Alberto Andrés Valdés González.
Degree: Mathematical Engineer.
Work position: Data Scientist.
Mail: anvaldes@uc.cl/alberto.valdes.gonzalez.96@gmail.com
Location: Santiago, Chile.

Small Language Models (SLM)

SLMs are essentially smaller versions of their LLM counterparts. They have significantly fewer parameters, typically ranging from a few million to a few billion, compared to LLMs with hundreds of billions or even trillions. This difference in size translates to several advantages:

- Efficiency: SLMs require less computational power and memory, making them suitable for deployment on smaller devices or even edge computing scenarios. This opens up opportunities for real-world applications like on-device chatbots and personalized mobile assistants.
- Accessibility: With lower resource requirements, SLMs are more accessible to a broader range of developers and organizations. This democratizes AI, allowing smaller teams and individual researchers to explore the power of language models without significant infrastructure investments.
- Customization: SLMs are easier to fine-tune for specific domains and tasks. This enables the creation of specialized models tailored to niche applications, leading to higher performance and accuracy.

How do Small Language Models Work?

Like LLMs, SLMs are trained on massive datasets of text and code. However, several techniques are employed to achieve their smaller size and efficiency:

- Knowledge Distillation: This involves transferring knowledge from a pre-trained LLM to a smaller model, capturing its core capabilities without the full complexity.
- Pruning and Quantization: These techniques remove unnecessary parts of the model and reduce the precision of its weights, respectively, further reducing its size and resource requirements.
- Efficient Architectures: Researchers are continually developing novel architectures specifically designed for SLMs, focusing on optimizing both performance and efficiency.

Benefits and Limitations

Small Language Models (SLMs) offer the advantage of being trainable with relatively modest datasets. Their simplified architectures enhance interpretability, and their compact size facilitates deployment on mobile devices.

A notable benefit of SLMs is their capability to process data locally, making them particularly valuable for Internet of Things (IoT) edge devices and enterprises bound by stringent privacy and security regulations.

However, deploying small language models involves a trade-off. Due to their training on smaller datasets, SLMs possess more constrained knowledge bases compared to their Large Language Model (LLM) counterparts. Additionally, their understanding of language and context tends to be more limited, potentially resulting in less accurate and nuanced responses when compared to larger models.

Examples:

- DistilBERT
- Orca 2
- Phi 2
- BERT Mini, Small, Medium, and Tiny
- GPT-Neo and GPT-J
- MobileBERT
- T5-Small
