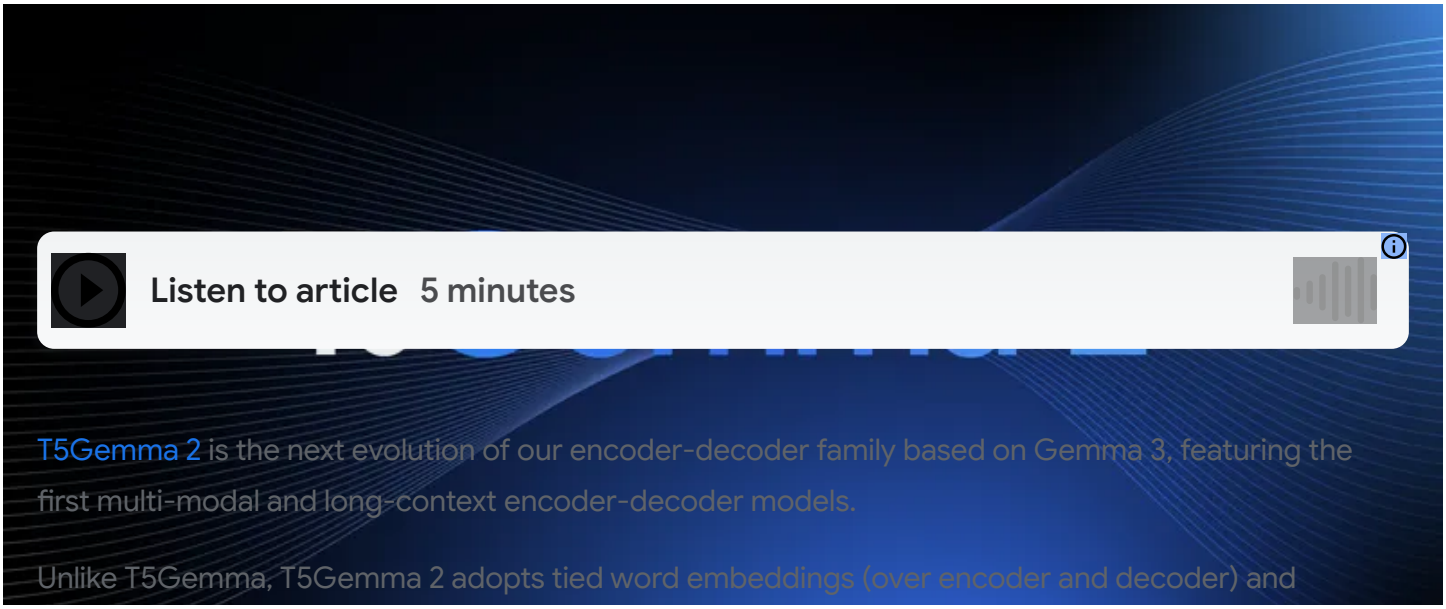
T5Gemma 2: The next generation of encoder-decoder models

Dec 18, 2025 · 3 min read

T5Gemma 2 is more than a re-training. It incorporates significant architectural changes while inheriting many of the powerful, next-generation features of the Gemma 3 family.

Biao Zhang
Research Scientist

Ben Hora
Product Manager



merged decoder self- and cross-attention to save model parameters. It offers compact pre-trained models at sizes of 270M-270M (~370M total, excluding vision encoder), 1B-1B (~1.7B) and 4B-4B (~7B) parameters, making them ideal for rapid experimentation and deployment in on-device applications.

Background

With the original T5Gemma, we demonstrated that we could successfully adapt modern, pre-trained decoder-only models into an encoder-decoder architecture, unlocking new versatility. By initializing with weights from a powerful decoder-only model and then applying continued pre-training, we created high-quality, inference-efficient models while bypassing the computational cost of training from scratch.

T5Gemma 2 extends this into the realm of vision-language models by incorporating key innovations from Gemma 3.

What's new

T5Gemma 2 is more than a re-training. It incorporates significant architectural changes while inheriting many of the powerful, next-generation features of the Gemma 3 family.

Architectural innovations for efficiency

To maximize efficiency at smaller scales, we have introduced key structural refinements:

- **Tied embeddings:** We now tie the embeddings between the encoder and decoder. This significantly reduces the overall parameter count, allowing us to pack more active capabilities into the same memory footprint — crucial for our new compact 270M-270M model.

- **Merged attention:** In the decoder, we adopt a merged attention mechanism, combining self- and cross-attention into a single, unified attention layer. This reduces model parameters and architectural complexity, improving model parallelization and benefiting inference.

Next-generation capabilities

Drawing from Gemma 3, T5Gemma 2 also represents a significant upgrade in model capabilities:

- **Multimodality:** T5Gemma 2 models can understand and process images alongside text. By utilizing a highly efficient vision encoder, the models can seamlessly perform visual question answering and multimodal reasoning tasks.

- **Extended long context:** We've dramatically expanded the context window. Leveraging Gemma 3's alternating local and global attention mechanism, T5Gemma 2 can handle context windows of up to 128K tokens.

- **Massively multilingual:** Trained on a larger, more diverse dataset, these models now support over 140 languages out of the box.

Performance

T5Gemma 2 sets a new standard for what compact encoder-decoder models can achieve. Our new models demonstrate strong performance across key capability areas, inheriting the powerful multimodal and long-context features from the Gemma 3 architecture.

Pre-training performance of Gemma 3, T5Gemma and T5Gemma 2 across five unique capabilities.

As shown in the charts above, T5Gemma 2 delivers:

- **Strong multimodal performance**, outperforming Gemma 3 on several benchmarks. We adapt text-only Gemma 3 base models (270M and 1B) into effective multimodal encoder-decoder models.

- **Superior long-context capability**, with substantial quality gains over Gemma 3 and T5Gemma. Using a separate encoder makes T5Gemma 2 better at handling long-context problems.

- **Improved general capabilities**. Across coding, reasoning and multilingual tasks, T5Gemma 2 generally surpasses its corresponding Gemma 3 counterpart.

Post-training performance. Note: we are not releasing any post-trained / IT checkpoints. These results here are only for illustration, where we performed a minimal SFT without RL for T5Gemma 2. Also note pre-training and post-training benchmarks are different, so scores are not comparable across plots.

Similar to the original T5Gemma, we find that the post-training performance of T5Gemma 2 generally yields better results than its decoder-only counterparts. This makes T5Gemma 2 suitable for both large language model research as well as downstream applications.

Getting started

We're looking forward to seeing what the community builds with T5Gemma 2. This release includes pre-trained checkpoints, designed to be post-trained by developers for specific tasks before deployment.

These pre-trained checkpoints are available now for broad use across several platforms:

- **Read the paper on arXiv**
- **Download on Kaggle**
- **Available on Hugging Face**
- **Explore via Colab**
- **Run Inference via Vertex AI**

POSTED IN:

Developer tools