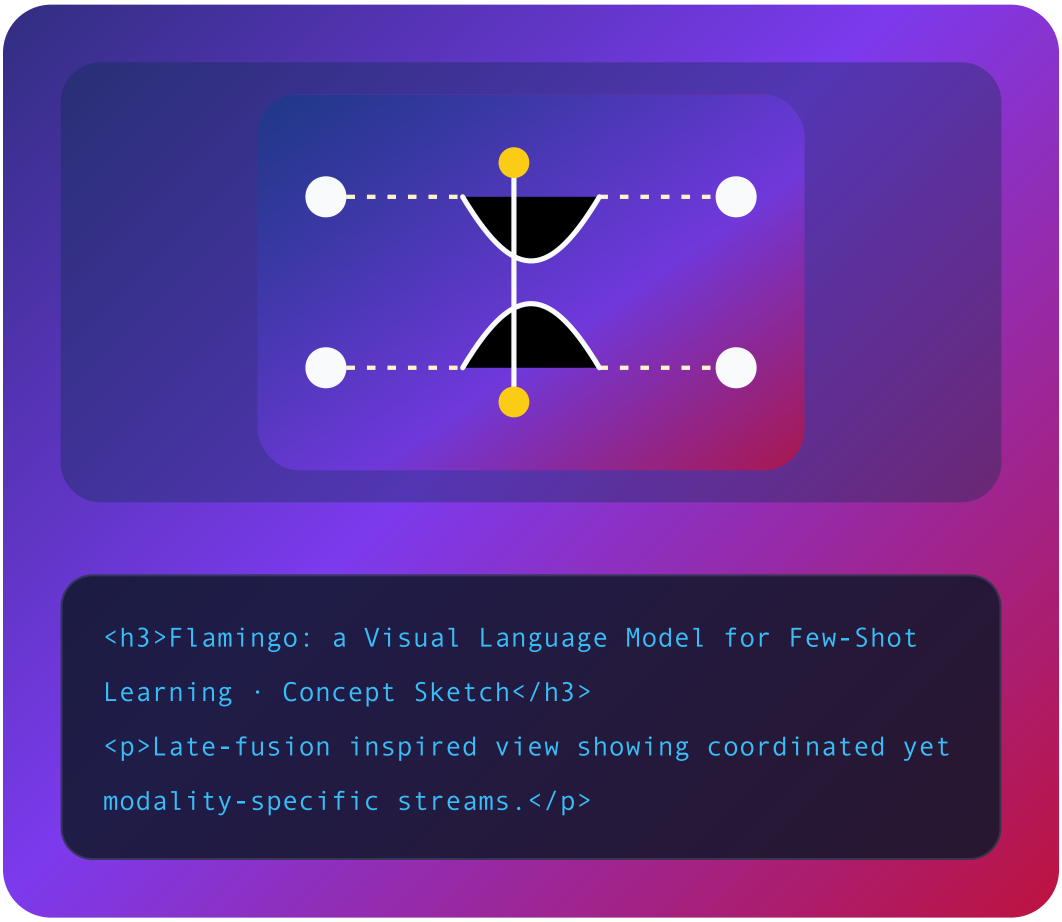

Flamingo: a Visual Language Model for Few-Shot Learning

Authors: Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan
Year: 2022
Venue: 36th Conference on Neural Information Processing Systems (NeurIPS)

1. Classification

- **Domain Category:**
  - **Vision / VLM / Multimodal FM.** Flamingo is a visual language model (VLM) that integrates vision and language for few-shot learning on image and video understanding tasks.
- **FM Usage Type:**
  - **Core FM development.** Flamingo introduces a new family of VLMs with architectural innovations (Perceiver Resampler, GATED XATTN-DENSE layers) that enable few-shot learning via in-context examples.
- **Key Modalities:**
  - **Images:** High-resolution images from web data.
  - **Videos:** Short video clips (average 22 seconds) with temporal information.
  - **Text:** Interleaved text descriptions, captions, questions, and answers.

