Title: Small language model

URL: https://en.wikipedia.org/wiki/Small_language_model

PageID: 77621280

Categories: Category:Language modeling, Category:Natural language processing stubs, Category:Statistical natural language processing, Category:Statistics stubs

---

Small language models or compact language models are artificial intelligence language models designed for human natural language processing including language and text generation . Unlike large language models , small language models are much smaller in scale and scope.

Typically, an large language models's number of training parameters is in the hundreds of billions, with some models even exceeding a trillion parameters. The size of any large language model is vast because it contains a large amount of information, which allows it to generate better content. However, this requires enormous computational power, making it impossible for an individual to train a large language model using just a single computer and graphical processing unit .

Small language models, on the other hand, use far fewer parameters, typically ranging from a few thousand to a few hundred million. This make them more feasible to train and host in resource-constrained environments such as a single computer or even a mobile device. [1] [2] [3] [4] [5]

Most contemporary (2020s) small language models use the same architecture as a large language model, but with a smaller parameter count and sometimes lower arithmetic precision. Parameter count is reduced by a combination of knowledge distillation and pruning . Precision can be reduced by quantization. Work on large language models mostly translate to small language models: pruning and quantization are also widely used to speed up large language models. Some notable models are: [2]

Below 1B parameters: Llama-Prompt-Guard-2-22M (detects prompt injection and jailbreaking, based on DeBERTa-xsmall), SmolLM2-135M, SmolLM2-360M

1–4B parameters: Llama3.2-1B, Qwen2.5-1.5B, DeepSeek-R1-1.5B, SmolLM2-1.7B, SmolVLM-2.25B, Phi-3.5-Mini-3.8B, Phi-4-Mini-3.8B, Gemma3-4B; closed-weights ones include Gemini Nano

4–14B parameters: Mistral 7B, Gemma 9B, Phi-4 14B.

(Phi-4 14B is marginally "small" at best, but Microsoft does market it as a small model.) [6]

Language model with small pre-training dataset

Traditional AI language systems need enormous computers and vast amounts of data.

Pre-training matters, even tiny models show significant performance improvements when pre-trained performance increases with larger pre-training datasets. Classification accuracy improves when pre-training and test datasets share similar tokens. Shallow architectures can replicate deep model performance through collaborative learning. [7]

See also

Edge computing

References

This statistics -related article is a stub . You can help Wikipedia by expanding it .

v

t

e

This article about natural language processing is a stub . You can help Wikipedia by expanding it .

v

t

e