You've told me that your training cutoff is July 2024, and that you don't know about DeepSeek-V3, released Jan. 2025. However, I have a related question which you can perhaps answer as a matter of general knowledge, or perhaps from knowledge of DeepSeek-R1. The DeepSeek-V3-Base folder on HuggingFace contains 163 *.safetensors files (with filenames such as "model-00001-of-000163.safetensors"), totaling about 687.9 GB in size. Would that generally define the size of a local installation on my laptop, for example? Would I need to download all 163 *.safetensors files?

Thought for 55 seconds

> Okay, let me try to figure this out. The user is asking about the size of a local installation for the DeepSeek-V3-Base model, which has 163 safetensors files totaling around 687.9 GB. They want to know if they need to download all those files for a local setup on their laptop.
>
> First, I remember that when models are stored on HuggingFace, they often split large files into smaller parts. This is probably to make downloading easier, especially with potential interruptions. So each of those 163 files is a part of the whole model. But when you actually use the model, you don't need all the parts at once, right? Wait, no. When loading the model, the framework (like PyTorch or TensorFlow) usually expects all the parts to be present because they contain different portions of the model's parameters. So if the model is split into 163 files, you do need all of them to load the model correctly.
>
> But wait, sometimes checkpoints are saved in multiple files, but maybe each file is a different layer or component. But in reality, frameworks like HuggingFace Transformers handle the loading by combining all the split files. So the total size on disk would indeed be the sum of all the parts because each file is a necessary piece of the model. So the user would need all 163 files, and the total disk space required would be around 688 GB. That's a lot for a laptop.
>
> But maybe there's a misunderstanding here. Some models have different formats. For example, sometimes safetensors files are sharded, meaning they split the tensors across multiple files. So if the model is 688 GB, that's the size of the parameters in half-precision (like float16 or bfloat16). Each parameter is 2 bytes, so 688e9 * 2 bytes would be... Wait, no. Wait, the number of parameters can be calculated. If the model is, say, 7B parameters, each parameter in float16 is 2 bytes. So 7e9 * 2 bytes = 14e9 bytes = ~14 GB. But this model is 688 GB. That would imply the model has 688e9 / 2 = 344 billion parameters, which is way larger than any existing model. Wait, that can't be right. So maybe the model is stored in a different format. Wait, maybe it's stored in 32-bit floats instead of 16-bit? Then each parameter is 4 bytes. 344e9 parameters would be 1.376e12 bytes, which is about 1.376 TB. But the total here is 688 GB, so maybe it's stored in 16-bit. Wait, 688 GB is 688 * 1e9 bytes. Divided by 2 bytes per parameter (for 16-bit) gives 344e9 parameters. That's 344 billion parameters. That's huge. But models like GPT-3 are 175B parameters. So 344B would be even larger. Maybe there's a mistake here. Alternatively, maybe the files are stored with redundancy or include other data like optimizer states? But no, the safetensors files typically just contain the model weights.
>
> Wait, perhaps the user is confusing the total size. Let me check: 163 files totaling 687.9 GB. So each file is roughly 4.22 GB (687.9 / 163 ≈ 4.22). That seems plausible for sharded model weights. But the model size in parameters would depend on the data type. For example, if it's a 344B model in 16-bit, that's 688 GB. But that's a very large model. Alternatively, maybe the model uses quantization. But if it's the base model, it's probably in float16 or bfloat16.
>
> But regardless of the parameter count, the user's main question is: do they need all 163 files for a local install. The answer is yes. When a model is sharded into multiple safetensors files, you need all of them to load the model because each contains a portion of the model's tensors. The HuggingFace transformers library, for example, automatically handles loading all the shards when you use `from_pretrained()`. So the total disk space required would be the sum of all shards, which is ~688 GB. That's a massive amount for a laptop, both in terms of download and storage.
>
> However, maybe the user is looking at a specific version or there's a mistake. For instance, sometimes repositories might have multiple branches or versions, and the 163 files could include different checkpoints or something else. But if it's the model itself, then yes, all shards are needed. Alternatively, maybe there's a different format available, like a single file or a quantized version that's smaller. The user should check the repository for alternative formats or quantization options. If not, they'd have to download all files.

> Another point: safetensors is a secure alternative to pickle, but the number of files doesn't change the total size. So regardless of the number of shards, the total size is the same. So the local installation would require approximately 688 GB of space. That's a lot, but maybe the model is intended for enterprise use with high-resource systems, not personal laptops.
>
> In summary: Yes, all 163 files are necessary, and the total local installation size would be around 687.9 GB. Unless there's a quantized or smaller version available, the user would need all shards. For a laptop, that's impractical, so they might need to look for a smaller model or use cloud services.

To use the DeepSeek-V3-Base model locally, **you would indeed need to download all 163 `.safetensors` files**, as they collectively contain the model's weights. The total size of **~688 GB reflects the full unquantized model in its original precision (likely 16-bit or 32-bit floats)**. Here's a breakdown:

---

1. Why All 163 Files Are Required

- **Sharded Weights**: Large models are split into multiple files (shards) to avoid hitting filesize limits on platforms like Hugging Face Hub. Each shard contains a portion of the model's parameters.
- **Framework Dependency**: Libraries like Hugging Face `transformers` or `safetensors` automatically load all shards when initializing the model. Missing any file will cause errors.

---

2. Disk Space Implications

- **Local Installation Size**: The total disk space required is **~688 GB**, matching the sum of the shards. This is typical for massive, unquantized models (e.g., 344B parameters in 16-bit precision).
- **Laptop Feasibility**: Storing and running such a model on a standard laptop is **impractical** due to:
  - Storage constraints (688 GB exceeds most laptop SSDs).
  - Memory requirements (loading the model into RAM/VRAM would need even more resources).

---

3. Alternatives for Local Use

- **Quantized Versions**: Check if a smaller, quantized variant (e.g., 4-bit, 8-bit) is available. Quantization reduces model size significantly (e.g., 344B → ~86 GB in 4-bit).
- **Cloud/API Access**: If local deployment is infeasible, consider cloud services (AWS, GCP) or API endpoints (if offered by DeepSeek).
- **Smaller Models**: Opt for a smaller model variant (e.g., DeepSeek-7B or DeepSeek-R1) that fits your hardware.

Key Takeaway

The **~688 GB size** reflects the full-precision model. For personal use, prioritize quantized versions or smaller models unless you have enterprise-grade hardware. Always check the Hugging Face repository for additional formats or documentation.

What would the approximate local size be for DeepSeek-R1 or DeepSeek-7B?

Thought for 42 seconds

> Okay, the user is asking about the approximate local size for DeepSeek-R1 or DeepSeek-7B models. Let me start by recalling the previous conversation. They were inquiring about the large 688GB model and whether they need to download all those files. Now they want to know about smaller models, likely because they realized the big one is impractical for their laptop.
>
> First, I need to figure out the parameter counts for these models. The names suggest that 7B is 7 billion parameters, which is common. But DeepSeek-R1 isn't as straightforward. Maybe R1 is another version, possibly similar in size to 7B or another standard size. Wait, I should confirm if DeepSeek-R1 is a specific model. Since my knowledge cutoff is July 2024, and DeepSeek-V3 was mentioned as a future release, maybe R1 is an older version. But the user is asking for general estimates.
>
> For a 7B model, the standard calculation for FP32 is 4 bytes per parameter, so 7B * 4 bytes = 28GB. If it's FP16, that's half, so 14GB. But models are often stored in FP16 or BF16, so the size would be around 14GB. But quantization can reduce this. For example, 4-bit quantization would be 7B * 0.5 bytes = ~3.5GB. Similarly, 8-bit would be 7B * 1 byte = 7GB. So the answer for 7B should cover these scenarios.
>
> Now, for DeepSeek-R1. If R1 is similar to 7B, then the same logic applies. But maybe R1 is a different size. If I can't find specific info, I can note that model sizes vary and provide examples based on common configurations. Alternatively, maybe R1 is another name for a different parameter count. Wait, sometimes models are named with R1 for "Release 1" but the parameter size might still be in the same range. If unsure, it's safer to explain the general approach and apply it to hypothetical parameter counts.
>
> Also, the user mentioned DeepSeek-V3-Base was 163 files totaling 688GB. That model is 344B parameters. Let me check: 344B parameters in FP16 would be 344*2 = 688GB, which matches. So the calculation is correct. Applying the same logic, a 7B model in FP16 is 14GB. However, when stored with safetensors, there might be some overhead, but usually, the main size is the parameters. Quantization reduces this further.
>
> The user's context is a local installation on a laptop. They probably want something manageable. So explaining the different quantization levels and their sizes is key. Also, mentioning that even 7B in FP16 might be too big for some laptops, but quantization helps. Maybe also note that some repos provide pre-quantized versions, so they don't have to do it themselves.
>
> Wait, the user might confuse model size with the total disk space. Need to clarify that the size is approximate and depends on the format. Also, check if Hugging Face models are often shared in different formats. For example, a model page might have multiple branches like FP16, 4-bit, etc. So suggesting they look for those variants is helpful.
>
> Putting it all together: start with the 7B model, explain FP32, FP16, and quantized sizes. Then address R1, assuming it's similar or different. If unsure about R1's parameters, mention that if it's a 7B model, the same applies. Alternatively, if R1 is a different size, adjust accordingly. Since I don't have specific info on R1, it's better to state that the size depends on parameters and provide examples. Also, mention checking the Hugging Face repo for exact sizes and available formats.
>
> Potential pitfalls: Not knowing R1's exact specs. Solution: Generalize the answer based on parameter count. Also, the user might not know about quantization, so explain it briefly. Ensure the answer is helpful even with limited info on R1.

