Online FORM Compiler: Usage Insights and AI-Assisted Code Generation

May 30, 2024

Bakar Chargeishvili

II. Institut für Theoretische Physik
Universität Hamburg

FORM and Symbolica developers meeting
Zürich, May 29-31, 2024

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

PRELUDE

- Last year in Madrid an online implementation of FORM compiler was presented
- https://capp.uni-hamburg.de
- Key features:
  - Online IDE
  - Syntax highlighting
  - Code completion using snippets
  - Real time shared sessions
  - Form exercises, automatic assessment
  - Public API endpoint: https://capp.uni-hamburg.de/api
- The portal was successfully used at CAPP2023: https://indico.desy.de/event/CAPP2023

Prelude: API

Following `RESTful API` is being provided:

- Request:

```
curl \
  -X POST -H "Content-Type:  multipart/form-data" \
  -H "Authorization:  Bearer <AUTHENTICATION TOKEN>" \
  --data-binary @YourFile.frm \
  "https://capp.uni-hamburg.de/api"
```

- Response:

```
{
  "success": <Success status>,
  "output": <FORM Output>
}
```

Prelude: Real life use case for the API

Interactive code snippets for the static webpages:

- Including the script below in your HTML source enables you to present the FORM code snippets interactively:
  - https://capp.uni-hamburg.de/FORM/FORMSnippet.js

- Live example at:
  - `https://capp.uni-hamburg.de/snippets.html`
  - Have a look at the HTML source of the page above for more details

INTERLUDE

Anonymous AI

INTERLUDE

Anonymous AI

When there is a free service, you are the product.

— Folklore

Interlude: Data harvesting

During the last year these services were passively collecting the data:

- 725 total visitors

Country	Visitors
Germany	201
United States	99
Russia	87
China	79
France	58
Italy	44
UK	36
Japan	29
Canada	22
India	20
. . .	. . .
Switzerland	4
. . .	. . .
Others	47

Interlude: Data harvesting

- 63% bounce rate (visitors who left the website without doing anything)
- 262 visitors wrote some code
- 45% of first compilations were erronous
- 17% left the website after the first compiler error
- It took a visitor on average 3 trials to successfully run a code
- 162 valid `FORM` files were harvested
  - Average file length: 20 lines
  - $\sim$ 5000 lines of code in total
- The most popular keyword is `id`, followed by `if`, followed by `.sort`

Interlude: What can we do with this data?

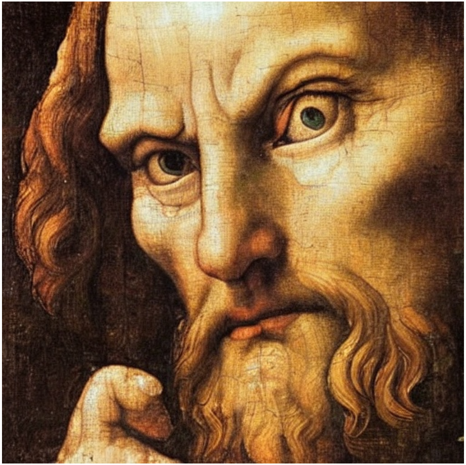
The best way to learn a language is to immerse yourself in it.

— *Unknown*

- Analyzing how users interact with the product and understanding their intentions can provide valuable insights to improve the product

- Can it help us liberate from the inherent problems inherent in the nature of our calculations?

TRIBULATION



Anonymous AI

As goods increase, so do those who consume them.
— Ecclesiastes 5:11

Perspective on theory calculations in pQCD

"In the next couple years yet another experiment will reach unprecedented accuracy, hence it is necessary to achieve higher precision in ...".

— arXiv:YourFavoriteNumber [hep-ph]

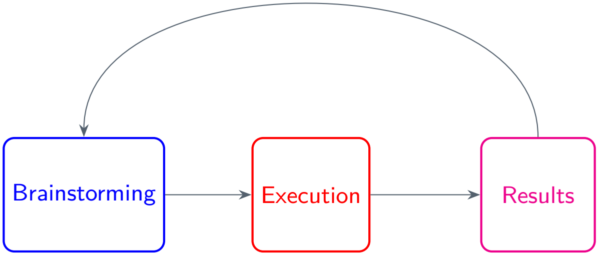
- Each advancement in precision is accompanied by the challenge of making further advancements increasingly difficult.

In the broader spectrum:

- As the difficulties of the problem increase, the effort required also increases, but the payoff of scientific benefits diminish

- Is it possible to systematically improve the situation?

Yet another day in the life of a theoretical researcher



- The Brainstorming part is essential
- The Results are primary motivation
- What role does the execution part play?
    - Major source of instability and inefficiency

Execution step for coding challenges

A typical problem in theoretical particle physics often does not require:

- Invention of new algorithms (from the point of view of Computer Science)

It does require:

- Adopting of known techniques to the particular problem

A typical programming task consists off:

- Trying to remember if you already solved something similar
- If yes, trying to find the existing code (in you head or in a personal database) and adopting it
- If no or if the problems were encountered in the step before, searching for the solutions in the manual and on the web

If carried out by humans, the steps above are susceptible to various imperfections inherent to human nature.

When carried out by computers, the efficiency can be increased in a controlled and predictable manner.

CRESCENDO

Anonymous AI

Qui non proficit, deficit. (He who does not advance, goes backward.)

— Latin proverb

Chatbots/Language models: state-of-the-art

Commercial:

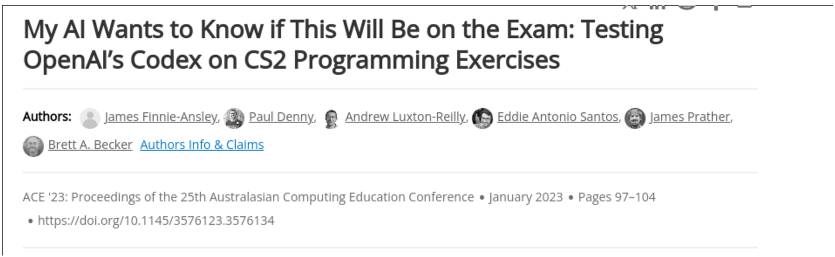
- OpenAI ChatGPT-4, ChatGPT-4o
- OpenAI/Microsoft Copilot, Codex
- Google Gemini
- Anthropic Claude3 opus
- ...

Prices vary between 10 – 30 € for the web access, unforeseeable (much higher) for the API access.

Free (?):

- Meta LLama2, Llama3 (open-source)
  - Alpaca (open-source, but commercial)
  - Vicuña (open-source)
- OpenAI ChatGPT 3.5
- Anthropic Claude3 nano, haiku
- ...

Chatbots: state-of-the art



My AI Wants to Know if This Will Be on the Exam: Testing OpenAI's Codex on CS2 Programming Exercises

Authors: James Finnie-Ansley, Paul Denny, Andrew Luxton-Reilly, Eddie Antonio Santos, James Prather, Brett A. Becker Authors Info & Claims

ACE '23: Proceedings of the 25th Australasian Computing Education Conference • January 2023 • Pages 97–104

• https://doi.org/10.1145/3576123.3576134

In early 2023 a Chatbot could perform better than 60% of students in the CS2 programming exam.

Chatbots: How do they work

A vast majority of large language models are based on transformers [arXiv:1706.03762].

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

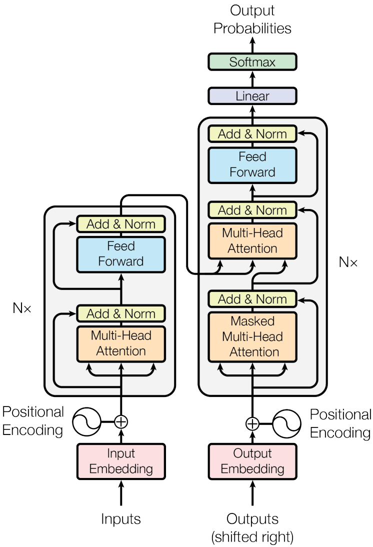
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

A paper written in 2017 with +100 000 citations.

Transformers



Provides the possibility to parallelize the training on huge datasets with (theoretically) unlimited context size.

Transformers for FORM: tokenization

The code is tokenized by the following rule:

1. Each keyword is registered as a separate token
2. Each shortcut is registered as a separate token
3. The whole Latin alphabet including special charachters are registered as a separate tokens

The tokenization follows in a greedy way starting from 1. going to 3.

Example:

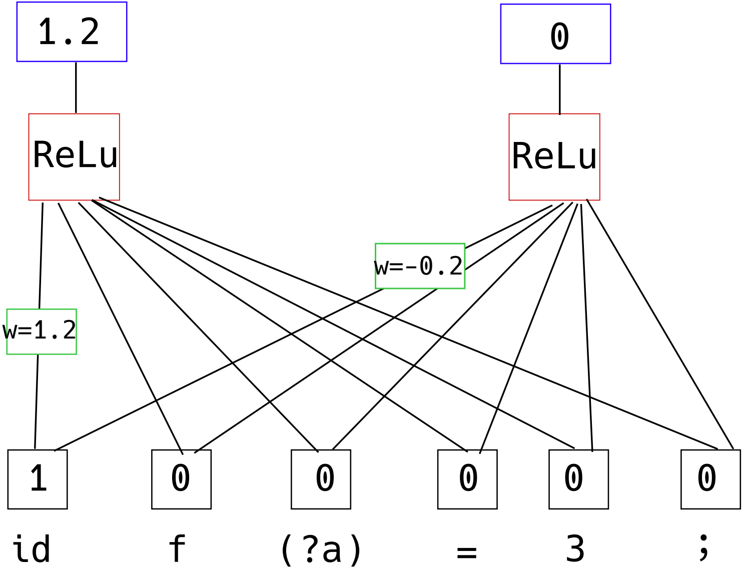
```
id f(?a) = 3;
```

After tokenization:

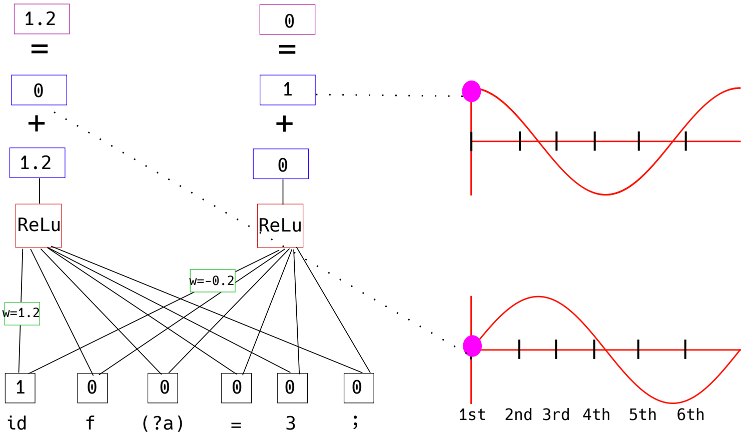
```
455 0 6 31 34 31 0 36 0 42 32
```

Our vocabulary contains 595 tokens.

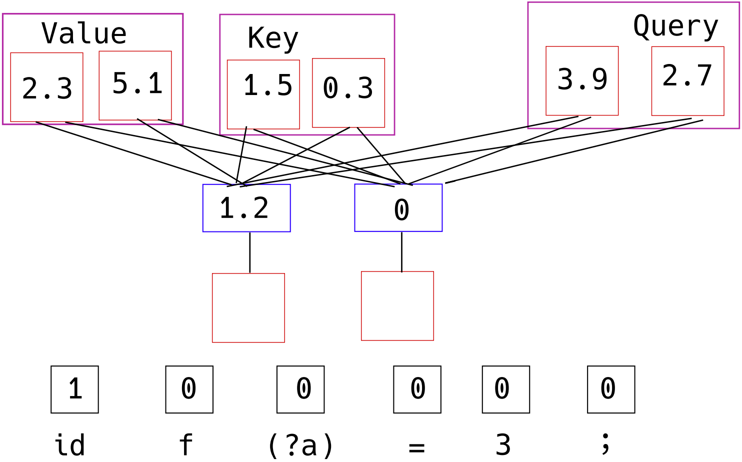
Transformers for FORM: Embedding



Transformers for FORM: Positional Encoding



Transformers for FORM: Attention



Transformers for FORM: Attention

$$\text{similarity}_i = \text{query} \cdot \text{key}_i$$

The similarity scores are mapped on probabilities using the softmax function defined as:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \,.$$

For example:

z_i	$\sigma_{(z_i)}$
-1	0.002
0	0.006
3	0.118
5	0.874

- The values of previous tokens are scaled using these probabilities
- The scaled values are added in pairs to produce a self-attention score of the given token
- The self-attention scores are unembedded by the inverse of the embedding matrix

LIBERATION

Anonymous AI

Befreier-1

- The FORM data was scrapped from all possible sources
- In total $\sim 50\,000$ lines of code was acquired.
- 90% of it was used in training, 10% for validation.
- A model with 50 000 parameters was trained with the context size of 100 tokens for 100 epochs.
- See the results in the attached video file.

Outlook

- First semi-stable language model of FORM was presented.
- Further improvements require:
    - Acquiring more FORM data.
    - Optimizing the generative model.
    - Scaling up the training.
- Investigate the possibility of fine-tuning of the existing open-source large language models.

Thanks for your attention!

UHH

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG