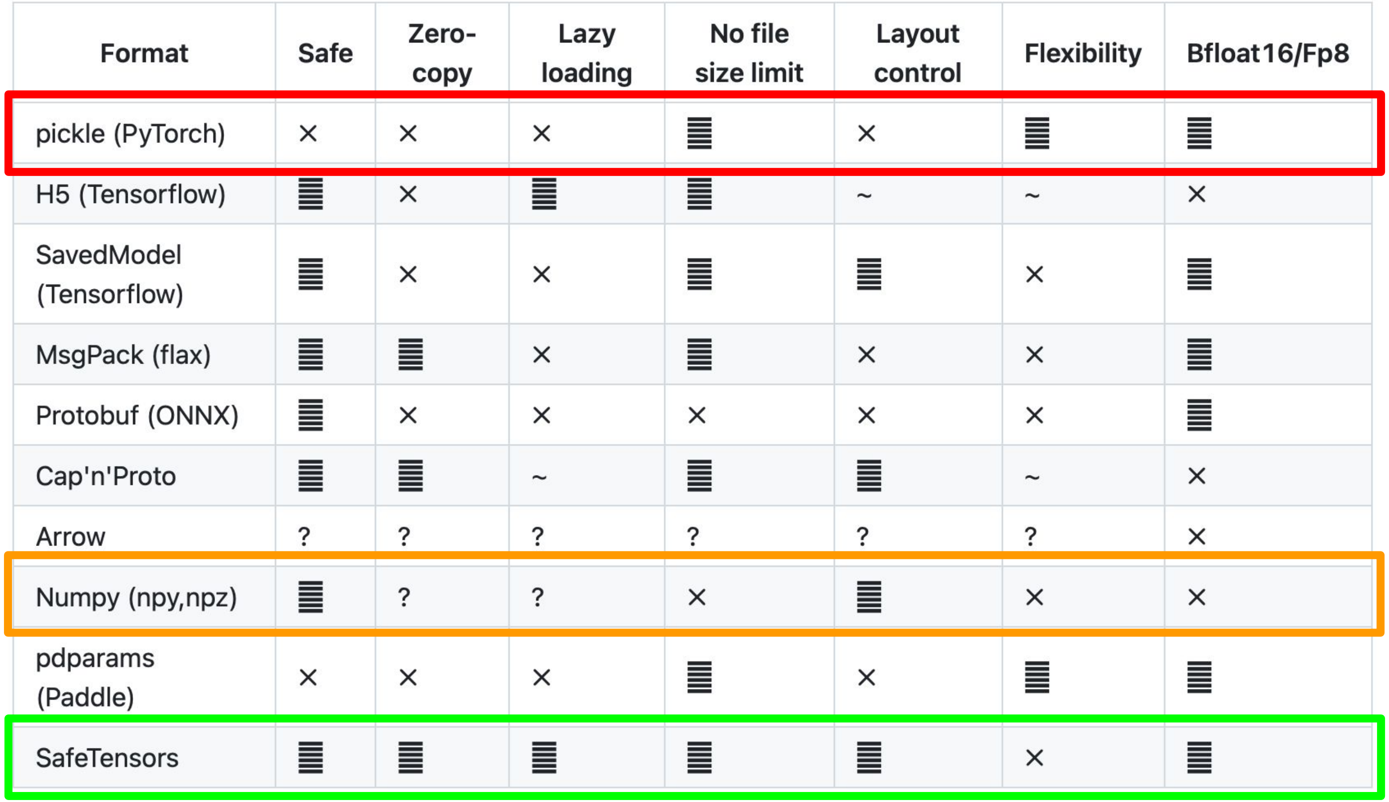
N = u64 int containing the size of the header

offsets: [BEGIN, END]

JSON utf-8 string representing the header

```
{
    "TENSOR_NAME_1": {
        "dtype": DATA_TYPE, // ex: "F16"
        "shape": List<Integer>, // ex: [1, 16, 256]
        "offsets": [BEGIN, END] // ex: [457, 8576]
    },
    "TENSOR_NAME_2": {...},
        ...,
    "__metadata__": {...} // special key for storing
                              free form text-to-text map
}
// DATA_TYPE can be one of ["F64","F32","F16","BF16",
"I64","I32","I16","I8","U8","BOOL"]
```

Comparison with different formats



Format	Safe	Zero-copy	Lazy loading	No file size limit	Layout control	Flexibility	Bfloat16/Fp8
pickle (PyTorch)	×	×	×	✓	×	✓	✓
H5 (Tensorflow)	✓	×	✓	✓	~	~	×
SavedModel (Tensorflow)	✓	×	×	✓	✓	×	✓
MsgPack (flax)	✓	✓	×	✓	×	×	✓
Protobuf (ONNX)	✓	×	×	×	×	×	✓
Cap'n'Proto	✓	✓	~	✓	✓	~	×
Arrow	?	?	?	?	?	?	×
Numpy (npy,npz)	✓	?	?	×	✓	×	×
pdparams (Paddle)	×	×	×	✓	×	✓	✓
SafeTensors	✓	✓	✓	✓	✓	×	✓

unsafe,
runs arbitrary code

no bfloat16 support
vulnerable to zip-bombs
not zero-copy

super cool