Paper Summaries: Generative AI, Web Caching, Web-LEGO

Yerdos 202311628

Contents

1 Generative AI (WebDiffusion)

1.1 1. What is the problem that the paper mainly discusses?

This paper discusses about generative AI being used to generate content, particularly images. And this paper talks about Stabel Diffusion that would generate images based on the "alt" text of the "img" tag in the html file.

1.2 2. Why is the problem important?

Images is one of the main components of a webpage, moreover the paper provides result that shows that 44% of the total size of a median webpage is images. And automating image generation could be used by Web devs to speed up webpage creation, or by browser vendors in the future to address issues like repairing broken webpages, conserving bandwidth, and improving privacy.

1.3 3. Why is the problem difficult to solve?

- Implementing generative AIs for image generation directly in the client is difficult. (I did my own research on this part) Even the best GPU for today, RTX 5090 (32GB GDDR7 VRAM, 21760 cores) cannot be compared to Tesla GPUs. A40 came out 4.5 years ago, but has 48GB GDDR6, 10752 cores and specifically designed for this. And we should also consider that RTX 5090 rare among regular consumers. Which leads us to downloading images from the server will be faster and more reliable than generating in the client-side. The paper also provides that for iPhone-13 it took 15.6s to generate a single image using Stable Diffusion.
- And even with powerful GPUs (A40, A100) can partially compete with classic image downloads. Generating high-quality images leads us to GENERATIVE LEARNING TRILEMMA (high-quality output, sample diversity, fast/inexpensive sampling - can't choose all three). Even though Stable Diffusion is designed to be efficient, still requires interference time, which impacts PLT. And it can also have difficulties generating complex things such as faces, hands, images containing text, which in result affects the quality.
- And also not all webpages have "alt" text inside the "img" tag. This makes it hard to generate, and so broken image remains broken. The paper says that out of 500 webpages from Tranco's top one million, the success rate was only 260 webpages. These webpages are the ones that met the requirements.

1.4 4. Summarize the existing methodologies for the problem.

There are General Adversarial Networks (GANs) or Stable Diffusion Models being used to generate images. And the paper sticks with Stable Diffusion, because it is outperforming GANs for image generation. Standard Diffusion Models work by gradually adding noise to an input and then reversing this process to produce an image, but it requires many iterations for denoising.

Latent Diffusion Models, the foundation of Stable Diffusion, are designed to solve slowness. They work by operating on compressed version of the image in a latent space. These steps help to generate high-quality images quickly and efficiently. And Stable Diffusion fulfills these requirements, making it suitable for practical applications.

1.5 5. Describe the main ideas of the proposed scheme in this paper in your own words.

The paper proposes WebDiffusion. The main idea is to simulate Web environment where images are generated by Stable Diffusion rather than just being downloaded. And it is tested in two modes: server-mode, and client-side-mode. Server-mode for web designers, they use AI to generate images; client-side-mode is when images are generated locally in the browser. The tools works by having a CRAWLER visit real webpages, make local copies, and generate text descriptions (prompts) for the images on the pages. It does this in two ways: a "client-based" method that only uses available contextual text around image ("alt" text or context around it), and a "server-based" method that also uses an AI model to describe the original image itself. A Proxy system then intercepts requests for images when these saved pages are "replayed" and sends the generated textual prompts to a Stable Diffusion server to generate an image, which are then returned to the browser. And this all allows researchers to evaluate the quality and performance of AI-generated images within actual webpages without needing to build the AI directly into a browser. And Finally, WebDiffusion includes a web interface connected to a crowdsourcing platform to gather user opinions on the quality of the AI-generated images and webpages.

1.6 6. What are the notable findings in the evaluation results?

- Textual prompts automatically generated from webpage content were found to be relevant to the original images 80% of the time for the client-based approach, and nearly 90% for the server-based approach. Even strong AIs describing the original image wasn't 100% relevant according to users, which makes client-based approach's accuracy relatively good.
- AI-generated images were rated as "fair" or higher by users for 70% of images using the client-based approach and 95% using the server-based

approach.

- Simple images of food, landscape, and object got the highest scores, while complex images containing text scored lowest. Images with faces also got a low score, relative to simple images.
- AI-generated webpages were rated "good" or higher 95% of the times. Possibly because the users view the page as a whole rather than focusing on image details.
- Tesla GPUs can partially compete with traditional downloading with 2.5 5s improvements in some cases. But this approach significantly slows down the loading of the entire page (PLT), even on fast networks.
- Locally generated images can result in significantly saved bandwidth, with a median of 1.3MB per webpage and over 5MB for the heavier webpages.
- Tesla GPUs are relatively fast, with median 1.1s on A40/V100 and about 500ms on A100. And it is consitent across different platforms.

