Flamingo: A Visual Language Model for Few-Shot Learning

**Presenter**: Turcan Tuna

**Mentor**: Ferjad Naeem

18.04.2023
Premise - Large Language Models

- What is a Large Language Model? (LLM)

Premise - Large Language Models

ChatGPT

Examples

- "Explain quantum computing in simple terms."
- "Get any creative ideas for a 10 year old's birthday!"
- "How do I make an HTTP request in JavaScript?"

Capabilities

- Remembers what user said earlier in the conversation
- Allows users to provide follow-up questions
- Trained to decline inappropriate requests

Limitations

- May occasionally generate incorrect information
- May occasionally produce harmful instructions or biased content
- Limited knowledge of world and events after 2021
Premise - Flamingo

- What is Flamingo\textsuperscript{[2]}?

> “...a visually-conditioned autoregressive text generation model able to ingest a sequence of text tokens interleaved with images and/or videos, and produce text as output”.

\[\text{Visual Features} \xrightarrow{} \text{A language model} \xrightarrow{} \text{Multi-Modal prompt}\]

Premise - Visually Conditioned Large Language Models

- How LLMs can gain the **ability to see**?

Visual Text Prompt: “What is this image?”

Tokenize → Visual Encoder → Visual Features → Large language Model

Compute

“*A picture of super excited students.*“
Premise - Example

Visual Text Prompt

“Describe me this image.”

*Visually Conditioned LLM (Flamingo)

“Students in the lecture hall”

* Deployed the re-implementation from https://github.com/dhansmair/flamingo-mini.
"Describe me this image."

Blip-2\textsuperscript{[3]}, 2023, SOTA.

“A group of people sitting at desks in a classroom”

Premise - Advanced Example

Lets have some fun!, Visual - Question Answering (VQA).

"Describe me this image."

"People in a room"

"How many people?"

"Too many"
Motivation

- LLMs are **very expensive** to train.
- No intuitive expansion support.
  - How to not forget?

- It **is** expensive to train. Really..

  “....training BERT on GPU is roughly equivalent to a trans-American flight.”[^4]

<table>
<thead>
<tr>
<th>Model</th>
<th>Hardware</th>
<th>Power (W)</th>
<th>Hours</th>
<th>kWh·PUE</th>
<th>Cloud compute cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer_base</td>
<td>P100x8</td>
<td>1415.78</td>
<td>12</td>
<td>27</td>
<td>$41–$140</td>
</tr>
<tr>
<td>Transformer_big</td>
<td>P100x8</td>
<td>1515.43</td>
<td>84</td>
<td>201</td>
<td>$289–$981</td>
</tr>
<tr>
<td>ELMo</td>
<td>P100x3</td>
<td>517.66</td>
<td>336</td>
<td>275</td>
<td>$433–$1472</td>
</tr>
<tr>
<td>BERT_base</td>
<td>V100x64</td>
<td>12,041.51</td>
<td>79</td>
<td>1507</td>
<td>$3751–$12,571</td>
</tr>
<tr>
<td>BERT_base</td>
<td>TPUv2x16</td>
<td>—</td>
<td>96</td>
<td>—</td>
<td>$2074–$6912</td>
</tr>
<tr>
<td>NAS</td>
<td>P100x8</td>
<td>1515.43</td>
<td>274,120</td>
<td>656,347</td>
<td>$942,973–$3,201,722</td>
</tr>
<tr>
<td>NAS</td>
<td>TPUv2x1</td>
<td>—</td>
<td>32,623</td>
<td>—</td>
<td>$44,055–$146,848</td>
</tr>
<tr>
<td>GPT-2</td>
<td>TPUv3x32</td>
<td>—</td>
<td>168</td>
<td>—</td>
<td>$12,902–$43,008</td>
</tr>
</tbody>
</table>

Motivation

- LLMs require huge amount of data.
  - This data can be found as text (*MassiveText*[^5])! But not for images.
  - 1.4 Trillion vs ~ 1.8 billion.

- ALIGN[^6] dataset **1.8 billion images** paired with text but *noisy*!

