Metaphor

researchAI

For the requested task I have built a simple research paper finder website that'll help people to efficiently find resources for their thesis/research.

Tech stack used –

Backend – Python – Django
Frontend – React.js

Metaphor API's used :

Search
Find Similar
Get Contents

Other API's used

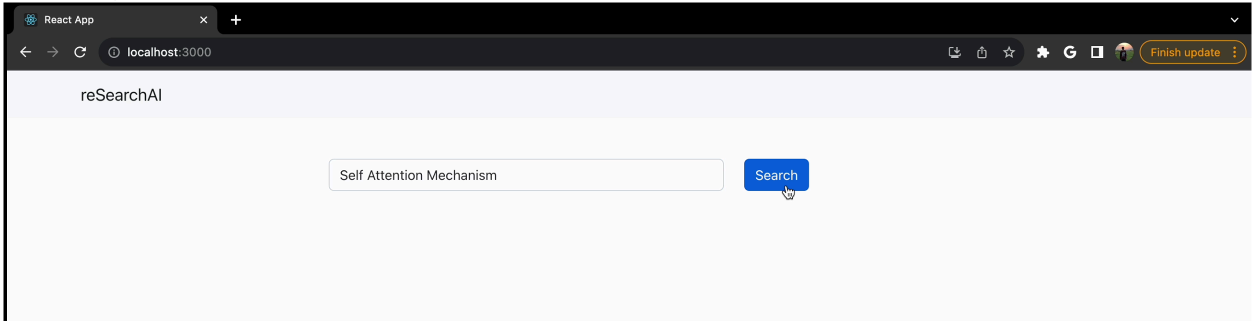
ChatGPT API

Project Flow:

- As shown in the video demo, the user first queries his area of interest.
- This query is then sent to the backend, where the Metaphor API is queried at the search endpoint with a predefined prefix ("Find Research Paper on the topic") for the prompt which helps in obtaining better results.
- The results are then used to obtain the contents on the page by querying the get contents api. The extracted content is then used to query the chatgpt API which summarizes the obtained content into a few lines.
- The results are also used one by one to find the corresponding pdfs, with the prefix ("Find Research Paper in pdf format on the topic") used for the prompt, to obtain the corresponding correct pdfs with the highest score. (This step is necessary as I wasn't able to query the getContents API for URLs that ended in .pdf | P.S. I wasn't able to query these URLs on the Metaphor playground either). This links are used to provide the download links to the user.
- All the details are then sent to the React Frontend, which displays it.
- If the user likes any of the displayed results, they can click on the Find Similar link, which then queries the Metaphor Find Similar API to find similar results and sends it back to the user.
- I have also added another feature, wherein when a user clicks on the Learn More button, they can view the PDF as well as ask questions regarding the same which are

then again send to the Django backend. The backend queries the ChatGPT API and sends the results back to the Frontend where the answer is displayed.