---

2. Executive Summary

This paper introduces Flamingo, a family of Visual Language Models (VLMs) that achieve state-of-the-art few-shot learning on a wide range of image and video understanding tasks, simply by being prompted with a few input/output examples—analogous to how GPT-3 performs few-shot learning with text. Flamingo bridges powerful pretrained vision-only and language-only models through novel architectural components: a Perceiver Resampler that converts variable-size visual feature maps into a fixed number of visual tokens, and GATED XATTN-DENSE layers that condition frozen language models on visual representations via gated cross-attention. The model can handle arbitrarily interleaved sequences of images/videos and text, enabling natural few-shot prompting where task examples are provided as (image, text) pairs followed by a query.

Flamingo is trained on a large-scale mixture of web-scraped multimodal data (interleaved image-text from webpages, image-text pairs, video-text pairs) totaling billions of examples, without using any task-specific annotations. After training, a single Flamingo model achieves new state-of-the-art few-shot performance on 16 diverse benchmarks (visual question-answering, captioning, classification, dialogue) and even outperforms fine-tuned models on 6 tasks despite using only 32 task-specific examples (around 1000× less data than fine-tuned SotA). The largest model (Flamingo-80B) sets new records across open-ended tasks like VQA and captioning. This work demonstrates how to effectively combine pretrained vision and language models for multimodal few-shot learning, and shows that large-scale web data training can enable powerful in-context learning capabilities previously seen only in text-only language models.

3. Problem Setup and Motivation

- **Scientific / practical problem**
    - Current vision-language models require extensive task-specific fine-tuning with thousands of annotated examples, making adaptation to new tasks expensive and slow.
    - Contrastive models (CLIP) enable zero-shot classification but lack generative capabilities for open-ended tasks like captioning and VQA.
    - Vision-conditioned language generation models exist but haven't shown strong few-shot learning abilities.
    - The goal is to build a model that can rapidly adapt to new vision-language tasks using only a few examples, similar to GPT-3's few-shot learning for text.

- **Why this is hard**
  - **Bridging vision and language:**
    - Vision encoders and language models are typically trained separately; connecting them effectively is non-trivial.
    - Need to preserve knowledge from both pretrained models while enabling cross-modal understanding.
  - **Handling interleaved multimodal sequences:**
    - Few-shot learning requires processing sequences like: ($image_1$, $text_1$), ($image_2$, $text_2$), …, (query_image, ?).
    - Standard architectures don't naturally handle such interleaved inputs.
  - **Variable-size visual inputs:**
    - Images and videos have variable resolutions and aspect ratios.
    - Language models expect fixed-size token sequences.
  - **Large-scale training data:**
    - Few-shot learning requires massive pretraining on diverse multimodal data.
    - Need to collect and process billions of image-text and video-text examples from the web.
  - **Training stability:**
    - Combining frozen pretrained models with new trainable components can be unstable.
    - Need careful initialization and gating mechanisms.

---

4. Data and Modalities

- **Datasets used**
  - **M3W (MultiModal MassiveWeb):**
    - ~43 million webpages with interleaved images and text extracted from HTML.
    - Images positioned relative to text based on DOM structure.