- A gap in the literature exists for **videos**!

Contributions

- A way to combined interleaved images and text
  - Gated cross-attention module

- A unique perceiver architecture with fixed output.
  - Perceiver sampler

- Evaluation and ablation
Related Work

CLIP[7]
- Close-ended task superiority.
- Trained from scratch.
- Very good zero-shot, few-shot performance.

CM3[8]
- Big model 24 days on 384 A100s
- Multimodal & Unimodal tasks.
- Directly works on HTML.
- Can output images.

VLKD[9]
- Visual-Language Knowledge Dist.
- Efficient and compact.
- Similar results to Flamingo

- Image-conditioned prompt learning.
- A good few-shot learner.
- Goal: generalizability

Method - Introduction

● Goal: Provide LLMs ability to see.
  ○ Convert LLM to VLM.

● Key ideas:
  ○ Extend a frozen Pre-trained Language Model.
  ○ Reducing visual input to a fixed number of tokens with Perceiver Sampler.
  ○ Cross attention layers to visually condition LLM.
  ○ Training on a different types of data.
Method - Introduction

Flamingo

Pretrained and frozen
Trained from scratch

Perceiver Resampler
Vision Encoder

Output: text
da very serious cat.

Vision Encoder

Perceiver Resampler

Processed text
This is a very cute dog. <image>
This is a very cute dog.

Interleaved visual/text data

Image

Vision Encoder

Perceiver Sampler

VLM

Text Tokens
Method - Introduction

- Multiple Models are Recycled.

![Diagram of NFNet-F6 with BERT](image)

- NFNet-F6\textsuperscript{[11]} 435M Param.
- + BERT\textsuperscript{[12]}

- Vision Encoder
- Perceiver Sampler
- ~Perceiver\textsuperscript{[13]} architecture

- Chinchilla\textsuperscript{[1]} 70B Param.


Method

Image → Visual Encoder → Perceiver Sampler → Text Tokens
Method - Visual Encoder

- VLM needs **text-conditioned** Visual tokens.

![Diagram of Visual Encoder](image)

**Visual Encoder (NFNet-F6)**

- Image
- Visual Tokens

$X_t$

- Time
- Vision Encoder
- Vision Encoder
- Vision Encoder

Image

Perceiver Sampler

Text Tokens
Method - Visual Encoder

- The NFNet-F6 architecture is from Normalizer-Free ResNet.
Method

Image → Visual Encoder → Perceiver Sampler → VLM → Text Tokens
Method - Perceived Sampler

- Latent queries:
  - Most important part of the data.
  - Model learns what to extract.

- No explicit spatial grid position encodings.

- Nb of *Output tokens* = Nb of *latent queries*. 

![Diagram of Perceived Sampler](image-url)
Method
Method - Gated Cross attention Layers

- **Gated cross-attention blocks**
  - Text conditioning on visual representations.

- Integrate new skills to LLMs without forgetting.
  - **Tanh gating.**
Method - Gated Cross attention Layers

(a) Attention tanh gating
(b) FFW tanh gating.

Image Encoder → Perceiver → Text Tokens
Method - Interleaved Visual / Text data support

- How to link **interleaved** multi-modal prompts?

\[ \phi : [1, L] \mapsto [0, N] \]
Method
Method - Training Flamingo Models

- Pre-processing & augmentation
  - Random flips, Increased resolution, Color augmentation
  - **Random text links 50% probability.**
  - 8 frames are sampled from training videos.

“*A cool cat next to a building*”
Method - Training Flamingo Models

- A big chunk is coming from **frozen** LLM. (Chinchilla\textsuperscript{[1]})
- Vision encoder (NFNet-F6) and Perceiver Resampler are same for all.

