M05e - SLMs Small Language Models

1 Einleitung

- Es existiert keine einheitliche Definition für SLMs. Einige definieren sie als Modelle mit weniger als einer Milliarde Parametern, während andere den Begriff relativ zu Large Language Models (LLMs) sehen.
- Der "Survey" schlägt eine generalisierte Definition vor: "Given specific tasks and resource constraints" werden Modelle als SLMs betrachtet. Modelle für mobile Geräte mit ca. 6GB Speicher haben oft unter einer Milliarde Parameter. Andere klassifizieren Modelle bis zu 10 Milliarden Parameter als klein, da ihnen "emergent abilities" fehlen.
- "Hugging Face" definiert SLMs typischerweise mit einer Spanne von **1 Million bis 10 Milliarden Parametern**.
- Im Gegensatz dazu haben LLMs Hunderte von Milliarden bis Billionen von Parametern ("Wikipedia", "Computer Weekly").

2 Vorteile von SLMs

- **Effizienz und Kosteneffizienz:** SLMs benötigen weniger Rechenleistung, Speicher und Energie sowohl für das Training als auch für die Inferenz ("Survey", "Forbes", "Kleine aber fein", "IT Reseller"). Das Hosting ist günstiger ("Kleine aber fein").
- **Flexibilität und Anpassbarkeit:** SLMs lassen sich besser an spezifische Aufgaben und Domänen anpassen ("Survey", "IT Reseller").
- **Lokaler Einsatz und Datenschutz:** SLMs ermöglichen den Einsatz auf Geräten (Edge Devices, Mobilgeräte) und lokalen Servern, was den Datenschutz verbessert und die Abhängigkeit von Cloud-Infrastrukturen reduziert ("Survey", "Forbes", "Kleine aber fein", "IT Reseller", "Computer Weekly"). Zitat aus "IT Reseller": "ermöglichen Echtzeitverarbeitung und verbesserten Datenschutz ohne grosse Abhängigkeit von Cloud-Infrastruktur."
- **Geringere Halluzinationen in spezifischen Kontexten:** Im Kontext von Retrieval-Augmented Generation (RAG) neigen SLMs möglicherweise weniger zu Halluzinationen als LLMs, da sie stärker auf den bereitgestellten Kontext fokussieren ("Kleine aber fein"). Zitat: "da die einfach nicht so viel Wissen haben s sind die einfach viel stärker auf okay ich hab die PDF ich schaue mir die PDF an und dementsprechend nehme ich nur die Inhalte der PDF als Vorgabe."

3 Nachteile und Herausforderungen von SLMs

- **Geringeres Allgemeinwissen:** SLMs sind schlechter im Zero-Shot Prompting, da sie weniger Wissen speichern können ("Kleine aber fein").

- **Potenzielle Leistungsgrenzen:** Für sehr komplexe, domänenübergreifende Aufgaben können LLMs überlegen sein ("Computer Weekly"). Zitat: "Sie können ihre größeren Pendants jedoch nicht immer ersetzen."
- **Herausforderungen beim Betrieb eigener SLMs:** Der Vortrag "KI:edu.nrw" thematisiert auch "die Herausforderungen beim Betrieb von eigenen SLMs".

4 Techniken zur Erstellung und Verbesserung von SLMs

- **Pruning (Beschneidung):** Reduzierung der Parameteranzahl durch Entfernen weniger wichtiger Verbindungen.
- **Knowledge Distillation (Wissensdestillation):** Übertragung von Wissen von einem großen (Lehrer-)Modell auf ein kleineres (Schüler-)Modell.
- **Quantisierung:** Reduzierung der numerischen Präzision der Modellgewichte und Aktivierungen. Beispiele: SqueezeLLM, QLoRA, BitNet.
- **Fortgeschrittene Trainingstechniken:** Innovative Methoden zur effizienteren Schulung von SLMs von Grund auf.
- **Distillationstechniken zur Verbesserung:** GKD, DistiLLM, Adapt-and-Distill.
- **Performanceverbesserung durch Quantisierung:** LLM.int8(), PB-LLM, OneBit.
- **Techniken aus LLMs, die SLMs zugutekommen:** RAG für SLMs, MoE (Mixture of Experts) für SLMs.