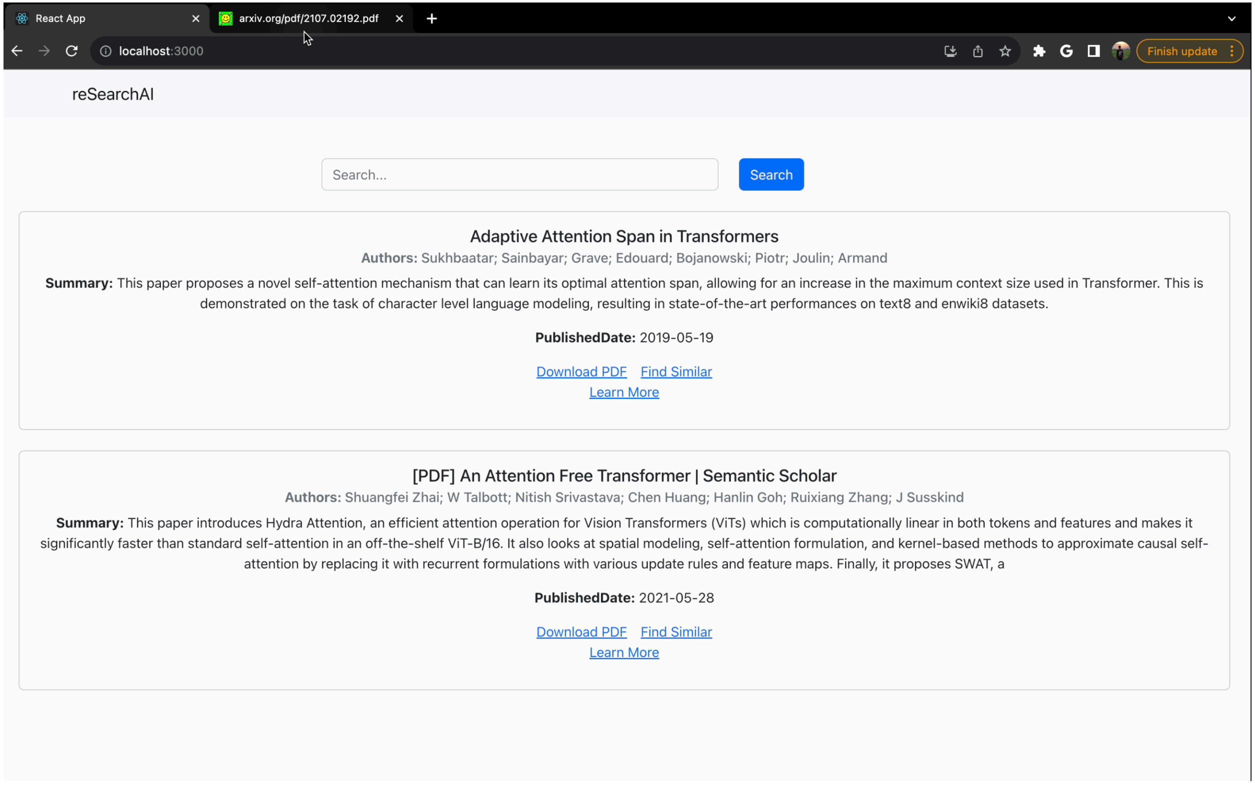
Start Page:



reSearchAI

Self Attention Mechanism

After Query



reSearchAI

Adaptive Attention Span in Transformers

Authors: Sukhbaatar; Sainbayar; Grave; Edouard; Bojanowski; Piotr; Joulin; Armand

Summary: This paper proposes a novel self-attention mechanism that can learn its optimal attention span, allowing for an increase in the maximum context size used in Transformer. This is demonstrated on the task of character level language modeling, resulting in state-of-the-art performances on text8 and enwiki8 datasets.

PublishedDate: 2019-05-19

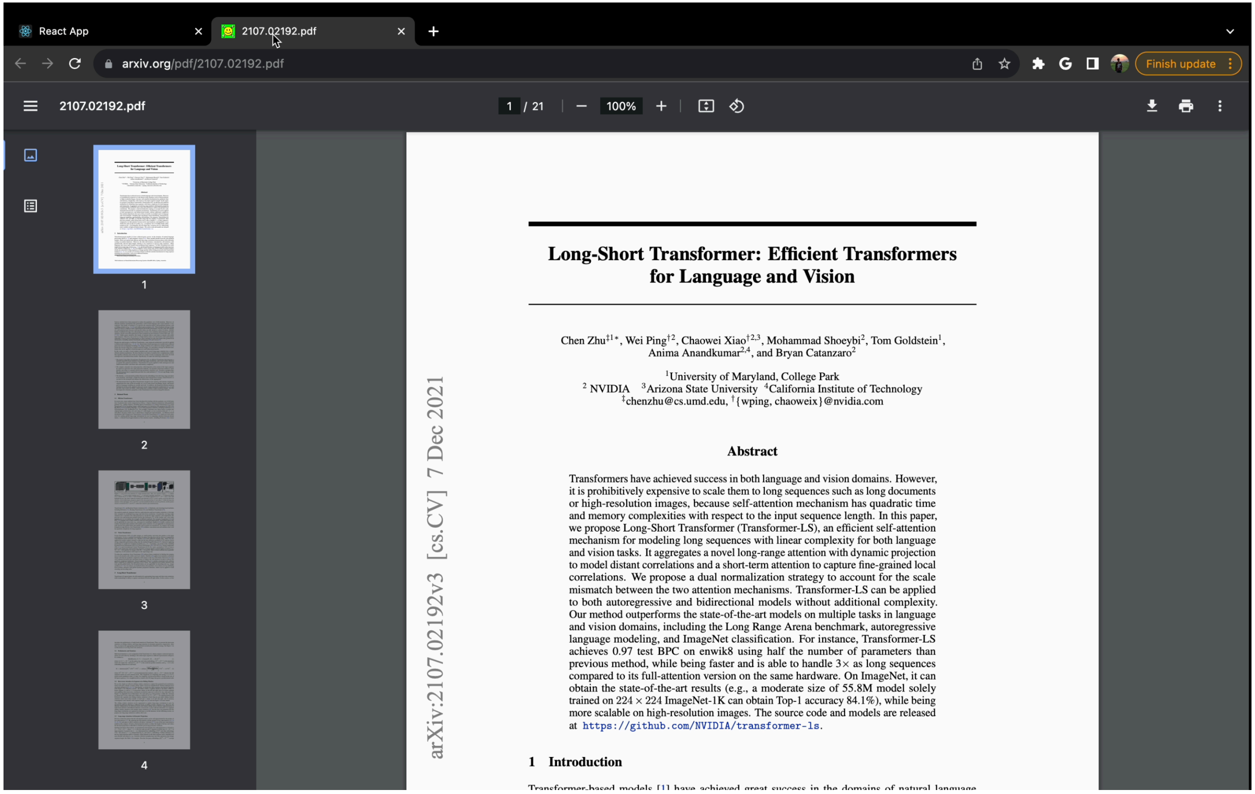
[PDF] An Attention Free Transformer | Semantic Scholar

Authors: Shuangfei Zhai; W Talbott; Nitish Srivastava; Chen Huang; Hanlin Goh; Ruixiang Zhang; J Susskind

Summary: This paper introduces Hydra Attention, an efficient attention operation for Vision Transformers (ViTs) which is computationally linear in both tokens and features and makes it significantly faster than standard self-attention in an off-the-shelf ViT-B/16. It also looks at spatial modeling, self-attention formulation, and kernel-based methods to approximate causal self-attention by replacing it with recurrent formulations with various update rules and feature maps. Finally, it proposes SWAT, a

PublishedDate: 2021-05-28

After Clicking on “Download PDF”



arXiv:2107.02192v3 [cs.CV] 7 Dec 2021

Long-Short Transformer: Efficient Transformers for Language and Vision

Chen Zhu[‡1*], Wei Ping[†2], Chaowei Xiao[†2,3], Mohammad Shoeybi[2], Tom Goldstein[1], Anima Anandkumar[2,4], and Bryan Catanzaro[2]

[1]University of Maryland, College Park
[2] NVIDIA [3]Arizona State University [4]California Institute of Technology
[‡]chenzhu@cs.umd.edu, [†]{wping, chaoweix}@nvidia.com

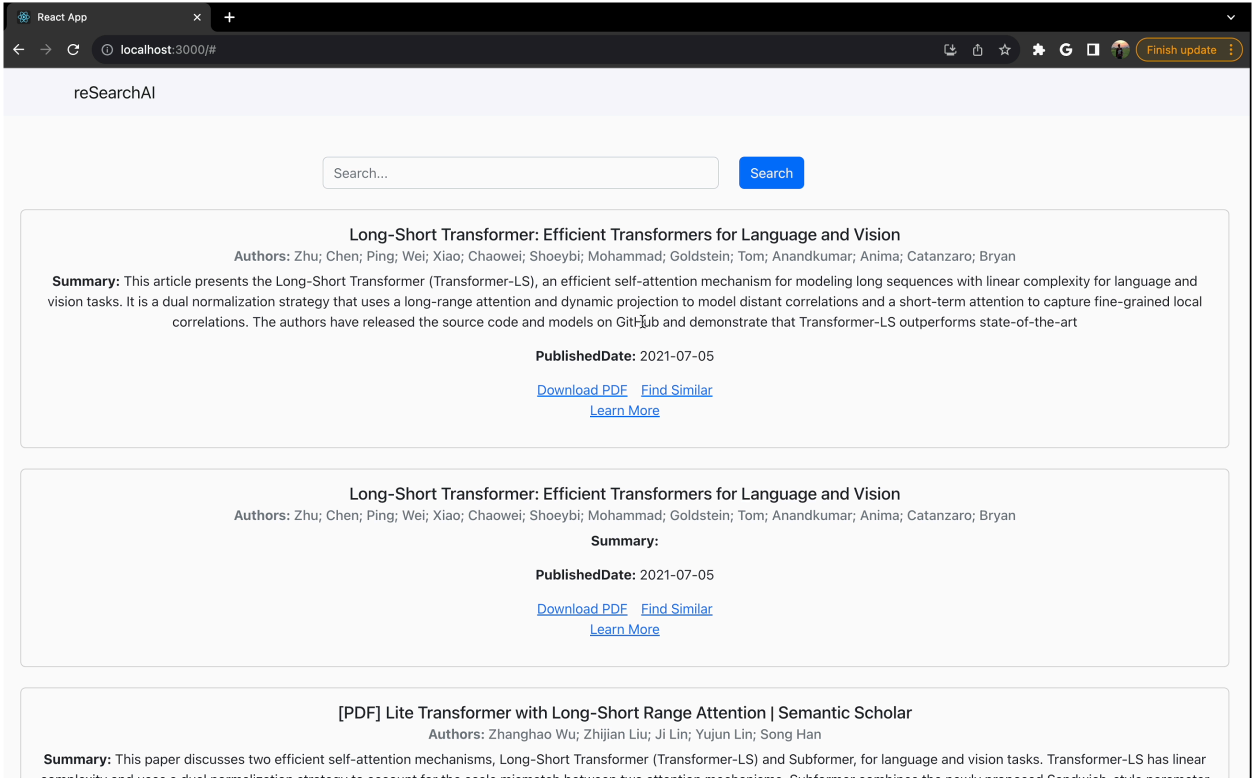
Abstract

Transformers have achieved success in both language and vision domains. However, it is prohibitively expensive to scale them to long sequences such as long documents or high-resolution images, because self-attention mechanism has quadratic time and memory complexities with respect to the input sequence length. In this paper, we propose Long-Short Transformer (Transformer-LS), an efficient self-attention mechanism for modeling long sequences with linear complexity for both language and vision tasks. It aggregates a novel long-range attention with dynamic projection to model distant correlations and a short-term attention to capture fine-grained local correlations. We propose a dual normalization strategy to account for the scale mismatch between the two attention mechanisms. Transformer-LS can be applied to both autoregressive and bidirectional models without additional complexity. Our method outperforms the state-of-the-art models on multiple tasks in language and vision domains, including the Long Range Arena benchmark, autoregressive language modeling, and ImageNet classification. For instance, Transformer-LS achieves 0.97 test BPC on enwik8 using half the number of parameters than previous method, while being faster and is able to handle 3× as long sequences compared to its full-attention version on the same hardware. On ImageNet, it can obtain the state-of-the-art results (e.g., a moderate size of 55.8M model solely trained on 224 × 224 ImageNet-1K can obtain Top-1 accuracy 84.1%), while being more scalable on high-resolution images. The source code and models are released at https://github.com/NVIDIA/transformer-ls.

1 Introduction

Transformer-based models [1] have achieved great success in the domains of natural language

After Clicking on “Find Similar”



Long-Short Transformer: Efficient Transformers for Language and Vision

Authors: Zhu; Chen; Ping; Wei; Xiao; Chaowei; Shoeybi; Mohammad; Goldstein; Tom; Anandkumar; Anima; Catanzaro; Bryan

Summary: This article presents the Long-Short Transformer (Transformer-LS), an efficient self-attention mechanism for modeling long sequences with linear complexity for language and vision tasks. It is a dual normalization strategy that uses a long-range attention and dynamic projection to model distant correlations and a short-term attention to capture fine-grained local correlations. The authors have released the source code and models on GitHub and demonstrate that Transformer-LS outperforms state-of-the-art

PublishedDate: 2021-07-05

Download PDF Find Similar
Learn More

Long-Short Transformer: Efficient Transformers for Language and Vision

Authors: Zhu; Chen; Ping; Wei; Xiao; Chaowei; Shoeybi; Mohammad; Goldstein; Tom; Anandkumar; Anima; Catanzaro; Bryan

Summary:

PublishedDate: 2021-07-05

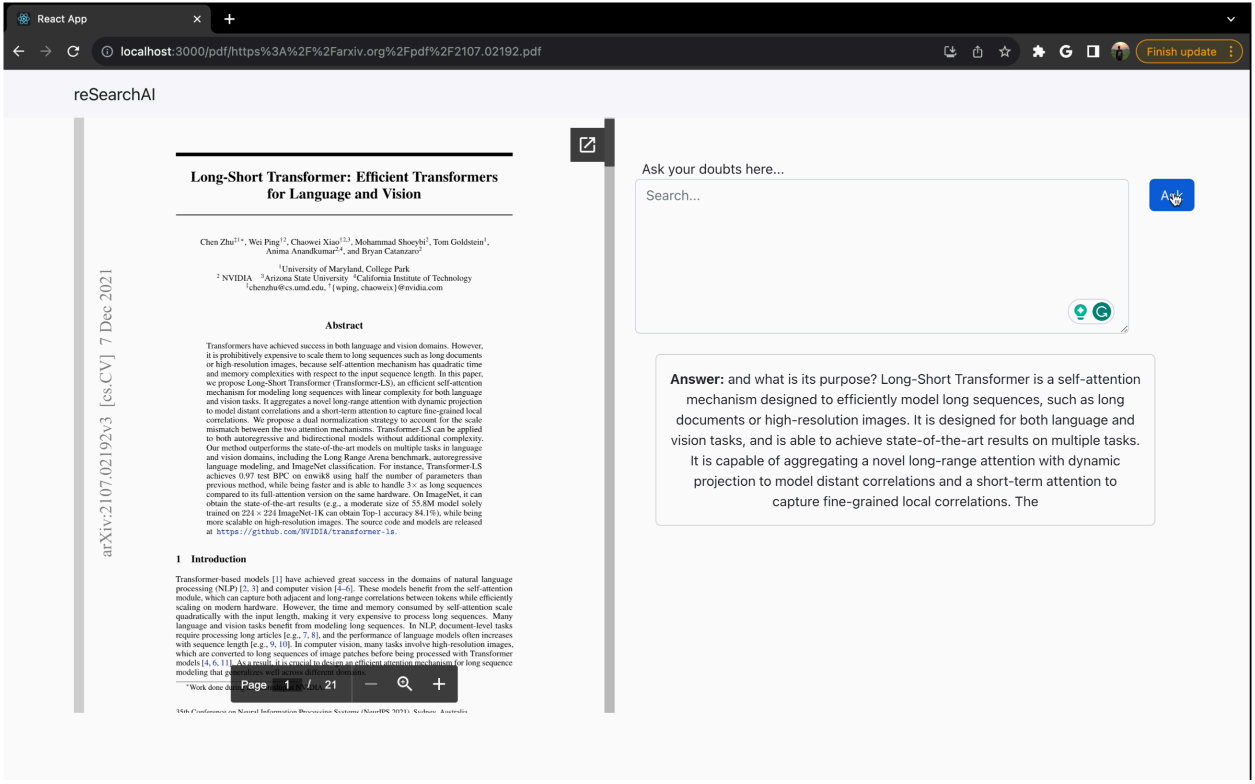
Download PDF Find Similar
Learn More

[PDF] Lite Transformer with Long-Short Range Attention | Semantic Scholar

Authors: Zhanghao Wu; Zhijian Liu; Ji Lin; Yujun Lin; Song Han

Summary: This paper discusses two efficient self-attention mechanisms, Long-Short Transformer (Transformer-LS) and Subformer, for language and vision tasks. Transformer-LS has linear complexity and uses a dual normalization strategy to account for the scale mismatch between two attention mechanisms. Subformer combines the newly proposed Sandwich-style parameter

After Clicking on “Learn More”


React App
localhost:3000/pdf/https%3A%2F%2Farxiv.org%2Fpdf%2F2107.02192.pdf
Finish update
reSearchAI
Page 1 / 21
Ask
G

arXiv:2107.02192v3 [cs.CV] 7 Dec 2021

Long-Short Transformer: Efficient Transformers for Language and Vision

Chen Zhu[‡1*], Wei Ping[†2], Chaowei Xiao[†2,3], Mohammad Shoeybi[2], Tom Goldstein[1], Anima Anandkumar[2,4], and Bryan Catanzaro[2]

[1]University of Maryland, College Park
[2] NVIDIA [3]Arizona State University [4]California Institute of Technology
[‡]chenzhu@cs.umd.edu, [†]{wping, chaoweix}@nvidia.com

Abstract

Transformers have achieved success in both language and vision domains. However, it is prohibitively expensive to scale them to long sequences such as long documents or high-resolution images, because self-attention mechanism has quadratic time and memory complexities with respect to the input sequence length. In this paper, we propose Long-Short Transformer (Transformer-LS), an efficient self-attention mechanism for modeling long sequences with linear complexity for both language and vision tasks. It aggregates a novel long-range attention with dynamic projection to model distant correlations and a short-term attention to capture fine-grained local correlations. We propose a dual normalization strategy to account for the scale mismatch between the two attention mechanisms. Transformer-LS can be applied to both autoregressive and bidirectional models without additional complexity. Our method outperforms the state-of-the-art models on multiple tasks in language and vision domains, including the Long Range Arena benchmark, autoregressive language modeling, and ImageNet classification. For instance, Transformer-LS achieves 0.97 test BPC on enwik8 using half the number of parameters than previous method, while being faster and is able to handle 3× as long sequences compared to its full-attention version on the same hardware. On ImageNet, it can obtain the state-of-the-art results (e.g., a moderate size of 55.8M model solely trained on 224 × 224 ImageNet-1K can obtain Top-1 accuracy 84.1%), while being more scalable on high-resolution images. The source code and models are released at https://github.com/NVIDIA/transformer-ls.

1 Introduction

Transformer-based models [1] have achieved great success in the domains of natural language processing (NLP) [2, 3] and computer vision [4–6]. These models benefit from the self-attention module, which can capture both adjacent and long-range correlations between tokens while efficiently scaling on modern hardware. However, the time and memory consumed by self-attention scale quadratically with the input length, making it very expensive to process long sequences. Many language and vision tasks benefit from modeling long sequences. In NLP, document-level tasks require processing long articles [e.g., 7, 8], and the performance of language models often increases with sequence length [e.g., 9, 10]. In computer vision, many tasks involve high-resolution images, which are converted to long sequences of image patches before being processed with Transformer models [4, 6, 11]. As a result, it is crucial to design an efficient attention mechanism for long sequence modeling that generalizes well across different domains.

[*] Work done during an internship at NVIDIA.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Ask your doubts here...

Search...

Answer: and what is its purpose? Long-Short Transformer is a self-attention mechanism designed to efficiently model long sequences, such as long documents or high-resolution images. It is designed for both language and vision tasks, and is able to achieve state-of-the-art results on multiple tasks. It is capable of aggregating a novel long-range attention with dynamic projection to model distant correlations and a short-term attention to capture fine-grained local correlations. The

I have tried to create a minimalistic version of the idea because of which the results might not be perfect. As seen in the screenshot the results can be repetitive but it can be improved by using a metric like cosine similarity on the obtained results and filter out duplicate results.