<table>
<thead>
<tr>
<th></th>
<th>Requires model sharding</th>
<th>Frozen Language</th>
<th>Frozen Vision</th>
<th>Trainable GATED XATTN-DENSE</th>
<th>Total count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flamingo-3B</td>
<td>×</td>
<td>1.4B</td>
<td>435M</td>
<td>1.2B (every)</td>
<td>3.2B</td>
</tr>
<tr>
<td>Flamingo-9B</td>
<td>×</td>
<td>7.1B</td>
<td>435M</td>
<td>1.6B (every 4th)</td>
<td>9.3B</td>
</tr>
<tr>
<td>Flamingo</td>
<td>✓</td>
<td>70B</td>
<td>435M</td>
<td>10B (every 7th)</td>
<td>80B</td>
</tr>
</tbody>
</table>
Method - Training Datasets

● Different data types

<table>
<thead>
<tr>
<th>Dataset</th>
<th>What Data?</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long Text &amp; Image Pairs (in-house)</td>
<td>312 million Image and text pairs</td>
<td>better quality and longer descriptions</td>
</tr>
<tr>
<td>MultiModal MassiveWeb(M3W)</td>
<td>Massive Web dataset, multimodal</td>
<td>MASSIVE, some ambiguous links</td>
</tr>
<tr>
<td>Video &amp; Text Pairs (VTP)</td>
<td>22 million short videos with paired text</td>
<td>On average 22 seconds</td>
</tr>
</tbody>
</table>
Method - Data Deduplication

- LTIP and ALIGN are deduplicated (M3W and VTP Not!).

<table>
<thead>
<tr>
<th>Datasets (EVALUATION)</th>
<th>Is deduplicated against?</th>
<th>Used for?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>✓</td>
<td>(train, valid)</td>
</tr>
<tr>
<td>COCO</td>
<td>✓</td>
<td>(train, valid, test)</td>
</tr>
<tr>
<td>OK-VQA</td>
<td>✓</td>
<td>(train, valid, test)</td>
</tr>
<tr>
<td>VQAv2</td>
<td>✓</td>
<td>(train, valid, test)</td>
</tr>
<tr>
<td>Flickr30k</td>
<td>✓</td>
<td>(valid, test)</td>
</tr>
<tr>
<td>VisDial</td>
<td>✓</td>
<td>(valid, test)</td>
</tr>
<tr>
<td>VizWiz</td>
<td>✗</td>
<td>(test)</td>
</tr>
<tr>
<td>HatefulMemes</td>
<td>✗</td>
<td>(test)</td>
</tr>
<tr>
<td>TextVQA</td>
<td>✗</td>
<td>(test)</td>
</tr>
</tbody>
</table>
Method - Training Flamingo Models

- Possible input combinations:

  - Image-Text Pairs dataset:
    - This is an image of a flamingo.

  - Video-Text Pairs dataset:
    - A kid doing a kickflip.

  - Multi-Modal Massive Web (M3W) dataset:
    - Welcome to my website!
    - This is a picture of my dog.
    - This is a picture of my cat.
Method - Training Flamingo Models

- **Optimizer**: AdamW\(^{[14]}\)
- **Linear-warm up** followed by a constant learning rate.
- **Dataset mixture weights**: 

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>M3W</td>
<td>1.0</td>
</tr>
<tr>
<td>LTIP</td>
<td>0.2</td>
</tr>
<tr>
<td>VTP</td>
<td>0.03</td>
</tr>
<tr>
<td>ALIGN</td>
<td>0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight Decay(^*)</td>
<td>0.1</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>10^-4</td>
</tr>
</tbody>
</table>

\(^*\): No weight decay for Perceiver Resampler

---

Method - Training Flamingo Models

- Loss function
  - Weighted sum of per-dataset expected negative log-likelihoods of text, given the visual inputs.

\[
\sum_{m=1}^{M} \lambda_m \cdot \mathbb{E}_{(x,y) \sim D_m} \left[ - \sum_{\ell=1}^{L} \log p(y_{\ell} | y_{<\ell}, x_{\leq \ell}) \right]
\]

The weights we discussed
Method - Ready for Evaluation!
Evaluation - Established Methods

- **Close-Ended Tasks**: Response from a **pre-defined** space.
  - Text after the query image is used.
  - Final Selection: Beam search.
  - ex. Classification
    - Zero-shot
    - Few-shot

