

Revolutionizing Research: Unleashing Creativity with Yi 34B and Langchain

Raghavan Muthuregunathan

Loom: <https://www.loom.com/share/a9d151a59a7c4fe68a8cc94b6a118d9d?sid=90a5d924-d995-4716-8a3b-c4da861e2e1d>

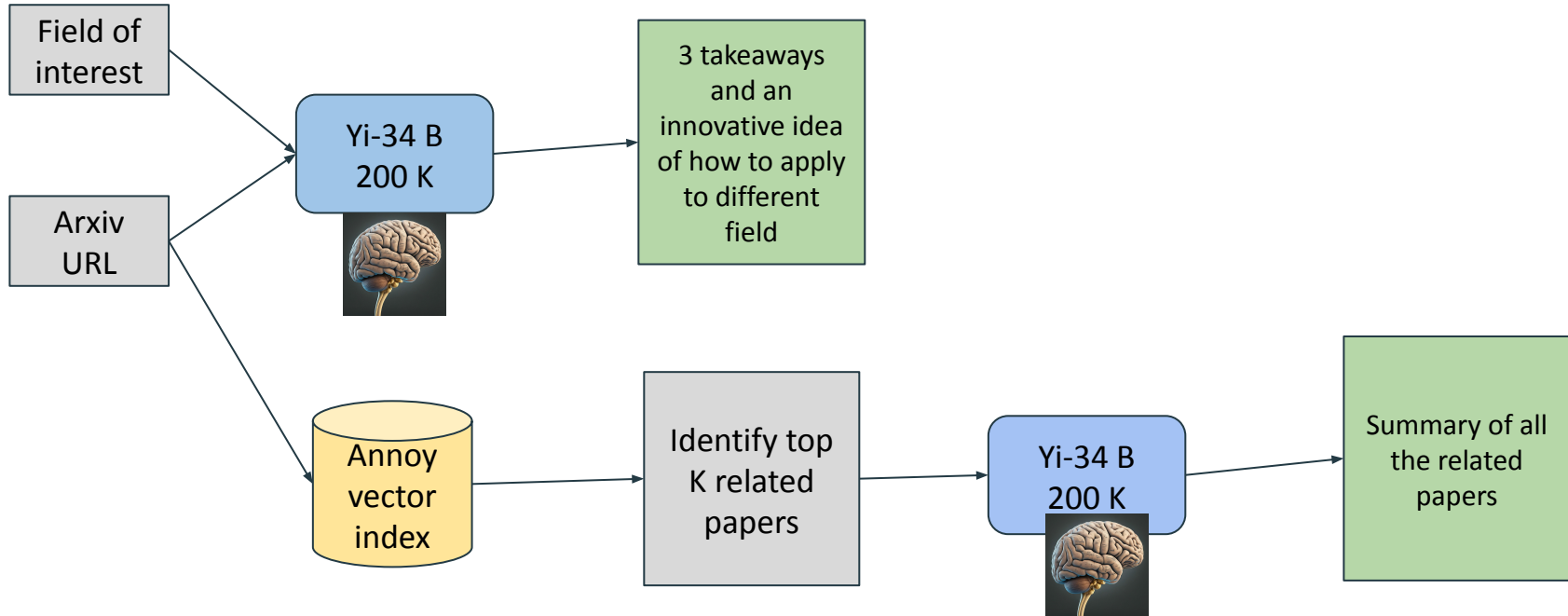
Problem

- 100s of new arxiv papers are uploaded everyday
- How do I get new ideas to apply them for my field ?
- How do i get a quick gist of the paper and related papers?

Loom video of Demo:

<https://www.loom.com/share/a9d151a59a7c4fe68a8cc94b6a118d9d?sid=90a5d924-d995-4716-8a3b-c4da861e2e1d>

Yi with its mighty 200K token to the rescue



Demo:
URL

<https://arxiv.org/abs/2310.01082>

Paper: linear attention is (may be) all
you need
field: **urban planning**

- Yi34B generated 4 key
takeaway by reading the
entire paper
- Generated an idea for “urban
planning”
- Summarized all the papers of
2023

← → 🔍 127.0.0.1:5000

Summary from Yi 34B

[Linear attention is \(may be\) all you need to understand transformer optimization](#)

Key Points From Research Article:

- Simplified Transformer Model:** The researchers introduce a simplified version of the transformer architecture called the “linear transformer” which retains some essential characteristics of the original transformer while being easier to analyze.
- Training Dynamics Reproduction:** By training linear transformers on regression tasks, they find that these simplified models can replicate many of the observed dynamics and challenges associated with training full transformers.
- Heavy-Tailed Gradient Noise:** Similar to what’s seen in actual transformer training, the linear transformer exhibits heavy-tailed and non-Gaussian gradient noise, challenging traditional optimization theories.
- Ill-Conditioned Loss Landscape:** The loss landscapes encountered during linear transformer training are found to be ill-conditioned, similar to those reported in complex transformer settings.

