M17c - Architekturformen

Anwendung Generativer KI

Stand: 05.2025

Künstliche Intelligenz hat in den letzten Jahren enorme Fortschritte gemacht – nicht zuletzt durch den Einsatz spezialisierter Modellarchitekturen. Drei Architekturformen stechen dabei besonders hervor: **Transformer**, **MoE (Mixture of Experts)** und **Diffusionsmodelle**. Sie bilden das technologische Rückgrat vieler moderner Anwendungen in Sprachverarbeitung, Bildgenerierung, Code-Assistenz und multimodaler KI.

1 Transformer – Die Grundlage moderner Sprachmodelle

Die Transformer-Architektur wurde 2017 mit dem bahnbrechenden Paper "Attention is All You Need" eingeführt und revolutionierte die Verarbeitung von Sequenzdaten wie Text. Im Gegensatz zu rekurrenten Netzwerken (RNNs), die Daten schrittweise verarbeiten, kann der Transformer alle Elemente einer Eingabesequenz gleichzeitig betrachten. Dies geschieht durch den Einsatz des sogenannten Self-Attention-Mechanismus, der Kontextinformationen über die gesamte Sequenz hinweg berücksichtigt.

Ein wesentlicher Vorteil ist die Möglichkeit zum parallelen Training, was die Effizienz beim Training großer Modelle erheblich steigert. Transformer können zudem besser mit langen Abhängigkeiten in Texten umgehen und sind flexibler in der Architekturgestaltung.

Kernkomponenten:

- **Self-Attention**: Berechnet die Wichtigkeit jedes Tokens im Kontext der gesamten Eingabe.
- **Positionskodierung**: Da der Transformer keine inhärente Reihenfolge kennt, werden Positionsinformationen addiert.
- **Encoder-Decoder-Struktur** (für Aufgaben wie maschinelle Übersetzung), **reine Decoder-Modelle** (z. B. GPT für Textgenerierung) oder **reine Encoder-Modelle** (z. B. BERT für Klassifikation und Verständnis).

Transformer sind heute die dominierende Architekturform für Sprachverarbeitung, Codierung, multimodale Aufgaben und sogar für spezielle Anwendungsgebiete wie die Vorhersage von Proteinstrukturen. Ihre Flexibilität und Skalierbarkeit haben sie zur Grundlage moderner KI-Anwendungen gemacht.

2 MoE (Mixture of Experts) – Effiziente Skalierung durch Spezialisten

MoE-Modelle (Mixture of Experts) setzen auf eine modulare Architektur, in der viele spezialisierte Subnetzwerke – sogenannte "Experten" – zur Verfügung stehen. Für jede Eingabe entscheidet ein sogenanntes **Gating-Modul**, welche wenigen Experten (z. B. 2 von 64) tatsächlich aktiviert werden. Damit wird die Rechenlast reduziert, ohne die Gesamtkapazität des Modells zu verringern.

Diese Architektur ermöglicht es, extrem große Modelle mit hoher Parameteranzahl effizient zu betreiben, da nur ein Bruchteil der Parameter für eine einzelne Inferenz genutzt wird. MoE erlaubt zudem eine gewisse Spezialisierung: Manche Experten werden häufiger für bestimmte Datenarten oder Aufgaben aktiviert, was zu einer besseren Gesamtauslastung und differenzierterem Lernen führen kann.

Vorteile:

- **Effizienz**: Geringerer Rechenaufwand durch selektive Aktivierung von Teilnetzwerken.
- **Skalierbarkeit**: Modelle mit mehreren Billionen Parametern werden realisierbar.
- **Spezialisierung**: Experten lernen unterschiedliche Aufgaben oder Datenmuster.

MoE-Architekturen werden vor allem in großskaligen Sprach- und Multimodellen eingesetzt. Google entwickelte mit Switch Transformer und GShard zwei prominente Beispiele. Auch GPT-4 wird mit hoher Wahrscheinlichkeit als MoE-Modell betrieben. Herausforderungen bestehen in der ausgewogenen Nutzung der Experten (Load Balancing), der Stabilität im Training und der effizienten Verteilung über Hardware-Infrastruktur.

3 Diffusionsmodelle – Hochwertige Bildsynthese durch Rauschen

Diffusionsmodelle stellen eine neuartige Herangehensweise zur Generierung komplexer Daten dar. Sie sind besonders bekannt für ihre Anwendung im Bereich der Bildsynthese, finden aber zunehmend auch Einsatz in Audio-, Video- und 3D-Generierung. Ihr Grundprinzip basiert auf zwei Phasen: einem Vorwärtsprozess, in dem ein Bild schrittweise verrauscht wird, und einem Rückwärtsprozess, in dem ein neuronales Netzwerk lernt, dieses Rauschen wieder zu entfernen.

Dabei erzeugt das Modell aus reinem Rauschen ein hochauflösendes und detailreiches Bild, das einem realistischen Eingabebild sehr nahekommt. Der Trainingsprozess ist stabiler als bei anderen generativen Ansätzen wie GANs und lässt sich gut kontrollieren.

Typischer Ablauf:

1. **Vorwärtsprozess**: Das ursprüngliche Bild wird über viele Schritte mit immer mehr Rauschen überlagert.
2. **Rückwärtsprozess**: Ein Modell wird trainiert, diese Rauschschritte rückgängig zu machen und die ursprünglichen Inhalte zu rekonstruieren.

