Back

TII Releases First SSLM With Falcon Mamba 7B

Falcon Mamba 7B: Highlights

- **Falcon Mamba 7B is the first open source released State Space Language Model (SSLM), a new revolutionary architecture for Falcon models**
- **Falcon Mamba 7B is the no. 1 globally performing open source SSLM in the world, as independently verified by Hugging Face**
- **SSLMs have a low memory cost and don't require additional memory to generate arbitrary long blocks of text**
- **Falcon Mamba 7B also outperforms traditional transformer architecture models such as Meta's Llama 3.1 8B and Mistral's 7B**
- **New model reflects the innovation and pioneering approach of Abu Dhabi in AI research and development**

This new Falcon Mamba 7B model is yet another example of the pioneering approach of Abu Dhabi in AI research and development.

The Falcon Mamba 7B is the no. 1 globally performing open source State Space Model Language (SSLM) in the world, as independently verified by Hugging Face. It outperforms traditional transformer architecture models such as Meta's Llama 3.1 8B and Mistral's 7B.

What's New

Falcon Mamba 7B is our first SSLM. We are really excited about this new model, which allows to process larger blocks of text. SSLMs have a low memory cost and don't require additional memory to generate arbitrary long sequences.

How can you use Falcon Mamba 7B?

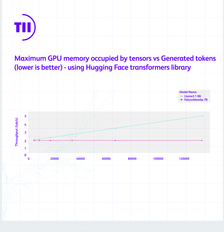
You can find the model on Hugging Face:
https://huggingface.co/tiiuae/falcon-mamba-7b

We have also provided an interactive playground as well for everyone to try the model:
https://huggingface.co/spaces/tiiuae/falcon-mamba-playground

How does the Falcon Mamba 7B fare?

For the transformer architecture models, Falcon Mamba 7B outperforms Meta's Llama 3.1 8B and Mistral's 7B. Meanwhile for the other SSLMs, Falcon Mamba 7B beats all other open source models in the old benchmarks and it will be the be first model on Hugging Face's new tougher benchmark leaderboard.

We test the largest sequence length that can fit on a single 24 GB A10 GPU. The batch size is fixed at 1 and we are using float32 precision. Transformer-based models use a resized vocabulary size to match Falcon Mamba model. It can be observed that we can fit larger sequences than SoTA transformer-based models while theoretically being able to fit infinite context length if one processes the entire context token by token, or by chunks of tokens with a size that fits on the GPU, denoted as sequential parallel.



We measure the generation throughput in a setting with a prompt of length 1 and up to 130k generated tokens, using batch size 1 and H100 GPU. We compare against Mistral architecture that is based on sliding window attention as this is more memory efficient at scale. We observe that our Falcon Mamba is generating all the tokens at constant throughput and without any increase in CUDA peak memory. For the transformer model, the peak memory grows, and generation speed slows down as the number of generated tokens grows.

Word of mouth

H.E. Faisal Al Bannai, Secretary General of ATRC and Adviser to the UAE President for Strategic Research and Advanced Technology Affairs

"The Falcon Mamba 7B marks TII's fourth consecutive top-ranked AI model, reinforcing Abu Dhabi as a global hub for AI research and development. This achievement highlights the UAE's unwavering commitment to innovation."

Dr. Najwa Aaraj, Chief Executive of TII

"The Technology Innovation Institute continues to push the boundaries of technology with its Falcon series of AI models. The Falcon Mamba 7B represents true pioneering work and paves the way for future AI innovations that will enhance human capabilities and improve lives."

Dr. Hakim Hacid, Acting Chief Researcher of the TII's AI Cross-Center Unit

"As we introduce the Falcon Mamba 7B, I'm proud of the collaborative ecosystem of TII that nurtured its development. This release represents a significant stride forward, inspiring fresh perspectives and further fueling the quest for intelligent systems. At TII, we're pushing the boundries of both SSLM and transformer models to spark further innovation in generative AI."

What's Next

We are moving forward fast with more fundamental research on Large Language Models, across the entire spectrum of the industry, from novel model architecture design to optimal training strategy to high quality data processing. Up next, we're looking to further optimize the model design and scale up the model to cover more application scenarios.