Applying These Insights to Urban Planning:

Idea: Developing Linear Models for Predicting Urban Growth Patterns

Urban planners often face complex decision-making processes involving predicting growth patterns, transportation demands, environmental impacts, etc. Traditional methods may not account for the intricacies of real-world data, leading to suboptimal outcomes.

By applying insights from the study of linear transformers, urban planners could develop simplified models that capture the essential relationships between variables influencing urban development. Such models would allow planners to forecast changes under different scenarios without the computational complexity of large-scale simulations.

The idea is to create linear models that are robust enough to handle the nonlinearities present in urban systems, akin to how the linear transformer captures the essence of the transformer’s behavior despite its simplicity. This approach could lead to faster, more accurate predictions that inform strategic urban planning decisions.

Summary of related work in 2023

Summary of Research Articles on Deep Learning and Related Topics

Result 0 - [id:2302.00722](#): A Survey of Deep Learning: From Activations to Transformers

- Authored by Johannes Schneider and Michalis Vlachos, this article provides a comprehensive review of deep learning developments since 2012.
- It covers a broad range of topics, including activations, normalizations, skip connections, transformers, and self-supervised learning methods.
- The authors aim to integrate and present influential works in these fields to foster new insights and connections within deep learning research.
- They discuss common patterns underlying many successful innovations and touch upon recently released commercial models like OpenAI’s GPT-4 and Google’s PaLM 2.

Result 1 - [id:2304.05133](#): Lecture Notes: Neural Network Architectures

- Written by Evelyn Herberg, these lecture notes offer a mathematical perspective on neural network architectures.
- They cover various types of neural networks, including feedforward, convolutional, ResNets, and recurrent neural networks.
- The notes serve as an educational resource for those seeking a fundamental understanding of how different neural network structures work.

Result 2 - [id:2304.02034](#): Effective Theory of Transformers at Initialization

- Emily Dinan, Sho Yaida, and Susan Zhang explore the initial configuration of Transformer models, which are variants of residual neural networks with multi-head self-attention blocks and MLP blocks.
- Their work proposes specific scaling relationships for the widths of initializations and training hyperparameters based on an effective theory analysis of signal propagation in wide and deep Transformers.
- The authors apply these findings to Vision and Language Transformers, demonstrating improved performance in practical settings.

Result 3 - [id:2307.13421](#): On the Learning Dynamics of Attention Networks

- Rahul Vaichait and Harsh G. Ramaswamy examine the learning dynamics of attention models trained using different loss functions—soft attention, hard attention, and latent variable marginal likelihood (LVML) attention.
- Each method approaches the selection of relevant parts of the input differently, leading to distinct dynamic behaviors during training.
- The authors analyze these differences and propose a hybrid approach that combines aspects of all three paradigms, showing promise across synthetic and real-world datasets.

Result 4 - [id:2302.13696](#): Moderate Adaptive Linear Units (MoLU)

- Hankyu Koh, Joon-hyuk Ko, and Wonho Jhe introduce the Moderate Adaptive Linear Unit (MoLU), a new activation function designed to improve the efficiency and speed of deep neural network training.
- MoLU is constructed from basic elements and is characterized by being an infinite diffeomorphism, meaning it is smooth and continuously differentiable throughout its domain.
- The authors claim that MoLU leads to shorter training times due to its simple construction.

Result 5 - [id:2306.00008](#): Brainformers: Trading Simplicity for Efficiency

- Led by Yangi Zhou, Nan Du, Yeping Huang, et al., this research introduces Brainformer, a complex block structure within Transformers that mixes various layer types and norms.
- Brainformer achieves superior performance and efficiency compared to traditional dense and sparse Transformers.
- Models with 8 billion activated parameters per token exhibit twice the training convergence rate and five times faster step times than counterparts, without compromising accuracy.

Result 6 - [id:2302.04342](#): Efficient Attention via Control Variates

- Lin Zheng, Jianbo Yuan, Cheng Wang, and Lingpeng Kong present a theoretical framework that characterizes the approximation gap between random-feature-based attention (RFA) and conventional softmax attention using control variates.
- This framework reveals that exact softmax attention can be approximated by manipulating individual control variates within RFA.
- Building on this concept, the authors devise a novel

urban planning ^ v Highlight All Match Case Match Diacritics Whole Words 1 of 2 matches

URL:

<https://arxiv.org/abs/2307.08691>

Paper is about flash attention-2 that is better parallelized

Yi34B summarized the 4 key points of the paper by reading the entire paper

Applied to the field of earthquake data processing and gave an actionable advice

Summarized all the related papers in 2023.