        - Up to 5 images per sequence, 256 tokens of text.
    - **ALIGN:**
        - 1.8 billion image-text pairs with alt-text descriptions.
    - **LTIP (Long Text & Image Pairs):**
        - 312 million image-text pairs with longer, higher-quality descriptions.
    - **VTP (Video & Text Pairs):**
        - 27 million short videos (average 22 seconds) with sentence descriptions.
    - **Evaluation benchmarks:**
        - 16 diverse tasks: VQAv2, OK-VQA, COCO captioning, TextVQA, VizWiz, MSRVTTQA, VATEX, VisDial, HatefulMemes, and more.
- **Modalities**
    - **Images:** High-resolution images from webpages and image-text pairs.
    - **Videos:** Short video clips (1 FPS sampling) with temporal embeddings.
    - **Text:** Captions, questions, answers, descriptions, interleaved with visual content.
- **Preprocessing / representation**
    - **Vision encoder:**
        - Pretrained NFNet-F6 (NormalizerFree ResNet) with contrastive pretraining.
        - Outputs 2D spatial grid of features, flattened to 1D sequence.
        - For videos: frames sampled at 1 FPS, encoded independently, temporal embeddings added.
    - **Perceiver Resampler:**
        - Takes variable number of visual features, produces fixed 64 visual tokens.
        - Uses learned latent queries that cross-attend to visual features.
    - **Text:**
        - Tokenized using language model's tokenizer.

        - Special tokens: `<image>`, `<EOC>` (end of chunk).

---

5. Model / Foundation Model

- **Model Type**
    - **Multimodal autoregressive language model** that generates text conditioned on interleaved visual and textual inputs.
    - Architecture: frozen vision encoder + trainable Perceiver Resampler + frozen language model with interleaved GATED XATTN-DENSE layers.
- **Is it a new FM or an existing one?**
    - **New FM.** Flamingo introduces a new family of VLMs with specific architectural innovations for few-shot learning, though it builds on pretrained vision and language models.
- **Key components and innovations**
    - **Perceiver Resampler:**
        - Converts variable-size visual feature maps into fixed 64 visual tokens.
        - Uses learned latent queries in a Transformer that cross-attends to visual features.
        - Enables efficient processing of high-resolution images and videos.
    - **GATED XATTN-DENSE layers:**
        - Inserted between frozen language model layers.
        - Structure: gated cross-attention (queries from language, keys/values from vision) + gated feed-forward.
        - Gating: tanh(α) multiplier, initialized at 0, so model starts as pure language model and gradually incorporates vision.
        - Preserves pretrained language model knowledge while enabling visual conditioning.

  - **Image-causal masking:**
    - At each text token, model only attends to visual tokens from the immediately preceding image (not all previous images).
    - Enables generalization to any number of images, regardless of training distribution.
  - **Model sizes:**
    - Flamingo-3B, Flamingo-9B, Flamingo-80B (based on Chinchilla 1.4B, 7B, 70B language models).
- **Training setup**
  - **Objective:** Autoregressive text generation conditioned on interleaved visual inputs.
  - **Loss:** Weighted sum of per-dataset negative log-likelihoods.
  - **Training strategy:**
    - Gradient accumulation over all datasets (outperforms round-robin).
    - Careful dataset weighting (λ_m) is crucial for performance.
  - **Few-shot adaptation:**
    - No fine-tuning; simply prompt with (image, text) example pairs followed by query.
    - Uses beam search for open-ended generation.

---

6. Multimodal / Integration Aspects (If Applicable)

- **Vision-language integration:**
  - Flamingo explicitly integrates images/videos and text through:
    - **Perceiver Resampler:** Bridges vision encoder and language model by converting visual features to tokens.
    - **GATED XATTN-DENSE layers:** Enable language model to condition on visual tokens via cross-attention.

        - **Interleaved sequences:** Support arbitrary mixing of visual and textual inputs.
- **Integration strategy:**
    - **Late fusion with cross-attention:**
        - Vision and language are processed separately (frozen encoders), then fused via cross-attention in GATED XATTN-DENSE layers.
        - This preserves pretrained knowledge while enabling multimodal understanding.
- **What this enables:**
    - **Few-shot learning:** Model can adapt to new tasks by seeing a few (image, text) examples.
    - **Open-ended generation:** Can generate captions, answers, descriptions conditioned on images/videos.
    - **Multi-image reasoning:** Can process sequences of multiple images with interleaved text (e.g., visual dialogue).
    - **Zero-shot capabilities:** Works out-of-the-box on tasks not seen during training.
- **Not biological multimodal:**
    - Flamingo focuses on natural images/videos and text, not biological data (brain imaging, genomics, etc.).
    - However, the architectural principles (Perceiver Resampler, gated cross-attention) could potentially be adapted to biological multimodal settings.