- **Open-Ended Tasks**: **Without** a pre-defined response space.
  - Final Selection: log-likelihood
  - ex. VQA, Open-ended dialog.
  - Zero or Few-Shot Generalization, new task learning
Evaluation - Few-shot Classification

- RICES (Retrieval In-Context Example Selection)
- Flamingo is **not trained with** a contrastive loss.
  - Requires **well distributed training data**.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Prompt size</th>
<th>shots/class</th>
<th>ImageNet top 1</th>
<th>Kinetics/OU avg top 1/5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SotA</td>
<td>Fine-tuned</td>
<td>-</td>
<td>full</td>
<td>91.0</td>
<td>89.0</td>
</tr>
<tr>
<td>SotA</td>
<td>Contrastive</td>
<td>-</td>
<td>0</td>
<td>85.7</td>
<td>69.6</td>
</tr>
<tr>
<td>NFNNetF6</td>
<td>Our contrastive</td>
<td>-</td>
<td>0</td>
<td>77.9</td>
<td>62.9</td>
</tr>
<tr>
<td>Flamingo-3B</td>
<td>RICES</td>
<td>8</td>
<td>1</td>
<td>70.9</td>
<td>55.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>1</td>
<td>71.0</td>
<td>56.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>5</td>
<td>72.7</td>
<td>58.3</td>
</tr>
<tr>
<td>Flamingo-9B</td>
<td>RICES</td>
<td>8</td>
<td>1</td>
<td>71.2</td>
<td>58.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>1</td>
<td>71.7</td>
<td>59.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>5</td>
<td>75.2</td>
<td>60.9</td>
</tr>
<tr>
<td>Random</td>
<td></td>
<td>16</td>
<td>≤ 0.02</td>
<td>66.4</td>
<td>51.2</td>
</tr>
<tr>
<td>Flamingo-80B</td>
<td>RICES</td>
<td>8</td>
<td>1</td>
<td>71.9</td>
<td>60.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>1</td>
<td>71.7</td>
<td>62.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>5</td>
<td>76.0</td>
<td>63.5</td>
</tr>
<tr>
<td>RICES+ensembling</td>
<td></td>
<td>16</td>
<td>5</td>
<td>77.3</td>
<td>64.2</td>
</tr>
</tbody>
</table>
Evaluation

- Superior in 6 tasks
  - Zero-shot, Few-shot image understanding.
Evaluation - Open Dialog

This is an apple with a sticker on it.

What does the sticker say?

The sticker says "iPod".

Where is the photo taken?

It looks like it's taken in a backyard.

Do you think it is printed or handwritten?

It looks like it's handwritten.

What color is the sticker?

It's white.
Text Completion and VQA Prompts

Question: What is odd about this image? Explain why it is unusual. Answer:

The image is odd because the elephant is in the back of the truck. It is unusual because elephants are not usually transported in the back of a truck.

The ingredients of this dish are: beef, potatoes, carrots, broccoli, and lemon.
Evaluation - Few Shot Generalization

- Examples are highly important.
  - Known to have example bias: **the last example has to be good** and relevant.

**Vision to Text tasks** (input=vision, output=text)

Support examples

- A cat wearing sunglasses.
- Elephants walking in the savanna.

Query

Processed prompt

**Visual Question Answering Task** (input=vision+text, output=text)

Support examples

- What's the cat wearing? sunglasses
- How many animals? 3

Query

Processed prompt

**Question:** What's the cat wearing? **Answer:** sunglasses

**Question:** How many animals? **Answer:** 3

**Question:** What is on the water? **Answer:**
Evaluation - Few Shot Generalization

- Outperforms **pre-trained** SOTA in some cases.
- \(\uparrow\) Shot, \(\uparrow\) Performance