1.7 7. Do you agree or disagree with the authors' proposal? Please explain your reasoning.

It is promising, it has a future. But nowadays we are limited by hardware power. I agree with the idea. It has lots of benefits such as bandwidth saving, fixing broken images, even security.

1.8 8. Describe any ideas to enhance the proposed scheme.

- I did some research, I found GFPGAN (Generative Facial Prior, Generative Adversarial Network). It is used to restore faces on images. The paper said that images with faces received relatively low score.
- And implementing and improving AIs to generate images containing texts.
- Also we could optimize models for less powerful GPUs on the client-side.
- And for the cases where the webpages loaded slowly due to images. We could prioritize which to generate first.
- And we could start telling web developers to add "alt" text or leave more context, so AIs would know what to generate.

2 Web Caching

2.1 1. What is the problem that the paper mainly discusses?

The paper discusses problem with the current web caching approach, particularly in modern high-speed networks. Specifically, the main problem with the current web caching is re-validation process introduces latency due to unnecessary RTTs, which reduces the full potential benefits of caching. The problem is really relevant in today's internet landscape where latency, rather bandwidth, has become the primary bottleneck for web performance.

2.2 2. Why is the problem important?

The problem is important because PLT is directly affected by caching, and PLT has a direct impact on business revenue. Paper brings reports where even a 100ms decrease in PLT can lead to significant increase in business revenue. Thus inefficient web caching negatively impact on economic state. And also unnecessary data transmission resulting from poor caching can lead to higher power consumption and data costs for mobile users.
The current caching approach's limitations, particularly the latency involved in re-validation, make it ineffective for mobile web browsing where latency is a determining factor for PLT.

2.3 3. Why is the problem difficult to solve?

- HTTP cache re-validation mechanism requires a RTT, which is significant in today's latency-costrained internet, even though bandwidth is high. RTT occurs even if the resource has not changed, just to confirm its validity.
- Deciding whether a resource should be cached and for how long is difficult, because it's unpredictable how long the resource will remain unchanged.
- Use of "no-cache" header means a resource is always considered stale and requires updating/re-validation before each use.. more RTT.
- The current cache mechanism was designed for when bandwidth was the bottleneck. Its benefits focused on eliminating download time. But nowadays, with high download speeds and relatively small resource sizes, just eliminating download time is insufficient. And re-validation RTT must also be eliminated.

2.4 4. Summarize the existing methodologies for the problem. Not just copy the sentences in the paper. Try to describe what you understand in your own words.

- There is a way of using a proxy to automatically set cache header by estimating how often objects change, even though it is difficult. Another

way, MICRO-CACHING is about caching the unchanging parts, which reduced download size, but it still has RTT for re-validation.

- RDR: This approach uses a proxy server, running on a cloud close to the origin web servers, to fetch all necessary resources for a page on behalf of the client. By moving the resource fetching closer too the server, it reduces the impact of the client's last-mile latency. The proxy then sends the complete page to the client in one bulk transfer. However, this method raises security and privacy concerns as the proxy acts as a man-in-the-middle for encrypted connections and requires access to sensitive information like cookies. It primarily benefits the initial page load and doesn't help with resources fetched dynamically after user interaction.
- HTTP/2 feature, SERVER PUSH, allows to send resource to the client before the client specifically requests them. Like sending resources that the client will likely need in the future and thus avoids RTT. But it is not easy to predict what the user will need. Pushing too much can waste bandwidth and increase client processing load, which leads slow PLT. And it has another disadvantage, it is less effective when a page includes resources from multiple domains, as the main server cannot securely push content from other sites.

2.5 5. Describe the main ideas of the proposed scheme in this paper in your own words.

Under current conditions with global 5G (60Mbps throughput, 40ms latency), the method shows a notable reduction in PLT. Evaluation suggests an average PLT reduction of 30
When the client first request the main HTML file, the server sends the file with a list of validation tokens for associated resources. The client's browser after receiving the HTML and the tokens, it can compare the received tokens with the validation tokens of resource in cache. If tokens match, the browser can use the cached resource without sending requests, matching tokens will validate the cached resource. The browser will need to send less requests to the server for resources that are not in the cache or whose tokens did not match.
This approach simplifies caching configuration because the server no longer needs to specify how long a resource should be seen as valid.

