Replication of "Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention"

Viet-Anh Do and Johannes Bergner

Version 1.0.2, 14.2.2022

1 Introduction

Transformers introduced by (Vaswani et al., 2017) are the state-of-the-art in tasks like language understanding and image processing. However their quadratic memory complexity within the self-attention mechanism weakens the efficiency when dealing with long sequences. Consequently a number of so called "X-$former$"- models have been proposed, that improve the original Transformer in terms of computational and memory efficiency.

Next to other approaches, like Performer (Choromanski et al., 2020), Linformer (Wang et al., 2020) and Big Bird (Zaheer et al., 2020) the authors of "Transformers are RNNs"developed a Linear Transformer (Katharopoulus et al., 2020) to show how the quadratic complexity $O(N^2)$ can be reduced to a linear complexity of $O(N)$ without reducing the accuracy.

Standard Transformers use the folllowing attention matrix:

$$Attention(Q, K, V){=}V{=}softmax\left(\frac{QK^T}{\sqrt{D_k}}\right)V, \tag{1}$$

with the queries Q $\in \mathbb{R}^{N\times D_k}$, keys K $\in \mathbb{R}^{M\times D_k}$ and values V $\in \mathbb{R}^{N\times D_v}$, where N and M represent the lenghts of queries and keys (or values) and D_k and D_v the dimensions of the keys (or queries) and values. The softmax function is applied rowwise to QK^T.

(Katharopoulus et al., 2020) simplify this attention mechanism by generalizing equation (1) and replacing the softmax with the kernel feature map $\phi(x){=}elu(x) + 1$:

$$V_i'{=}\frac{\phi(Q_i)^T \sum_{j=1}^{i} \phi(K_j)V_j^T}{\phi(Q_i)^T \sum_{j=1}^{i} \phi(K_j)} \tag{2}$$

Because of that $\sum_{j=1}^{i} \phi(K_j)V_j^T$ and $\sum_{j=1}^{i} \phi(K_j)$ can be computed once and reused for every query, which leads to a time and memory complexity of $O(N)$. Furthermore by representing these cumulated sums as S_i and Z_i, they can be seen as states of an RNN for causal attention and computed from S_{i-1} and Z_{i-1} in constant time.

In tests the Linear Transformer of (Katharopoulus et al., 2020) performed an image generation task based on the CIFAR-10 dataset over 4,000 times faster than the normal transformer with

the same accuracy. These results seem extraordinary, which is why the Linear Transformer has been recognized in further scientific research and the paper "$Transformers\ are\ RNNs$" has been recited about 250 times in nearly two years.

Nevertheless papers that benchmark the improvements in the Transformer architecture,like (Yi et al., 2020) show, that Linear Transformer perform faster then the original Transformer but not as high as 4,000 times faster. In this replication we will try to replicate a Linear Transformer in Julia in order to examnie if five-digit speed improvements are really possible.

2 Umfang der Replikation/Reproduktion

3 Methoden

3.1 Modellbeschreibung

Beischreiben sie die Modelle, die im Originalartikel genutzt werden, einschließlich der Architektur, der Zielfunktion und der Parameter.

3.2 Datenbeschreibung

Beschreiben sie die Datenmengen die sie genutzt haben und wie sie sie bekommen haben.

3.3 Hyperparameter

Beschreiben sie, wie sie Hyperparameter gesetzt haben. Welche Quellen haben sie für die konkreten Werte genutzt (z.B. den Forschungsartikel, Code oder sie hatten eine wohlbegründete Vermutung, educated guess).

3.4 Implementierung

Beschreiben sie, ob sie vorhandenen Code oder eigenen Code genutzt haben. Stellen sie Links zum Code bereit und beschreiben sie welche Programmiersprachen und Pakete genutzt wurden. Ihr Github oder Gitlab-Repositittory sollte öffentlich sein. Das Repositittory sollte klar dokumentiert werden.

3.5 Aufbau der Experimente

Erklären sie, wie sie ihre Experimente durchgeführt haben. Was für Ressourcen haben sie verwendet, z.B. GPU/CPU-Ressourcen. Verlinken sie ihren Code und Notebooks.

3.6 Ressourcen für Berechnungen

Beschreiben sie die Anforderungen für die Berechnungen für jedes ihrer Experimente, z.B. die Anzahl der CPU/GPU-Stunden oder die Voraussetzungen für den Hauptspeicher und GPU-Speicher. Geben sie für Zeit und Speicher eigene Abschätzungen an, bevor die Experimente gelaufen sind und vergleichen sie dies mit den tatsächlich verbrauchten Ressourcen. Sie müssen vor den Experimenten einplanen, dass diese Informationen auch durch ihren Code gemessen und gespeichert werden.

4 Ergebnisse

Starten sie mit einem Überblick über die Ergebnisse. Bestätigen ihre Ergebnisse die aufgeführten Behauptungen? Dieser Abschnitt sollte hauptsächlich Fakten nennen und so präzise wie möglich geschrieben werden. Die Bewertung und Diskussion kann im späteren Kapitel "Diskussion" folgen.