<table>
<thead>
<tr>
<th>Method</th>
<th>FT Shot</th>
<th>OKVQA (%)</th>
<th>NQA-v2 (%)</th>
<th>COCO (%)</th>
<th>M3VQA (%)</th>
<th>VATEX (%)</th>
<th>VqA-v1 (%)</th>
<th>Flick3K (%)</th>
<th>MSVRCVQA (%)</th>
<th>TVQA (%)</th>
<th>VqAvqa/2 (%)</th>
<th>STAR (%)</th>
<th>VqA-bul (%)</th>
<th>TestVQA (%)</th>
<th>NestQ (%)</th>
<th>HateLibMemes-100 (%)</th>
<th>RareAct (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero/Few shot SOTA</td>
<td>X</td>
<td>43.3 (16)</td>
<td>38.2 (4)</td>
<td>32.2 (0)</td>
<td>35.2 (0)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>58 (0)</td>
<td>19.2 (0)</td>
<td>-</td>
<td>19.2 (0)</td>
<td>12.2 (0)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>66.1 (0)</td>
</tr>
<tr>
<td>Flamingo-3B</td>
<td>X</td>
<td>40.7 (0)</td>
<td>49.2 (2)</td>
<td>73.0 (0)</td>
<td>27.5 (0)</td>
<td>40.1 (0)</td>
<td>28.9 (0)</td>
<td>60.6 (0)</td>
<td>11.0 (0)</td>
<td>32.7 (0)</td>
<td>55.8 (0)</td>
<td>39.6 (0)</td>
<td>46.1 (0)</td>
<td>30.1 (0)</td>
<td>21.3 (0)</td>
<td>53.7 (0)</td>
<td>58.4 (0)</td>
</tr>
<tr>
<td>Flamingo-9B</td>
<td>X</td>
<td>40.7 (0)</td>
<td>49.2 (2)</td>
<td>73.0 (0)</td>
<td>27.5 (0)</td>
<td>40.1 (0)</td>
<td>28.9 (0)</td>
<td>60.6 (0)</td>
<td>11.0 (0)</td>
<td>32.7 (0)</td>
<td>55.8 (0)</td>
<td>39.6 (0)</td>
<td>46.1 (0)</td>
<td>30.1 (0)</td>
<td>21.3 (0)</td>
<td>53.7 (0)</td>
<td>58.4 (0)</td>
</tr>
<tr>
<td>Flamingo</td>
<td>X</td>
<td>40.7 (0)</td>
<td>49.2 (2)</td>
<td>73.0 (0)</td>
<td>27.5 (0)</td>
<td>40.1 (0)</td>
<td>28.9 (0)</td>
<td>60.6 (0)</td>
<td>11.0 (0)</td>
<td>32.7 (0)</td>
<td>55.8 (0)</td>
<td>39.6 (0)</td>
<td>46.1 (0)</td>
<td>30.1 (0)</td>
<td>21.3 (0)</td>
<td>53.7 (0)</td>
<td>58.4 (0)</td>
</tr>
<tr>
<td>Pretrained FT SOTA</td>
<td>✔</td>
<td>54.4 (10K)</td>
<td>80.2 (444K)</td>
<td>143.3 (500K)</td>
<td>47.9 (27K)</td>
<td>76.3 (500K)</td>
<td>57.2 (20K)</td>
<td>67.4 (500K)</td>
<td>46.8 (20K)</td>
<td>35.4 (133K)</td>
<td>138.7 (20K)</td>
<td>36.7 (133K)</td>
<td>75.2 (20K)</td>
<td>54.7 (133K)</td>
<td>25.2 (20K)</td>
<td>79.1 (133K)</td>
<td></td>
</tr>
</tbody>
</table>
Evaluation - Zero Shot Generalization

- Prompt engineering is important.
  - The way you present the question matters.

<table>
<thead>
<tr>
<th></th>
<th>Flickr30K image-to-text</th>
<th>Flickr30K text-to-image</th>
<th>COCO image-to-text</th>
<th>COCO text-to-image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1  R@5  R@10</td>
<td>R@1  R@5  R@10</td>
<td>R@1  R@5  R@10</td>
<td>R@1  R@5  R@10</td>
</tr>
<tr>
<td>Florence</td>
<td>90.9  99.1  -</td>
<td>76.7  93.6  -</td>
<td>64.7  85.9  -</td>
<td>47.2  71.4  -</td>
</tr>
<tr>
<td>ALIGN</td>
<td>88.6  98.7  99.7</td>
<td>75.7  93.8  96.8</td>
<td>58.6  83.0  89.7</td>
<td>45.6  69.8  78.6</td>
</tr>
<tr>
<td>CLIP</td>
<td>88.0  98.7  99.4</td>
<td>68.7  90.6  95.2</td>
<td>58.4  81.5  88.1</td>
<td>37.7  62.4  72.2</td>
</tr>
<tr>
<td>Flamingo</td>
<td>89.3  98.8  99.7</td>
<td>79.5  95.3  97.9</td>
<td>65.9  87.3  92.9</td>
<td>48.0  73.3  82.1</td>
</tr>
</tbody>
</table>
Evaluation - Video / Text Input

- Videos are **sequence of single images**.
Ablation Studies

- Ablation study on Flamingo-3B Model.

<table>
<thead>
<tr>
<th>Ablated setting</th>
<th>Flamingo-3B original value</th>
<th>Changed value</th>
<th>Param. count ↓</th>
<th>Step time ↓</th>
<th>COCO CIDEr↑</th>
<th>OKVQA top1↑</th>
<th>VQA v2 top1↑</th>
<th>MSVDQA top1↑</th>
<th>VATEX CIDEr↑</th>
<th>Overall score↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flamingo-3B model</td>
<td>3.2B 1.74s</td>
<td>86.5 42.1 55.8 36.3 53.4</td>
<td>70.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(i) Training data</td>
<td>All data</td>
<td>w/o Video-Text pairs</td>
<td>3.2B 1.42s</td>
<td>84.2 43.0 53.9 34.5 46.0</td>
<td>67.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>w/o Image-Text pairs</td>
<td>3.2B 0.95s</td>
<td>66.3 39.2 51.6 32.0 41.6</td>
<td>69.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Image-Text pairs → LAION</td>
<td>3.2B 1.74s</td>
<td>79.5 41.4 53.5 33.9 47.6</td>
<td>66.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>w/o M3W</td>
<td>3.2B 1.02s</td>
<td>54.1 36.5 52.7 31.4 23.5</td>
<td>53.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ii) Optimisation</td>
<td>Accumulation</td>
<td>Round Robin</td>
<td>3.2B 1.68s</td>
<td>76.1 39.8 52.1 33.2 40.8</td>
<td>62.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(iii) Tanh gating</td>
<td>3.2B 1.74s</td>
<td>78.4 40.5 52.9 35.9 47.5</td>
<td>66.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(iv) Cross-attention architecture</td>
<td>GATED XATTN-DENSE</td>
<td>VANILLA XATTN GRAFTING</td>
<td>2.4B 1.16s</td>
<td>80.6 41.5 53.4 32.9 50.7</td>
<td>66.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.3B 1.74s</td>
<td>79.2 36.1 50.8 32.2 47.8</td>
<td>63.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(v) Cross-attention frequency</td>
<td>Every</td>
<td>Single in middle</td>
<td>2.0B 0.87s</td>
<td>71.5 38.1 50.2 29.1 42.3</td>
<td>59.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Every 4th</td>
<td>2.3B 1.02s</td>
<td>82.3 42.7 55.1 34.6 50.8</td>
<td>68.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Every 2nd</td>
<td>2.6B 1.24s</td>
<td>83.7 41.0 55.8 34.5 49.7</td>
<td>68.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(vi) Resampler</td>
<td>Perceiver</td>
<td>MLP Transformer</td>
<td>3.2B 1.85s</td>
<td>78.6 42.2 54.7 35.2 44.7</td>
<td>66.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.2B 1.81s</td>
<td>83.2 41.7 55.6 31.5 48.3</td>
<td>66.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(vii) Vision encoder</td>
<td>NFNNet-F6</td>
<td>CLIP ViT-L/14 NFNNet-F0</td>
<td>3.1B 1.58s</td>
<td>76.5 41.6 53.4 33.2 44.5</td>
<td>64.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.9B 1.45s</td>
<td>73.8 40.5 52.8 31.1 42.9</td>
<td>62.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(viii) Freezing LM</td>
<td>✓</td>
<td>(random init)</td>
<td>3.2B 2.42s</td>
<td>74.8 31.5 45.6 26.9 50.1</td>
<td>57.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(pretrained)</td>
<td>3.2B 2.42s</td>
<td>81.2 33.7 47.4 31.0 53.9</td>
<td>62.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Ablation - Dataset Combining Strategy

