Building Machine Learning Models like Open-Source Software

Colin Raffel
From “PaLM: Scaling Language Modeling with Pathways” by Chowdhery et al. and “Scaling Vision Transformers” by Zhai et al.
From “PaLM: Scaling Language Modeling with Pathways” by Chowdhery et al. and “Scaling Vision Transformers” by Zhai et al.
From “PaLM: Scaling Language Modeling with Pathways” by Chowdhery et al. and “Scaling Vision Transformers” by Zhai et al.
From “PaLM: Scaling Language Modeling with Pathways” by Chowdhery et al. and “Scaling Vision Transformers” by Zhai et al.
From “PaLM: Scaling Language Modeling with Pathways” by Chowdhery et al. and “Scaling Vision Transformers” by Zhai et al.
OpenAI technology, just an HTTPS call away

Apply our API to any language task — semantic search, summarization, sentiment analysis, content generation, translation, and more — with only a few examples or by specifying your task in English.
Hugging Face

Search models, datasets, users...

Models 33,490

- **bert-base-uncased**
  - Fill-Mask • Updated May 18 • ↓ 30M • ♥ 54

- **roberta-large**
  - Fill-Mask • Updated May 21 • ↓ 13.1M • ♥ 20

- **distilbert-base-uncased**
  - Fill-Mask • Updated Aug 29 • ↓ 4.83M • ♥ 26

- **xlm-roberta-base**
  - Fill-Mask • Updated Sep 16 • ↓ 4.78M • ♥ 11

- **bert-base-cased**
  - Fill-Mask • Updated Sep 6 • ↓ 4.02M • ♥ 6

- **roberta-base**
  - Fill-Mask • Updated Jul 6 • ↓ 3.45M • ♥ 6

- **gpt2**
  - Text Generation • Updated May 19 • ↓ 3.34M • ♥ 24
How can we enable collaborative and continual development of machine learning models?
How can we enable collaborative and continual development of machine learning models?

Contributors need to be able to cheaply communicate patches to a model.
Fisher-Induced Sparse Unchanging (FISH) Mask
\[(W \times x) \bullet l\]
\((W \circ l)(x)\)
How can we enable collaborative and continual development of machine learning models?

Maintainers need to be able to merge updates from different contributors.
Pre-training  →  Downstream
Pre-training → Donor → Downstream
**Pre-training**  
**Donor**  
**Downstream**

- Pre-trained  
- Fine-tuned  
- Pre-trained  
- Fine-tuned  
- Pre-trained  
- Fine-tuned  
- Donor  
- Intermediate  
- Fine-tuned
\[
\arg \max_\theta \sum_{i=1}^{M} \lambda_i \log p(\theta|\mathcal{D}_i)
\]
arg max_{\theta} \sum_{i=1}^{M} \lambda_i \log p(\theta|D_i)

Log posterior for model i
\[
\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log p(\theta|D_i)
\]

Hyperparameter controlling the importance of model i
\[
\arg\max_{\theta} \sum_{i=1}^{M} \lambda_i \log p(\theta|D_i)
\]
\[
\arg \max_{\theta} \sum_{i=1}^{M} \lambda_i \log \mathcal{N}(\theta | \theta_i, \hat{F}_i^{-1})
\]
\[
\text{arg max}_\theta \sum_{i=1}^{M} \lambda_i \log \mathcal{N}(\theta|\theta_i, \hat{F}_i^{-1})
\]

\[
\theta^*(j) = \frac{\sum_{i=1}^{M} \lambda_i \hat{F}_i(j) \theta_i(j)}{\sum_{i=1}^{M} \lambda_i \hat{F}_i(j)}
\]

Fisher Merging
\[
\arg \max_{\theta} \sum_{i=1}^{M} \lambda_i \log \mathcal{N}(\theta|\theta_i, I)
\]

\[
\theta^*(j) = \frac{\sum_{i=1}^{M} \lambda_i \theta_i^{(j)}}{\sum_{i=1}^{M} \lambda_i}
\]

Isotropic Merging
How can we enable collaborative and continual development of machine learning models?

Users who lack resources need to be able to train and run large models.
Distribute across peers with different compute budgets

- Peer 1: Layer 1, Layer 2, Layer 3
- Peer 2: Layer 4, Layer 5
- Peer 3: Layer 6, Layer 7
- Peer 4: Layer 8, Layer 9, Layer 10, Layer 11, Layer 12
Compress weights
Find the most efficient chain of peers
<table>
<thead>
<tr>
<th>Network</th>
<th>Inference (steps / s)</th>
<th>Parallel (tokens / s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sequence length</td>
<td>Batch size</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>Latency</td>
<td>128</td>
</tr>
<tr>
<td>1 Gbit/s</td>
<td>&lt; 5 ms</td>
<td>1.22</td>
</tr>
<tr>
<td>100 Mbit/s</td>
<td>&lt; 5 ms</td>
<td>1.19</td>
</tr>
<tr>
<td>100 Mbit/s</td>
<td>100 ms</td>
<td>0.89</td>
</tr>
<tr>
<td>3 local servers on 3xA100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 desktops &amp; servers in Europe and North America</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real world</td>
<td></td>
<td>0.63</td>
</tr>
</tbody>
</table>
How can we enable collaborative and continual development of machine learning models?

Maintainers need to be able to quickly **vet** community contributions.
How can we enable collaborative and continual development of machine learning models?

We need to be able to combine modular components to enable new capabilities.
A Call to Build Machine Learning Models like Open-Source Software
Colin Raffel

Training Neural Networks with Fixed Sparse Masks
Yi-Lin Sung*, Varun Nair*, and Colin Raffel

Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning
Haokun Liu*, Derek Tam*, Mohammed Muqeeth*, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel

Merging Models with Fisher-Weighted Averaging
Michael Matena and Colin Raffel

PETALS: Collaborative Inference of Large Models

Please give me feedback: