Scaling Laws

SOTA2 - Group C

Anton Kesy, Étienne Muser, Katharina Schindler, Lukas Fehrenbacher, Nico Ruschmann

Offenburg University of Applied Sciences

WS 2024/2025

Woher kommen Scaling Laws?

Leistung von Machine-Learning-Modellen, insbesondere von großen LLMs, vorhersagbar mit der ***Skalierung von Modellgröße, Datensatzgröße und Rechenaufwand zusammenhängt***

- aus der empirischen Beobachtung abgeleitet (GPT3 und dann DeepMind)
- zuerst Natural Language Processing und später auch in andere Bereiche (Bild-, Video- und Multimodal-Modellierung)

Warum sind Scaling Laws wichtig?

- Vorhersage der Modellleistung
  - Leistung von Modellen vorherzusagen, bevor sie trainiert werden
    - Planung von Ressourcen
- Optimierung der Ressourcennutzung
  - optimalen Rechenaufwand (Modellgröße & Datensatzgröße) für bestimmte Leistungsziel zu setzen
- Verständnis der Modellkapazität
  - Einblicke in die Komplexität der Daten und die Fähigkeit von Modellen
- Forschungsförderung
  - theoretischer Forschung, um die zugrunde liegenden Mechanismen dieser Skalierungsbeziehungen zu verstehen und neue Architekturen und Trainingsmethoden zu entwickeln
    - "Kommen wir noch weiter mit den aktuellen Methoden und Architekturen?"

Skalierungsgesetze für Transformer-Sprachmodelle

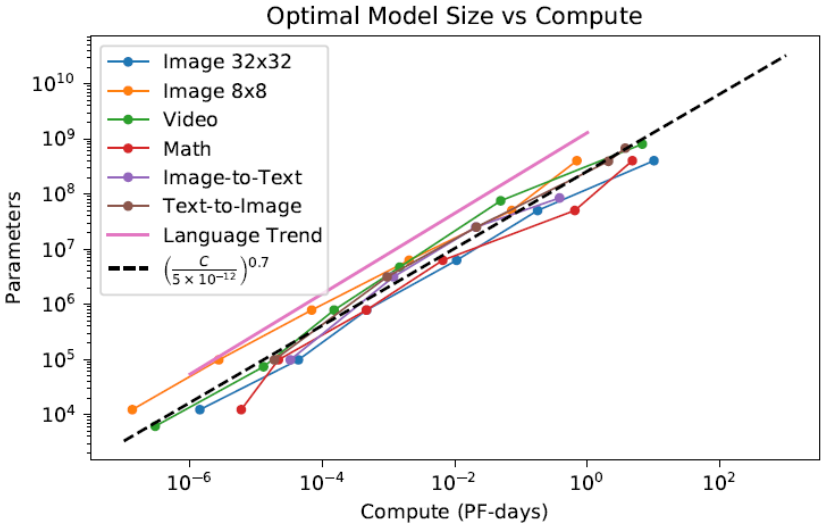
- Leistung hängt stark von Skalierung, schwach von Modellform ab
  - Anzahl Modellparameter N (ohne Embeddings)
  - Größe des Datensatzes D
  - Menge der Rechenleistung C, die für Training verwendet wird
- Leistung hat eine Potenzgesetzbeziehung mit jedem der drei Skalierungsfaktoren N, D, C, wenn sie nicht von den anderen beiden eingeschränkt wird
- Die Leistung verbessert sich vorhersehbar, solange N und D gemeinsam skaliert werden, aber sie tritt in einen Bereich mit abnehmenden Erträgen ein, wenn entweder N oder D konstant gehalten wird, während das andere erhöht wird
  - $\frac{N}{D} \approx \frac{8}{5} \Rightarrow$ Vermeidung von Leistungseinbußen

Optimale Modellgröße

Was ist die optimale Modellgröße? - Die optimale Modellgröße N_{opt} ist diejenige, die das Verhältnis von Leistung und Ressourcenverbrauch maximiert. - Ein Modell wird oft größer, um mehr Daten und Komplexität abzubilden, aber ab einer bestimmten Größe sinken die Vorteile (diminishing returns).

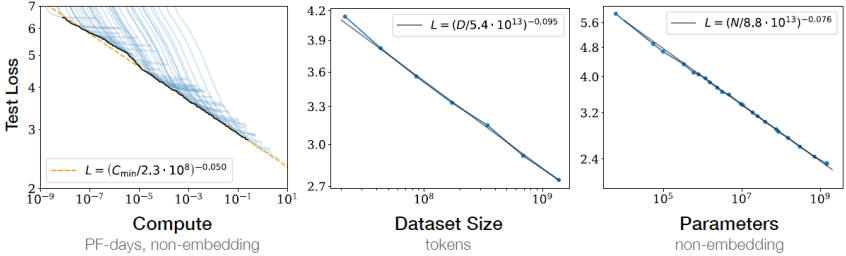
Wichtige Erkenntnisse aus den Scaling Laws: - Die Modellgröße N sollte sich gemäß einer **Power-Law-Beziehung** zum Rechenbudget C verhalten:
$N_{opt} \propto C^{0.7}$ für maximale Effizienz in allen getesteten Bereichen. - Dies gilt universell für **verschiedene Datenarten** (z. B. Bild, Text) und kann als Orientierung für die Skalierung genutzt werden.