Vorteile:

- **Hervorragende Bildqualität**, oft besser als bei GANs.
- **Stabile Trainingsdynamik** ohne die typischen Instabilitäten adversarieller Verfahren.
- **Hohe Flexibilität** für Anwendungen wie Inpainting, Text-zu-Bild-Generierung, Stilübertragungen oder sogar Animation.

Bekannte Modelle, die auf dieser Technik beruhen, sind **Stable Diffusion**, **DALL·E 2**, **Imagen** und viele neuere multimodale Generatoren. Diffusionsmodelle gelten als die vielversprechendste Richtung im Bereich generativer Medien.

4 Small Language Models (SLMs) – Kompakte KI für lokale Anwendungen

Small Language Models (SLMs) repräsentieren einen gegenläufigen Trend zur kontinuierlichen Skalierung der Modellgröße. Statt auf immer größere Parameteranzahlen zu setzen, fokussieren sich SLMs auf Effizienz, lokale Ausführbarkeit und spezifische Anwendungsfälle. Diese Modelle mit typischerweise weniger als 10 Milliarden Parametern bieten ein ausgewogenes Verhältnis zwischen Leistungsfähigkeit und Ressourcenbedarf.

Die Entwicklung dieser kompakten Modelle wurde durch Fortschritte in Kompressionstechniken, effizienteren Trainingsmethoden und architektonischen Optimierungen ermöglicht. Sie adressieren zentrale Herausforderungen großer Modelle wie hohe Betriebskosten, Latenzprobleme, Datenschutzbedenken und eingeschränkte Zugänglichkeit.

4.1 Wichtigste Techniken:

1. **Modellkompression**: Durch Verfahren wie Knowledge Distillation, Pruning und Quantisierung werden große Modelle kompakter gemacht, ohne signifikant an Leistung zu verlieren.
2. **Architekturoptimierung**: Anpassungen wie Sparse Attention, effiziente Aktivierungsfunktionen und optimierte Embedding-Schichten reduzieren den Rechenaufwand.
3. **Domänenspezifisches Training**: Fokussierung auf bestimmte Anwendungsgebiete statt Generalisierung über alle Domänen hinweg.
4. **Parameter-Effizienz**: Techniken wie LoRA (Low-Rank Adaptation) oder Adapter erlauben effiziente Feinabstimmung mit wenigen trainierbaren Parametern.

4.2 Vorteile:

- **Lokale Ausführung**: Können direkt auf Endgeräten ohne Cloud-Anbindung laufen.
- **Datenschutz**: Verarbeitung der Daten bleibt auf dem Gerät, was Privacy-by-Design ermöglicht.
- **Geringere Kosten**: Reduzierter Energie- und Ressourcenverbrauch bei Training und Inferenz.
- **Niedrigere Latenz**: Schnellere Antwortzeiten durch kompaktere Berechnungen.
- **Zugänglichkeit**: Demokratisieren KI-Technologie durch geringere Hardwareanforderungen.

Zu den bekanntesten Vertretern dieser Kategorie zählen TinyLlama (1,1B Parameter), Phi-2 (2,7B), Mistral-7B, Google Gemma (2B/7B) und FLAN-T5-Small. Diese Modelle zeigen, dass auch mit deutlich weniger Parametern beeindruckende Ergebnisse erzielt werden können, insbesondere bei spezifischen Aufgaben und nach effektivem Finetuning.

SLMs werden bereits erfolgreich in Bereichen wie mobilen Anwendungen, Embedded Systems, IoT-Geräten und in Umgebungen mit begrenzter Konnektivität oder hohen Datenschutzanforderungen eingesetzt. Sie stellen einen wichtigen Entwicklungszweig dar, der die praktische Anwendbarkeit von KI-Technologien wesentlich erweitert.

5 Zusammenfassung

Architektur	Einsatzgebiet	Kernidee	Vorteile	Beispiele
Transformer	Text, Sprache, Codierung, Multimodalität	Self-Attention zur parallelen Verarbeitung von Sequenzen	Paralleles Training, lange Abhängigkeiten, flexibel	GPT-4, BERT, T5, LLaMA
MoE (Mixture of Experts)	Großskalige Sprach- und Multimodalmodelle	Spezialisierte Teilmodelle, nur wenige pro Abfrage aktiv	Effizienz, Skalierbarkeit, Spezialisierung	Switch Transformer, GShard, GPT-4
Diffusionsmodell	Bild-, Audio-, Video-, und 3D-Generierung	Rauschen schrittweise in sinnvolle Daten zurückverwandeln	Hohe Bildqualität, stabile Trainingsdynamik, vielseitig	Stable Diffusion, DALL·E 2, Imagen
Small Language Models	Mobile Anwendungen, Edge Computing, Privacy-kritische Bereiche	Kompakte, effiziente KI-Modelle für lokale Ausführung	Datenschutz, geringe Latenz, Ressourceneffizienz, Zugänglichkeit	TinyLlama, Phi-2, Mistral-7B, Gemma

Fazit

Diese vier Architekturformen – Transformer, MoE, Diffusionsmodelle und Small Language Models – prägen heute maßgeblich die Entwicklung und Anwendung moderner KI-Systeme. Sie bilden das Fundament für Anwendungen in Sprachverarbeitung, Codegenerierung, Bildsynthese und multimodaler künstlicher Intelligenz und werden in der Forschung wie in der Industrie intensiv weiterentwickelt. Während Transformer, MoE und Diffusionsmodelle die Grenzen des technisch Machbaren erweitern, sorgen Small Language Models für die praktische Anwendbarkeit von KI-Technologien in ressourcenbeschränkten Umgebungen und tragen so zur Demokratisierung dieser wichtigen Zukunftstechnologie bei.