- **Data merged**: Merging examples from each dataset.
- **Round-robin**\(^{[15]}\): Alternate examples from each dataset.
- **Accumulation**: The gradients from each dataset are weighted and summed.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Combination strategy</th>
<th>ImageNet accuracy top-1</th>
<th>COCO image-to-text</th>
<th>COCO text-to-image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>LTIP</td>
<td>None</td>
<td>40.8</td>
<td>38.6</td>
<td>66.4</td>
</tr>
<tr>
<td>ALIGN</td>
<td>None</td>
<td>35.2</td>
<td>32.2</td>
<td>58.9</td>
</tr>
<tr>
<td>LTIP + ALIGN</td>
<td>Accumulation</td>
<td>45.6</td>
<td>42.3</td>
<td>68.3</td>
</tr>
<tr>
<td>LTIP + ALIGN</td>
<td>Data merged</td>
<td>38.6</td>
<td>36.9</td>
<td>65.8</td>
</tr>
<tr>
<td>LTIP + ALIGN</td>
<td>Round-robin</td>
<td>41.2</td>
<td>40.1</td>
<td>66.7</td>
</tr>
</tbody>
</table>

Bigger but more noisy.

5x Smaller, higher in quality.

### Ablation - Additional

- Additional studies

<table>
<thead>
<tr>
<th>Ablated setting</th>
<th>Flamingo 3B value</th>
<th>Changed value</th>
<th>Param. count ↓</th>
<th>Step time ↓</th>
<th>COCO CIDEr↑</th>
<th>OKVQA topp1↑</th>
<th>VQA v2 topp1↑</th>
<th>MSVDQA topp1↑</th>
<th>VATEX CIDEr↑</th>
<th>Overall score↑</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Flamingo 3B model (short training)</strong></td>
<td></td>
<td></td>
<td>3.2B</td>
<td>1.74s</td>
<td>86.5</td>
<td>42.1</td>
<td>55.8</td>
<td>36.3</td>
<td>53.4</td>
<td>70.7</td>
</tr>
<tr>
<td>(i) Resampler size</td>
<td>Medium</td>
<td>Small</td>
<td>Large</td>
<td>3.1B</td>
<td>1.58s</td>
<td>81.1</td>
<td>40.4</td>
<td>54.1</td>
<td>36.0</td>
<td>50.2</td>
</tr>
<tr>
<td></td>
<td>Only</td>
<td>Last</td>
<td>All previous</td>
<td>3.2B</td>
<td>1.74s</td>
<td>70.0</td>
<td>40.9</td>
<td>52.0</td>
<td>32.1</td>
<td>46.8</td>
</tr>
<tr>
<td>(ii) Multi-img att</td>
<td>0.5</td>
<td>0.0</td>
<td>1.0</td>
<td>3.2B</td>
<td>1.74s</td>
<td>85.0</td>
<td>41.6</td>
<td>55.2</td>
<td>36.7</td>
<td>50.6</td>
</tr>
<tr>
<td>(iii) $p_{next}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(iv) *LM pretraining</td>
<td>MassiveText</td>
<td>C4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(v) Freezing Vision</td>
<td>✓</td>
<td>✓ (random init)</td>
<td>✓ (pretrained)</td>
<td>3.2B</td>
<td>4.70s*</td>
<td>74.5</td>
<td>41.6</td>
<td>52.7</td>
<td>31.4</td>
<td>35.8</td>
</tr>
<tr>
<td>(vi) Co-train LM on MassiveText</td>
<td>✓</td>
<td>✓ (random init)</td>
<td>✓ (pretrained)</td>
<td>3.2B</td>
<td>5.34s*</td>
<td>69.3</td>
<td>29.9</td>
<td>46.1</td>
<td>28.1</td>
<td>45.5</td>
</tr>
<tr>
<td>(vii) Dataset and Vision encoder</td>
<td>M3W+ITP+VTP and NFNetF6</td>
<td>LAION400M and CLIP</td>
<td>M3W+LAION400M+VTP and CLIP</td>
<td>3.1B</td>
<td>0.86s</td>
<td>61.4</td>
<td>37.9</td>
<td>50.9</td>
<td>27.9</td>
<td>29.7</td>
</tr>
</tbody>
</table>
Discussion - Limitations

