Scaling Autoregressive Multi-Modal Models: Pretraining and Instruction Tuning

arXiv 2023 (44 Citations)


Presented by:
Michael Cruz, Christopher Lee, Saurabh Aggarwal, Robert Martin
Problem Statement

- Lack of a computationally efficient, multi-modal language model that excels at both text-to-text and image-to-text generation.
- Diffusion Models are computationally efficient while auto-regressive models give better outputs at the cost of being computationally expensive.
- CM3Leon bridges this gap by providing a computationally efficient autoregressive model with SOTA performance in multimodal generative tasks.
CM3Leon

- A retrieval-augmented, token-based, decoder-only multi-modal language model
- Capable of generating and infilling both text and images
- Uses CM3 multi-modal architecture with upscaled training and more diverse dataset
- Pretraining follows the retrieval-augmented CM3 approach with new large-scale Shutterstock dataset
- Supervised fine-tuning phase follows multi-task instruction tuning for text-only models
  - Allows for arbitrary mixtures of image and text tokens in both the input and output
Image Text Tasks
CM3Leon Architecture
CM3: Causal Multimodal Model
CM3: Causal Multimodal Model
RA-CM3 Architecture
RA-CM3 Architecture
Retrieval Augmentation: Dense Retriever
RA-CM3 Architecture
Example Input

1 caption-image pair

1 image-caption pair + $k$ retrieved pairs
RA-CM3 Architecture
Retrieval Augmentation: Retrieval Strategy

Objective function:

- $L = L_{\text{main}} + \alpha L_{\text{retr}}$
- $L = -\log p(x|m_1, ..., m_K) - \alpha \log p(m_1, ..., m_K)$
RA-CM3 Architecture
RA-CM3 Alterations

- Reduced scope of objective function
Reduced Scope of Objective Function

Objective function:

- \( L = L_{\text{main}} + \alpha L_{\text{retr}} \)
- \( L = -\log p(x|m_1, \ldots, m_K) - \alpha \log p(m_1, \ldots, m_K) \)
RA-CM3 Alterations

- Reduced scope of objective function
- Reduced decoder hyperparameters
Reduced Scope of Objective Function

1 Text

1 image-caption pair + $k$ retrieved pairs
RA-CM3 Alterations

- Reduced scope of objective function
- Reduced decoder hyperparameters
- Modified Dataset
Decoding Strategies

- **Temperature Sampling:** involves modifying the SoftMax temperature during the sampling stage to control the randomness of predictions.
- **TopP Sampling:** involves sampling from the smallest set of top-ranked tokens with a cumulative probability exceeding a predefined threshold.
- **Classifier Free Guidance (CFG):** CFG refers to directing an unconditional sample towards a conditional sample.
- **Contrastive Decoding TopK (CD-K):** a variant of the contrastive decoding algorithm.
Decoding Strategies: Classifier Free Guidance (CFG)

- In this technique, the text with the `<mask>` token is replaced to facilitate unconditional sampling.
- This allows for CFG without the need for fine tuning.
- During inference, two concurrent token streams are generated:
  - Conditional token stream contingent on input text
  - Unconditional token stream conditioned on a mask token
- CFG:
  - $\text{logits}_{\text{cond}} = T(t_y | t_x)$, $\text{logits}_{\text{uncond}} = T(t_y | \text{< mask >})$
  - $\text{logits}_{\text{cf}} = \text{logits}_{\text{uncond}} + \alpha_c \cdot (\text{logits}_{\text{cond}} - \text{logits}_{\text{uncond}})$
  - $T = \text{Transformer}$
  - $t_y = \text{output token}$
  - $t_x = \text{conditional input}$
  - $\alpha_c = \text{scaling factor}$
Decoding Strategies: Contrastive Decoding Method (CD)

- The logit subtraction from CFG resembles the log probability subtraction in CD methods.
- For this reason, a variant of the CD algorithm was used as an alternative to CFG.
- CD per token score:
  \[
  CD(t_{yi}; t_{y<i}) = \begin{cases} 
  \log \frac{P_{EXP}(t_{yi}|t_{y<i})}{P_{AMA}(t_{yi}|t_{y<i})}, & \text{if } t_{yi} \in v(t_{y<i}), \\
  -\infty, & \text{otherwise}
  \end{cases}
  \]
  - \(P_{EXP}\) = strong model trained with more compute (or larger model size)
  - \(P_{AMA}\) = weak model trained with less compute (or smaller model size)
  - \(v(t_{y<i}) = \{t_{yi} \in v : P_{EXP}(t_{yi}|t_{y<i}) \geq \alpha \max_w P_{EXP}(w|t_{y<i})\}\)
    - set of potential subsequent tokens whose probabilities are at least times the maximum probability value:
Decoding Strategies: Contrastive Decoding TopK (CD-K)

- Variant of the CD algorithm
- $P_{\text{EXP}}$ is selected having text conditioning
- $P_{\text{AMA}}$ is selected to have no text conditioning
- $v(t_{y<i})$ constraint was too strict and would consistently become greedy decoding
- CD-K:
  - $v(t_{y<i}) = \{t_{yi} \in v : P_{\text{EXP}}(t_{yi}|t_{y<i}) \cdot \alpha \cdot \max_{k\leq w} \left( P_{\text{EXP}}(w|t_{y<i}) \right) \}$
  - Instead of taking the largest probability, the k-th largest probability is taken
Supervised Fine Tuning

- Trains a model to understand future instructions or prompts
- Amplifies multi-modal model performance across various tasks
- CM3Leon was fine-tuned on an array of mixed image and text tasks
- Tasks are organized as a series of interleaved image and text examples
- Fine tuning process employs the CM3 objective by combining the task instruction with the output
Long Form Generation Task

**Instruction:** Question: Which property do these two objects have in common?
Context: Select the best answer.
Options: (A) shiny (B) salty
Answer: Let's think step-by-step:

**CM3Leon Output:** Think about each object. Potato chips have a salty taste. The pretzel is salty. So the answer is (B).

**Instruction:** Describe all the objects in the given image in very detail.

**CM3Leon Output:** A street sign is on a metal pole. The sign is blue with white writing. There is a red light on the pole. The sky is a bright blue. There are cars parked on the street. There are trees on the side of the street. There are buildings on the side of the street.
How-to-Write Task
Spatially Grounded Image Generation

Generate high quality image of "a room that has a sink and a mirror in it" with bottle at location (199, 130) -> (204, 150) and with sink at location (149, 133) -> (190, 154) and with bed at location (0, 169) -> (67, 255).
Image Captioning and Visual Question Answering Tasks

**Instruction:** Describe the given image.
**CM3Leon Output:** A man is standing on a beach with a surfboard.

**Instruction:** Question: What are people flying?
**CM3Leon Output:** Kites
Supervised Fine Tuning

<table>
<thead>
<tr>
<th>Model</th>
<th>MS-COCO CIDEr (test)</th>
<th>VQA2 Acc. (test-dev)</th>
<th>VizWiz Acc. (test-dev)</th>
<th>OKVQA Acc. (val)</th>
<th>Image Paragraph CIDEr (test)</th>
<th>VisDial NDCG (val)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenFlamingo-9B† (0-shot)</td>
<td>65.5</td>
<td>43.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Flamingo-9B (0-shot)</td>
<td>79.4</td>
<td>51.8</td>
<td>28.8</td>
<td>44.7</td>
<td>-</td>
<td>48.4</td>
</tr>
<tr>
<td>SFT-CM3Leon-7B (0-shot)</td>
<td>61.6</td>
<td>47.6</td>
<td>37.6</td>
<td>23.8</td>
<td>10.5</td>
<td>22.6</td>
</tr>
</tbody>
</table>
Evaluation Datasets

- Shutterstock Dataset
- MS-COCO
- Flickr30k
- Image Paragraph
- Localized Narratives
- VQA2
- VizWiz
- OKVQA
- ScienceQA
Quantitative Evaluations
## Zero-Shot: MSCOCO

<table>
<thead>
<tr>
<th>Method</th>
<th>Retrieval in Training</th>
<th>Responsible</th>
<th># of Retrieved Documents</th>
<th>Dataset Size</th>
<th>Model Size</th>
<th>Zero-shot FID-30K</th>
</tr>
</thead>
<tbody>
<tr>
<td>RA-CM3</td>
<td>✓</td>
<td>x</td>
<td>2</td>
<td>150M</td>
<td>2.7B</td>
<td>15.70</td>
</tr>
<tr>
<td>StableDiffusion</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>400M</td>
<td>800M</td>
<td>12.60</td>
</tr>
<tr>
<td>KNN-Diffusion</td>
<td>✓</td>
<td>x</td>
<td>10</td>
<td>70M</td>
<td>400M</td>
<td>12.50</td>
</tr>
<tr>
<td>MUSE</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>500M</td>
<td>3B</td>
<td>7.88</td>
</tr>
<tr>
<td>PARTI</td>
<td>x</td>
<td>x</td>
<td>-</td>
<td>5B</td>
<td>20B</td>
<td>7.23</td>
</tr>
<tr>
<td>RE-IMAGEN</td>
<td>✓</td>
<td>x</td>
<td>2</td>
<td>450M</td>
<td>3.6B</td>
<td>5.25</td>
</tr>
</tbody>
</table>
Inference Latency

<table>
<thead>
<tr>
<th>Model</th>
<th>Resolution</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imagen</td>
<td>256 × 256</td>
<td>9.1s</td>
</tr>
<tr>
<td>Imagen</td>
<td>1024 × 1024</td>
<td>13.1s</td>
</tr>
<tr>
<td>LDM (50 steps)</td>
<td>512 × 512</td>
<td>3.7s</td>
</tr>
<tr>
<td>LDM (250 steps)</td>
<td>512 × 512</td>
<td>18.5s</td>
</tr>
<tr>
<td>Parti (3B)</td>
<td>256 × 256</td>
<td>6.4s</td>
</tr>
<tr>
<td>MUSE (3B)</td>
<td>256 × 256</td>
<td>0.5s</td>
</tr>
<tr>
<td>CM3Leon (7B, BF16)</td>
<td>256 × 256</td>
<td>11.8s</td>
</tr>
<tr>
<td>CM3Leon (7B, INT8)</td>
<td>256 × 256</td>
<td>9.1s</td>
</tr>
</tbody>
</table>
Limitations

- Dependency on large scale data
- Complexity in model architecture and training
- Generalization issues
- Bias in models
- Ethical and responsible AI concerns
Conclusion & Future Work

- CM3Leon is a retrieval-augmented, token-based, decoder-only, multi-modal language model for text-to-image and image-to-text generation tasks.
- Achieves state-of-the-art performance in generating high-quality images from textual prompts and vice versa.
- Future work could focus on optimizing the training process, enhancing efficiency and scalability.
- There are countless possibilities in extending these methods in zero-shot tasks.
References