---

7. Experiments and Results

Main findings

- **State-of-the-art few-shot performance:**
  - Flamingo-80B sets new SotA on 9 of 16 tasks with few-shot learning (4-32 shots).
  - Outperforms fine-tuned models on 6 tasks despite using only 32 examples (vs. thousands for fine-tuning).
- **Performance by task type:**
  - **Visual question-answering:** Flamingo-80B achieves 57.8% on VQAv2 (32-shot) vs. 80.2% for fine-tuned SotA, but with 1000× less data.
  - **Captioning:** 113.8 CIDEr on COCO (32-shot) vs. 143.3 for fine-tuned SotA.
  - **Video understanding:** Strong performance on MSRVTTQA, VATEX, NextQA.
  - **Visual dialogue:** Competitive on VisDial, TextVQA.
- **Scaling with model size:**
  - Performance improves with model size (3B → 9B → 80B) and number of shots (0 → 4 → 32).
- **Zero-shot performance:**
  - Flamingo-80B achieves strong zero-shot results on many tasks, though few-shot prompting further improves performance.

Ablation studies

- **Perceiver Resampler:**
  - Outperforms plain Transformer and MLP alternatives for vision-language connection.
- **GATED XATTN-DENSE layers:**
  - Gating mechanism (tanh initialization) improves training stability and final performance.

    - Frequency of layer insertion trades off efficiency vs. expressivity.
- **Image-causal masking:**
    - Single-image cross-attention (only attend to immediately preceding image) outperforms attending to all previous images.
- **Dataset weighting:**
    - Careful tuning of per-dataset weights (λ_m) is crucial; gradient accumulation outperforms round-robin sampling.

Key insights

- **Large-scale web data enables few-shot learning:**
    - Training on billions of interleaved image-text examples from the web (without task-specific annotations) enables powerful few-shot capabilities.
- **Frozen pretrained models + trainable connectors:**
    - Preserving pretrained vision and language knowledge while adding minimal trainable components (Perceiver, GATED layers) is effective.
- **Interleaved sequences are key:**
    - Ability to process arbitrarily interleaved visual and textual inputs enables natural few-shot prompting.

---

8. Strengths and Limitations

Strengths

- **Powerful few-shot learning:**
    - Achieves SotA on many tasks with just 4-32 examples, dramatically reducing annotation requirements.
- **Open-ended generation:**
    - Can generate free-form text (captions, answers) unlike contrastive models (CLIP) that only do classification.

- **Handles diverse tasks:**
    - Single model works on classification, captioning, VQA, dialogue, video understanding.
- **Leverages pretrained models:**
    - Effectively combines frozen vision and language models, preserving their knowledge.
- **Scalable architecture:**
    - Works across model sizes (3B to 80B) with consistent improvements.
- **Large-scale training:**
    - Demonstrates value of web-scale multimodal data for foundation model training.

Limitations

- **Still behind fine-tuned models:**
    - On some tasks, fine-tuned models with thousands of examples outperform Flamingo's few-shot performance.
- **Compute intensive:**
    - Training on billions of examples and 80B parameters requires massive compute resources.
- **Limited to vision-language:**
    - Doesn't handle other modalities (audio, 3D, etc.) or biological data.
- **Hallucination:**
    - Can generate plausible but incorrect captions or answers, especially in zero-shot settings.
- **Evaluation gaps:**
    - Some benchmarks may have data leakage or limited diversity; held-out evaluation is important.