5 Anwendungsbereiche von SLMs

- **Aufgabenspezifische Anwendungen:** Frage-Antwort-Systeme (QA), Textgenerierung, etc.
- **Domänenspezifische SLMs:** Modelle für Medizin (Hippocrates, BioMedLM), Finanzen (MindLLM), Chemie (ChemLLM), etc. Zitat aus "Survey": "achieving performance comparable to LLMs for domain-specific problems".
- **SLMs für LLMs:** Unterstützung von LLMs in Bereichen wie zuverlässige Generierung, Prompt-Extraktion, Feinabstimmung, Anwendung und Evaluation. Beispiele: Kalibrierung der LLM-Konfidenz, Halluzinationsdetektion, Verbesserung von RAG.
- **Personalisierte Empfehlungssysteme:** RecLoRA nutzt SLMs/LLMs mit personalisierten LoRA-Gewichten.
- **Mobile Task Automation:** AutoDroid nutzt SLMs zur Steuerung von Android-Apps.
- **Betrugserkennung im Finanzdienstleistungsbereich:** Europäische Banken setzen lokale SLMs zur Transaktionsüberwachung ein ("Computer Weekly").
- **Vertragsanalyse im Rechtswesen:** Anwaltskanzleien nutzen SLMs zur Überprüfung von Vertraulichkeitsvereinbarungen ("Computer Weekly").
- **Netzwerkmanagement in der Telekommunikation:** SLMs werden in Netzwerkknoten zur Bedrohungserkennung eingesetzt ("Computer Weekly").

6 Beispiele für SLMs

- Llama 3.2 (bis zu 3B Parameter)
- Qwen (bis zu 7B Parameter in kleineren Varianten)
- Gemma
- StableLM
- TinyLlama
- Phi-3-mini (3.8B Parameter)
- DistilBERT
- OpenELM (bis zu 3B Parameter)
- MiniCPM

7 Wachsender Markt für SLMs

- "IT Reseller" berichtet, dass der weltweite Umsatz mit SLMs von 0,93 Milliarden Dollar im Jahr 2025 auf 5,45 Milliarden Dollar im Jahr 2032 ansteigen soll, was einem jährlichen durchschnittlichen Wachstum von 28,7 Prozent entspricht. Zitat: "Der Markt für sogenannte Small Language Models (SLMs) wächst. Und das nicht zu knapp."
- Das schnellste Wachstum wird bei Modellen mit weniger als zwei Milliarden Parametern erwartet.
- Der asiatisch-pazifische Raum soll die am schnellsten wachsende Region sein.

Trustworthiness (Vertrauenswürdigkeit) von SLMs

- Der "Survey" untersucht Trustworthiness-Aspekte von SLMs (bis ca. 7B Parameter), einschließlich Robustheit, Privatsphäre, Zuverlässigkeit (Halluzinationen, Sycophancy), Sicherheit (Misinformation, Toxicity) und Fairness.
- Viele bestehende Arbeiten zur Trustworthiness konzentrieren sich auf LLMs.
- Der "Survey" zitiert Benchmarks und Studien zur Evaluierung der Trustworthiness von LMs, darunter HELM, Do-Not-Answer, PromptRobust, HaluEval, PrivLM-Bench, FFT, ROBBIE, TrustLLM und andere.
- Einige Studien zeigen, dass kleinere LMs in Bezug auf Trustworthiness manchmal besser abschneiden als größere.
- Der "Survey" wirft Fragen für die zukünftige Forschung auf, z.B. wie die Trustworthiness bei der Kompression von LLMs zu SLMs erhalten werden kann und wie nicht-vertrauenswürdige SLMs durch Fine-Tuning robuster gemacht werden können.

Fazit

Small Language Models (SLMs) stellen eine vielversprechende Alternative und Ergänzung zu Large Language Models (LLMs) dar. Ihre Vorteile in Bezug auf Effizienz, Kosten, Flexibilität, lokalen Einsatz und Datenschutz machen sie besonders attraktiv für

ressourcenbeschränkte Umgebungen und spezifische Anwendungsfälle. Während LLMs in Bezug auf allgemeines Wissen und sehr komplexe Aufgaben überlegen sein können, zeigen SLMs in domänenspezifischen Bereichen und in Kombination mit Techniken wie RAG und Agentenarchitekturen bemerkenswerte Leistungen. Der wachsende Markt und die kontinuierliche Weiterentwicklung von Techniken zur Erstellung und Verbesserung von SLMs deuten auf eine zunehmende Bedeutung dieser Modelle in der Zukunft der KI hin. Die Auseinandersetzung mit den Möglichkeiten und Herausforderungen von SLMs, wie sie auch von Initiativen wie KI:edu.nrw gefördert wird, ist entscheidend, um ihr Potenzial voll auszuschöpfen.

Quellen:

- "A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Col."
- "KI:edu.nrw meets KI-NEL: Small Language Models – KI für die Hosentasche?"
- "KI:edu.nrw-Themenreihe - Ruhr-Universität Bochum"
- "Kleine aber fein: Small Language Models"
- "Scaling Small Language Models (SLMs) For Edge Devices: A New Frontier In AI - Forbes"
- "Small Language Models (SLM): A Comprehensive Overview - Hugging Face"
- "Small language model - Wikipedia"
- "Umsatz mit Small Language Models soll sich mehr als verfünffachen - IT Reseller"
- "Warum Small Language Models (SLM) auf dem Vormarsch sind - Computer Weekly"