← → C 127.0.0.1:5000

Summary from Yi 34B

FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning

Key Points From Research Article "FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning":

- Memory Efficiency:** FlashAttention reduces the memory complexity of self-attention computations from quadratic to linear in the sequence length without any approximations. This allows for processing longer sequences within the limited memory capacity of GPUs.
- Speed Improvement:** Compared to traditional attention layers, FlashAttention provides a significant speedup ranging from 2x to 4x, which can be attributed to its exploitation of the asymmetric GPU memory hierarchy.
- Parallelization Enhancements:** FlashAttention-2 introduces improved work partitioning strategies that lead to higher warp and thread block level parallelism. This results in reduced latency and increased throughput during the attention computation.
- Closer Alignment with Matrix Multiply Performance:** By optimizing the distribution of workload among threads and minimizing non-matmul operations, FlashAttention-2 achieves closer to the theoretical peak floating point operations per second (FLOPS) achieved by highly optimized GEMM libraries. This brings the efficiency of attention calculations closer to that of fundamental matrix multiplication operations.

Application to Earthquake Analysis:

The principles behind FlashAttention-2 could potentially be applied to the field of earthquake analysis to accelerate seismic data processing tasks, particularly those involving large datasets or complex algorithms. For example, in seismic signal processing, where vast amounts of sensor data must be analyzed to detect and characterize earthquakes, the ability to process longer sequences efficiently could lead to improvements in detection accuracy and response times. Additionally, in the realm of machine learning for seismology, where deep learning models are trained on massive datasets to predict earthquake occurrences or to enhance seismic imaging, the enhanced parallelism and efficiency offered by FlashAttention-2 might enable larger batch sizes or longer sequence inputs, ultimately leading to more accurate and robust predictive models.

Summary of related work in 2023

Summary of Research Articles:

Result 0 - Latency-aware Unified Dynamic Networks for Efficient Image Recognition

- Key Insights:
 - Dynamic computation can optimize inference efficiency by activating computational units selectively.
 - Existing dynamic models often fail to meet theoretical efficiency expectations due to fragmentation in research, reliance on algorithms alone, and limitations in measuring practical latency.
 - "Latency-aware Unified Dynamic Networks" (LAUDNet) addresses these issues by integrating dynamic paradigms and optimizing scheduling strategies alongside algorithm design.
- LAUDNet demonstrates significant reductions in latency across various vision tasks when implemented on platforms like V100, RTX3090, and TX2 GPUs.
- Contributions:
 - Introduced LAUDNet, a framework that combines spatial adaptation, dynamic layer skipping, and dynamic channel skipping.
 - Developed a latency prediction technique to accurately gauge dynamic operator latency.
 - Open-sourced code at <https://www.github.com/LeapLabTHU/LAUDNet>.

Result 1 - High-Level Features Parallelization for Inference Cost Reduction through Selective Attention

- Key Insights:
 - Proposes a method to parallelize high-level features within deep networks to reduce inference costs.
 - Challenges exist in efficiently focusing on selected class-specific features without retraining.
 - Serial-Parallel Hybrid Architecture separates generic low-level features from parallelizable high-level features.
- Approach enables cutting out irrelevant portions of the network and adapting to changes in task complexity.
- Contributions:
 - Presented a method to selectively skip or activate class-specific features without retraining.
 - Demonstrated up to 75% parameter savings and 35% reduced GMACs compared to baseline methods.

Result 2 - WPC: Whole-Picture Workload Characterization

- Key Insights:
 - Identified a need for understanding the contribution of each critical component (languages, frameworks, runtimes, ISAs, OSs, microarchitectures) to bottlenecks in workload performance.
 - Proposed the Whole-Picture Workload Characterization (WPC) methodology as an iterative cycle involving observation, reference, fusion, and exploration.
- WPC tool quantitatively reveals the influence of each component across the technology stack on pipeline efficiency.
- Contributions:
 - Formulated a systemic methodology to guide software and hardware co-design and optimization.
 - Open-sourced the WPC tool.

Result 3 - SuperScaler: Supporting Flexible DNN Parallelization via a Unified Abstraction

- Key Insights:
 - Contemporary parallelization plan generators rely on heuristic rules that limit flexibility and optimal resource utilization.
 - SuperScaler introduces a principled approach that breaks down the parallelization process into distinct phases: transformation, scheduling, and maintaining data dependencies.
- Enables the construction of highly flexible parallelization plans, resulting in improved performance and efficiency for diverse DNN models.
- Contributions:
 - Proposed a system that connects parallelization plan generators to compiler models and runtime levels of task execution.

Applications

1. Researchers can get a quick summary of the paper
2. Researchers can get new ideas by applying the paper to a completely new field
3. Researchers can use the 200K limit and get summary of all the related papers too at the same time

Github

- https://github.com/Raghavan1988/vi34b200k_hackathon
- To run on terminal
 - Pip install -r requirements.txt
 - Python arxiv_summarizer_terminal.py <URL> <field>
- To run flask application
 - Pip install - requirements.txt
 - export FLASK_APP=arxiv_summarizer.py
 - Flask run

Both requires **replicate token**

export REPLICATE_API_TOKEN=<your TOKEN> i can provide the token for judges

Thank you

My linkedin: <https://www.linkedin.com/in/raghavanmit/>

Reach out to me for any questions.

Loom Video:

<https://www.loom.com/share/a9d151a59a7c4fe68a8cc94b6a118d9d?sid=90a5d924-d995-4716-8a3b-c4da861e2e1d>