Open questions / future directions

- **How to improve zero-shot performance?**
  - Can better training strategies or architectures close the gap with fine-tuned models?
- **Other modalities:**
  - Can Flamingo-style architectures handle audio, 3D, or biological data?
- **Longer context:**
  - Can models handle longer video sequences or more interleaved examples?
- **Interpretability:**
  - How does the model reason about visual and textual inputs? Can we interpret cross-attention patterns?
- **Efficiency:**
  - Can smaller models achieve similar few-shot performance with better architectures or training?

---

9. Context and Broader Impact

Relation to other work

- **Compared to CLIP (Radford et al., 2021):**
  - CLIP uses contrastive learning for zero-shot classification but can't generate text.
  - Flamingo enables open-ended generation and few-shot learning via in-context examples.
- **Compared to BLIP-2 (Li et al., 2023):**
  - BLIP-2 also bridges vision and language but focuses on fine-tuning rather than few-shot learning.

    - Flamingo’s few-shot capabilities are more flexible for rapid adaptation.
- **Compared to GPT-3 (Brown et al., 2020):**
    - GPT-3 showed few-shot learning works for text; Flamingo extends this to vision-language.
    - Uses similar prompting paradigm but with multimodal inputs.
- **Connection to Perceiver (Jaegle et al., 2021):**
    - Flamingo’s Perceiver Resampler adapts Perceiver architecture for vision-language connection.
    - Enables handling variable-size visual inputs efficiently.

Broader scientific and practical impact

- **Enables rapid adaptation:**
    - Few-shot learning reduces annotation costs and enables faster deployment to new tasks.
- **Demonstrates web-scale training value:**
    - Shows that large-scale web data (without task-specific annotations) can enable powerful capabilities.
- **Influences future VLMs:**
    - Architectural innovations (Perceiver Resampler, gated cross-attention) influence subsequent models (GPT-4V, LLaVA, etc.).
- **Opens new applications:**
    - Few-shot learning enables applications where collecting large labeled datasets is impractical.

Open questions for future research

- **How to scale further?**
    - Can even larger models or more data improve few-shot performance?
- **Other modalities:**
    - Can similar architectures handle audio, 3D, or biological multimodal data?

- **Efficiency:**
  - Can smaller, more efficient models achieve similar capabilities?
- **Robustness:**
  - How do models handle out-of-distribution images, adversarial examples, or biased data?

---

10. Key Takeaways

1. **Few-shot learning works for vision-language:**
   Just as GPT-3 showed few-shot learning for text, Flamingo demonstrates it's possible for multimodal tasks by prompting with (image, text) examples.

2. **Bridging pretrained models is effective:**
   Rather than training from scratch, combining frozen vision and language models with minimal trainable connectors (Perceiver, gated cross-attention) preserves knowledge while enabling new capabilities.

3. **Interleaved sequences enable natural prompting:**
   Ability to process arbitrarily interleaved visual and textual inputs makes few-shot prompting natural and flexible.

4. **Large-scale web data is valuable:**
   Training on billions of web-scraped image-text pairs (without task annotations) enables powerful few-shot capabilities.

5. **Gating mechanisms stabilize training:**
   Initializing new layers with tanh gating (starting at 0) ensures model begins as pure language model and gradually incorporates vision, improving stability.

6. **Perceiver Resampler handles variable inputs:**
   Converting variable-size visual feature maps to fixed tokens enables efficient processing of diverse images and videos.

7. **Single-image cross-attention is sufficient:**
   Attending only to the immediately preceding image (not all previous images) works well and enables generalization to any number of images.

8. **Few-shot can match fine-tuning:**
   On some tasks, 32-shot Flamingo matches or exceeds fine-tuned models trained on thousands of examples.

9. **Open-ended generation is powerful:**
   Unlike contrastive models, Flamingo can generate free-form text, enabling captioning, VQA, and dialogue.

10. **This is foundational work:**
    Flamingo establishes few-shot learning as a viable paradigm for vision-language tasks, influencing subsequent VLMs and demonstrating the power of web-scale multimodal training.