BigMAC: Big Model Adaptation for Computer vision

Speakers:
- Neil Houlsby, Google Brain
- Maria Attarian, Google Brain, U. of Toronto
- Ludwig Schmidt, U. of Washington
- Ishan Misra, Meta AI
- Aditi Raghunathan, Carnegie Mellon University
- Sayak Paul, HuggingFace
- Carl Vondrick, Columbia University

Organised by:
- Yuki M. Asano
- Tengda Han
- Mathilde Caron
- Phillip Isola
- Serge Belongie
<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 am - 9:15 am</td>
<td>Welcome and Introduction</td>
<td></td>
</tr>
<tr>
<td>9:15 am - 9:45 am</td>
<td>Neil Houlsby</td>
<td>Google Brain</td>
</tr>
<tr>
<td>9:45 am - 10:15 am</td>
<td>Maria Attarian</td>
<td>Google Brain, University of Toronto</td>
</tr>
<tr>
<td>10:15 am - 10:45 am</td>
<td>Ludwig Schmidt</td>
<td>University of Washington</td>
</tr>
<tr>
<td>10:45 am - 11:00 am</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>10:00 am - 11:30 am</td>
<td>Ishan Misra</td>
<td>Meta AI</td>
</tr>
<tr>
<td>11:30 am - 12:00 pm</td>
<td>Aditi Raghunathan</td>
<td>Carnegie Mellon University</td>
</tr>
<tr>
<td>12:00 pm - 12:30 pm</td>
<td>Sayak Paul</td>
<td>HuggingFace</td>
</tr>
<tr>
<td>12:30 pm - 1:00 pm</td>
<td>Carl Vondrick</td>
<td>Columbia University</td>
</tr>
<tr>
<td>1:00 pm</td>
<td>Closing remarks</td>
<td></td>
</tr>
</tbody>
</table>
Why now?
Big Models have arrived. So we need to figure out how to use them.

- tremendous growth in LLM sizes
- Multiple growth trends

Sevilla et al. Compute Trends Across Three Eras of Machine Learning, 2022
That's just NLP... or is it? No.

With Transformers in vision, scale has arrived here too.

Sevilla et al. Compute Trends Across Three Eras of Machine Learning. 2022
Dehghani et al. Scaling Vision Transformers to 22 Billion Parameters. ICML 2023
Woo et al. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. CVPR 2023

"Compute Requirements: ViT-22B was trained on 1024 TPU V4 chips [...]"
Visual Language Models further increases the #params, by a lot

Visual Language Models further increases the #params, by a lot

Vision systems' getting bigger

- Flamingo, BLIP, CM3, Frozen, CoCa, ALIGN, Fromage, VisualLLM, ...

Yu et al. Contrastive Captioners are Image-Text Foundation Models. TMLR 2022
Koh et al. Grounding Language Models to Images for Multimodal Generation. 2023
Li et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023
Aghajanyan et al. CM3: A Causal Masked Multimodal Model of the Internet. 2022
The base/foundation vs the adaptation. (Too a large extent) a well known recipe.
What to do with those big models?
Main ways of adapting models (1/2)

**Full-finetuning**

![Diagram showing full-finetuning process]

- Model \( \rightarrow \mathcal{L} \rightarrow \text{target labels} \)

**Limited-finetuning (e.g. linear probing)**

![Diagram showing limited-finetuning process]

- Model \( \rightarrow \mathcal{L} \rightarrow \text{target labels} \)

**No-finetuning (e.g. used for retrieving similar instances)**

![Diagram showing no-finetuning process]

- Model \( \rightarrow \text{embedding}_1 \rightarrow \ldots \rightarrow \text{embedding}_n \rightarrow \text{e.g. retrieval, clustering} \)
Parameter-efficient Finetuning (PEFT) ideas

More params inside model: Adapters, LoRA etc.
- e.g.: 1x1 convs, Residual-MLPs, only BN or bias params, binary masks, low-rank adaptation of matmuls

Prompt/prefix learning
- similar to prompt manual engineering [like "step-by-step" or "trending on artstation"]

https://github.com/adapter-hub/adapter-transformers
https://github.com/huggingface/peft
## Schedule

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 am - 9:15 am</td>
<td>Welcome and Introduction</td>
<td></td>
</tr>
<tr>
<td>9:15 am - 9:45 am</td>
<td>Neil Houlsby</td>
<td>Google Brain</td>
</tr>
<tr>
<td>9:45 am - 10:15 am</td>
<td>Maria Attarian</td>
<td>Google Brain, University of Toronto</td>
</tr>
<tr>
<td>10:15 am - 10:45 am</td>
<td>Ludwig Schmidt</td>
<td>University of Washington</td>
</tr>
<tr>
<td>10:45 am - 11:00 am</td>
<td>Coffee Break</td>
<td></td>
</tr>
<tr>
<td>10:00 am - 11:30 am</td>
<td>Ishan Misra</td>
<td>Meta AI</td>
</tr>
<tr>
<td>11:30 am - 12:00 pm</td>
<td>Aditi Raghunathan</td>
<td>Carnegie Mellon University</td>
</tr>
<tr>
<td>12:00 pm - 12:30 pm</td>
<td>Sayak Paul</td>
<td>HuggingFace</td>
</tr>
<tr>
<td>12:30 pm - 1:00 pm</td>
<td>Carl Vondrick</td>
<td>Columbia University</td>
</tr>
<tr>
<td>1:00 pm</td>
<td>Closing remarks</td>
<td></td>
</tr>
</tbody>
</table>