For models like **DeepSeek-7B** or **DeepSeek-R1** (assuming it is a similarly sized model), the local disk space required depends on **precision** (full-precision vs. quantized) and **format** (e.g., safetensors, PyTorch). Here's a general breakdown:

---

1. DeepSeek-7B (7 Billion Parameters)

- **Full Precision (FP32/FP16/BF16):**
  - **FP32 (32-bit float):** ~28 GB (7B params × 4 bytes/param).
  - **FP16/BF16 (16-bit):** ~14 GB (7B × 2 bytes/param).
  - **Typical Download Size:** Most Hugging Face repos share 16-bit models, so expect **~14-15 GB** for all shards.
- **Quantized Versions (Smaller & Laptop-Friendly):**
  - **8-bit (int8):** ~7 GB (7B × 1 byte/param).

  - **4-bit (int4):** ~3.5-4 GB (7B × 0.5 bytes/param).
  - **GPTQ/AWQ Quantized:** ~3-6 GB (varies by method and calibration).

---

2. DeepSeek-R1 (Assumed Similar to 7B)

If DeepSeek-R1 is a **7B-class model**, the sizes above apply. If it's smaller (e.g., **1-3B parameters**), adjust accordingly:

- **3B model (FP16):** ~6 GB.
- **1B model (FP16):** ~2 GB.

If DeepSeek-R1 is larger (e.g., **13B**):

- **13B FP16:** ~26 GB.
- **13B 4-bit:** ~6.5 GB.

---

3. Key Notes

1. **Check the Repo:** Always verify the Hugging Face model card:
   - Look for **quantized variants** (e.g., `-4bit`, `-8bit`, `-GPTQ` in filenames).
   - Example: `DeepSeek-7B-4bit` might be ~4 GB.
2. **Overhead:** Add ~0.5-1 GB for tokenizers, configs, and metadata.
3. **Hardware Limits:**
   - **RAM/VRAM:** To *load* a 7B FP16 model, you need ~14 GB of RAM/VRAM. For 4-bit, ~4 GB suffices.
   - **Laptop Feasibility:** Quantized 7B/13B models work on modern laptops with 16+ GB RAM.

Example Workflow for Local Use

1. Download a **4-bit quantized version** (smallest size).
2. Use libraries like `llama.cpp`, `AutoGPTQ`, or Hugging Face `transformers` + `bitsandbytes`.
3. For DeepSeek-R1, replace "7B" with its actual parameter count in the calculations above.

Let me know if you need help finding specific quantized variants!

New chat

Message DeepSeek

DeepThink (R1)   Search

AI-generated, for reference only