Beschreiben sie dann detailliert jedes einzelne Ergebnis, das sie haben. Zeigen sie wie es mit einer oder mehreren Behauptungen in Beziehung steht. Erklären sie konkret was der Kern ihres Ergebnis ist. Gruppieren sie die Ergebnisse in logische Abschnitte. Beschreiben sie klar, wo sie über den Originalartikel hinausgegangen sind, wo sie zusätzliche Experimente durchgeführt haben und wie diese mit den ursprünglichen Behauptungen in Beziehung stehen.

Tipp 1: Drücken sie sich genau aus und verwenden sie eine klare und einfache Sprache, z.B.

> "we reproduced the accuracy to within 1% of reported value, that upholds the paper's conclusion that it performs much better than baselines." oder
>
> "We konnten die Klassifikationsrate bis auf 1% des angegebenen Werts reproduzieren. Dies unterstützt die Schlussfolgerung der Artikels, dass der Ansatz leistungsfähiger als die Baselines ist."

Oft kann man nicht die exakt gleiche numerische Zahl als Ergebnis bekommen. Deshalb müssen sie das Ergebnis bewerten, um zu entscheiden, ob ihr Ergebnis die Behauptung der Originalartikels unterstützt.

Tipp 2: Nutzen sie Tabellen und Abbildungen, um ihre Ergebnisse darzustellen.

4.1 Ergebnis 1

4.2 Ergebnis 2

4.3 Zusätzliche Ergebnisse, die nicht im Originalartikel enthalten waren

Beschreiben sie alle zusätzlichen Experimente, die über den Originalartikel hinausgehen. Dies können Experimente zu weiteren Datenmengen sein oder sie probieren andere Methoden bzw. weitere Vereinfachungen des Modells aus oder passen die Hyperparameter an. Beschreiben sie für jedes zusätzliche Experiment, was sie genau durchgeführt haben, was die Ergebnisse sind und diskutieren sie was diese Ergebnisse zeigen.

5 Diskussion

Beschreiben sie die weiterführenden Implikationen der experimentellen Ergebnisse. War der Originalartikel replizierbar bzw. reproduzierbar. Falls nicht, welche Faktoren haben dazu geführt, dass die Experimente nicht reproduziert werden konnten.

Bewerten sie, ob sie die Evidenz, die sie durch das Durchführen der Experimente erhalten haben, auch überzeugt, dass die Behauptungen des Originalartikels dadurch gestützt werden. Diskutieren sie die Stärken und Schwächen ihres Ansatzes, vielleicht haben sie aus Zeitgründen nicht alle Experimente durchführen können, oder vielleicht haben zusätzliche Experimente durchgeführt, die den Originalartikel weiter stärken.

5.1 Was war einfach?

Beschreiben sie welche Teile der Replikation/Reproduktion sich leicht umsetzen ließen. Lief der Code der Autoren problemlos? War es aufgrund der Beschreibung im Originalartikel nicht

aufwändig die Methoden zu reimplementieren? Dieser Abschnitt soll den Lesenden zeigen, welche Teile des Originalartikels sich leicht für eigene Ansätze verwenden lässt.

Tipp: Machen sie keine pauschalen Verallgemeinerungen. Was für sie leicht ist, muss für andere nicht leicht sein. Geben sie genügend Kontext und erklären sie warum manche Sachen leicht waren, z.B. der Code hatte eine umfangreiche Dokumentation der Schnittstellen und viele Beispiele aus der Dokumentation passten zu den Experimenten im Artikel.

5.2 Was war schwer?

Beschreiben sie welche Teile ihrer Replikation/Reproduktion aufwändig oder schwierig waren oder viel mehr Zeit in Anspruch genommen haben, als sie erwarteten. Vielleicht waren Daten nicht verfügbar, so dass sie einige Experimente nicht verifizieren konnten, oder der Code der Autoren funktionierte nicht und musste erst debugged werden. Vielleicht dauerten auch einige Experimente zu lange und sie konnten sie deshalb nicht verifizieren. Dieser Abschnitt soll den Lesenden zeigen, welche Teile des Originalartikels schwer wiederverwendbar sind, bzw. signifikante Zusatzarbeiten und Ressourcen erfordern.

Tipp: Setzen sie sorgfältig ihre Diskussion in den richtigen Kontext, z.B. sagen sie nicht " die Mathematik war schwer verständlich" sondern sagen sie " die Mathematik erfordert fortgeschrittene Kenntnisse in Analysis für das Verständnis".

5.3 Empfehlungen für die Replizierbarkeit / Reproduzierbarkeit

Geben sie Empfehlungen, wie die Autoren des Originalartikels oder andere Forschende in diesem Feld die Replizierbarkeit / Reproduzierbarkeit verbessern können.

6 Kommunikation mit den Autoren

Dokumentieren sie das Ausmaß (oder das Fehlen) der Kommunikation mit Autoren. Stellen sie sicher, dass der Bericht eine faire Beurteilung der Forschungsarbeiten ist. Versuchen sie deshalb mit den Autoren Kontakt aufzunehmen. Sie können ihnen konkrete Fragen stellen oder falls sie keine Fragen haben, den Bericht zusenden und um Feedback bitten.