Optimale Modellgröße



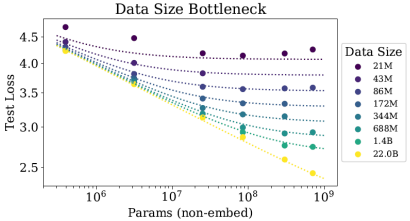
Faktoren, die die optimale Modellgröße bestimmen

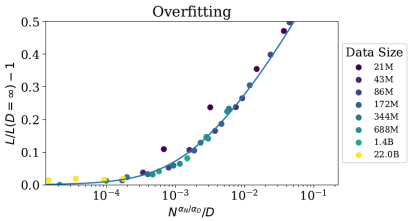
1. **Rechenbudget C**
   - Das Modell wird basierend auf dem verfügbaren Compute optimiert: Größere Budgets erlauben größere Modelle.
   - Die exponentielle Beziehung $N_{opt} \propto C^{0.7}$ zeigt, dass das Modell mit zunehmendem Compute langsamer skaliert (sublinear).



Faktoren, die die optimale Modellgröße bestimmen

2. **Aufgabenkomplexität:** Komplexere Aufgaben erfordern größere Modelle; einfache Probleme profitieren oft weniger von größerer Skalierung.
3. **Overfitting vermeiden:** Eine sublineare Skalierung der Datenmenge $D \propto N^{0.4}$ wird empfohlen, um Overfitting bei zunehmender Modellgröße zu minimieren.





Kritische Batch-Größe

- Batch größe, ab der das Training nicht mehr effizienter wird
- Unter BCrit: Je höher B, desto recheneffizienter ist das Training
- Über BCrit: Je höher B, desto weniger Steps nötig, aber nicht recheneffizienter
- BCrit nicht abhängig von Modellgröße, lediglich vom Loss

Verallgemeinerung auf andere Datensätze

- Transformer-Sprachmodelle weisen eine starke Fähigkeit zur Verallgemeinerung auf andere Datensätze auf
- Verallgemeinerungsleistung verbessert sich auf anderen Datensätze mit **zunehmender Modellgröße gleichmäßig**
- Parallel zur Verbesserung auf dem Trainingsdatensatz

Dies bedeutet, dass größere Modelle nicht nur auf den Daten, auf denen sie trainiert wurden, besser abschneiden, sondern auch besser auf neue, **unbekannte Daten generalisieren** können.

Verallgemeinerung

- **Skalierungsgesetze**: Die Leistung von Sprachmodellen, einschließlich ihrer Generalisierungsfähigkeit, verbessert sich mit zunehmender Größe, Datenmenge und Rechenleistung.
- **Unabhängigkeit von der Trainingsdauer**: Die Generalisierung hängt nicht von der Trainingsdauer oder der Nähe zur Konvergenz ab. Ein Modell mit niedrigem Verlust auf dem Trainingsdatensatz generalisiert wahrscheinlich gut, selbst bei frühzeitigem Abbruch des Trainings.
- **Validierungsverlust als Indikator**: Die Generalisierung hängt hauptsächlich vom Validierungsverlust innerhalb der Verteilung ab.Ein Modell mit guter Leistung auf den Trainingsdaten schneidet in der Regel auch auf anderen Datensätzen gut ab.
- **Geringe Abhängigkeit von der Modelltiefe**: Die Modelltiefe hat keinen signifikanten Einfluss auf die Generalisierung. Sowohl flache als auch tiefe Modelle können bei gleicher Parameterzahl ähnliche Generalisierungseigenschaften aufweisen.

Reducible and Ireducible Loss

- **Ireducible Loss**(Feste Grenze):
  - **Domänenabhängige Konstante** -> kann **nicht** durch **Skalierung** des Modells **reduziert** werden
  - Entropie der zugrunde liegenden Datenverteilung (S(True))
    - spiegelt Zufälligkeit & Komplexität der Daten wider

- **Reducible Loss**(Skalierbar):
  - **skaliert mit** Modellgröße, Datensatzgröße und Rechenaufwand
  - schätzt die KL-Divergenz zwischen der wahren Datenverteilung und der vom Modell vorhergesagten Verteilung
  - repräsentiert die Menge an Informationen, die das Modell noch nicht über die Daten gelernt hat

Irreducible Loss Beispiel

Beispiel:

Die Vorlesung SOTA-II ist ...

1. ... super!

Irreducible Loss Beispiel

Beispiel:

Die Vorlesung SOTA-II ist ...

1. ... super!
2. ... toll.

Irreducible Loss Beispiel

Beispiel:

Die Vorlesung SOTA-II ist ...

1. ... super!
2. ... toll.
3. ... gut.

Irreducible Loss Beispiel

Beispiel:

Die Vorlesung SOTA-II ist ...

1. ... super!
2. ... toll.
3. ... gut.
4. ... im Wintersemester des Studiengangs INFM an der Hochschule Offenburg zu finden, welche sich im südwesten Deutschlands befindet und neben den Informatikstudiengängen auch viele weitere Studiengänge in den Bereichen Biologie, Medien, Wirtschaft und weiteren anbietet.