- Inherits **weaknesses** of LLMs.
  - Hallucinations and ungrounded guesses.
  - Fixed number of tokens.
  - Bad sample efficiency.
Discussion - Limitations

- Limited Visual and language interface
  - No visual context about the output prompt.

“How many giraffe?”

“How 3”
Conclusion

- A framework on how to extend LLMs -> VLMs.
  - Cross-attention allows for VLM extension!

- Perceiver Sampler based fixed tokens successful!.
  - Video input enabled!

- Data size matters.
  - Data quality matters more.

- Could perform better than Fine Tuned SOTA!

- Direct inheritance of bad habits of LLMs
  - Racism
  - Hallucinations
Thank you for listening!

“How many Flamingos?”

hmm

“How many?”

“25”
References


Appendix - Compute

- A big chunk is coming from LLM.
- Vision encoder is same for all. (NFNet-F6)

<table>
<thead>
<tr>
<th></th>
<th>Perceiver Resampler</th>
<th>GATED XATTN-DENSE</th>
<th>Frozen LM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L \ D \ H \ Act.</td>
<td>L \ D \ H \ Act.</td>
<td>L \ D \ H \ Act.</td>
</tr>
<tr>
<td>Flamingo-3B</td>
<td>6 \ 1536 \ 16 \ Sq. ReLU</td>
<td>24 \ 2048 \ 16 \ Sq. ReLU</td>
<td>24 \ 2048 \ 16 \ GeLU</td>
</tr>
<tr>
<td>Flamingo-9B</td>
<td>6 \ 1536 \ 16 \ Sq. ReLU</td>
<td>10 \ 4096 \ 32 \ Sq. ReLU</td>
<td>40 \ 4096 \ 32 \ GeLU</td>
</tr>
<tr>
<td>Flamingo</td>
<td>6 \ 1536 \ 16 \ Sq. ReLU</td>
<td>12 \ 8192 \ 64 \ Sq. ReLU</td>
<td>80 \ 8192 \ 64 \ GeLU</td>
</tr>
</tbody>
</table>

L: Layers, D: Transformer Hidden Size, H: Number of heads
Appendix - Training the Image Encoder

- Details:
  - ALIGN and LTIP datasets
  - Resolution: 288 x 288
  - Embedding Size: 1376
  - Adam Opt.
  - Gradient clipping

- Evaluation
  - Zero-shot image classification -> Image-text retrieval

- Why use BERT?
  - To be able to extract contextual features rather than pure geometric features.
  - If trained with a LLM it generalizes better as a visual conditioner.
Appendix - Training the Image Encoder

● Trained from scratch with BERT language encoder
  ○ Text-to-image contrastive loss
    \[
    L_{\text{contrastive:txt2im}} = -\frac{1}{N} \sum_{i}^{N} \log \left( \frac{\exp \left( L_i^{\top} V_i \beta \right)}{\sum_{j}^{N} \exp \left( L_i^{\top} V_j \beta \right)} \right)
    \]
  ○ Image-to-text contrastive loss
    \[
    L_{\text{contrastive:im2txt}} = -\frac{1}{N} \sum_{i}^{N} \log \left( \frac{\exp \left( V_i^{\top} L_i \beta \right)}{\sum_{j}^{N} \exp \left( V_i^{\top} L_j \beta \right)} \right)
    \]
  Trainable inverse temperature parameter
Appendix - Visually Conditioned Large Language Models

- Examples are from Flamingo [2]