The paper also suggests implementing Service Workers in existing browsers. The server sees a Service Worker that intercepts client requests and responses. The Service Worker caches resources and receives the latest validation tokens from the server with the initial HTML response. When the browser requests a resource, the SW intercepts it, checks its cache, and uses the server-provided token list to see if its cached version is still valid. If it is, it server the resource; otherwise it forwards the request to the server.

2.6 6. What are the notable findings in the evaluation results?

Even though the method designed to reduce PLT, it cannot significantly reduce it at low throughput.
But at high throughput, it is seen a significant reduction in PLT. It is because in high throughput, latency is the bottleneck. The method is designed to eliminate latency caused by unnecessary RTTs.
And also, at a constant throughput, the improvement in PLT is more pronounced for higher latencies. It means that the method provides greater benefits for users who are geographically farther from web servers or on networks with higher inherent latency.

Under current conditions with global 5G (60Mbps throughput, 40ms latency), the method shows a notable reduction in PLT. Evaluation suggests an average PLT reduction of 30%.

2.7 7. Do you agree or disagree with the authors' proposal? Please explain your reasoning.

If we believe the results and evaluation, the researchers are definitely moving to a promising direction. Improvements are needed, since the existing methods are outdated.
The method targets on eliminating re-validation RTTs and also eliminating bottlenecks wherever they may be.
Though the method does not work on low throughput, it works on current high throughput. So it shows that the method is up-to-date.

2.8 8. Describe any ideas to enhance the proposed scheme.

It is not easy to identify resources whose links are generated or discovered through JS execution, as this process isn't always deterministic. To enhance they need to develop a method to accurately determine this set of resources without imposing excessive load on the server. Authors suggested a method, where servers record the resources fetched by a specific user during their first visit to a page. The list can later help telling which tokens to provide on subsequent visits.

3 Web-LEGO

3.1 1. What is the problem that the paper mainly discusses?

The paper discusses slow webpage loading problem, which leaves bad impressions (low QoE), and improving QoE(Quality of Experience). The solution to the problem discussed in the paper is to speed up webpages that do not have Content Distribution Networks (CDNs).

3.2 2. Why is the problem important?

The problem of slow loading webpages affects the content providers and end-users. Slow loading time leaves bad impressions, and end-users are not likely to come back to the webpage that loads slowly. The paper bring study from Amazon, that have shown that even 100ms increase in load time can cost 1% in sales. Which proves the relation between page load time and sales.

3.3 3. Why is the problem difficult to solve?

There is an existing solution to this problem, CDNs are effective at speeding up web content delivery, but the cost can be expensive for providers who just started and they typically cannot afford it. Even though there are free CDN services, 77% of the top one million websites do not use CDN. There are solutions such as CDN adoption or updating to HTTP/2, but there are unlikely to be implemented by these CDN-less websites. And this creates a vicious circle, where providers cannot afford CDN -¿ Low QoE -¿ low profit -¿ no money to afford CDN...

3.4 4. Summarize the existing methodologies for the problem. Not just copy the sentences in the paper. Try to describe what you understand in your own words.

CDN: This is used and a valuable asset. They involve distributing copies of popular web content accross the globe to many servers. And users are directed to the closest server, which results in low latency. However, again, they are not universally adopted due to cost.
Network Protocols and Web Technologies: Protocols such as QUIC, SPDY, HTTP/2 aim to improve the efficiency of data transfer over the internet.
Webpage Optimization Techniques: Optimization is usually good for performance. Systems prioritize the content or understanding the page's structure to load essential elements faster.

3.5 5. Describe the main ideas of the proposed scheme in this paper in your own words.

The main idea is a system called Web-LEGO, that is designed to speed up webpages primarily for users visiting websites that are not using CDN or other server-side speedups.
Web-LEGO replaces the original content with faster alternative content, which is either identical or similar. In the paper it says "trading strictness for performance".
It works through three main components: a client, a SIMILARITY STORAGE SERVER(3S), and REVERSE FILE SEARCH SERVER (RFSS).

- Client intercepts requests and sends the original request to the target server and sedns the requests for supported content types to the 3S.
- 3S stores information about alternative resources for web objects. And when the client asks it, it looks for alternative URLs in its DB. If it finds, it provides a list of alternative resources. But if it cannot find, it forwards the request to the RFSS.
- RFSS is responsible for finding alternative identical or similar content on the internet. It returns these alternative URLs to the 3S to populate its DB. It check the preference of providers regarding sharing or replacing content based on similarity control header.

The client sends multiple requests: the original request and requests to the alternative resources provided by 3S. The first response is returned to the browser. For JS and CSS files, Web-LEGO searches IDENTICAl copies, but for images SIMILAR ones are okay, too.
And this method is an OPT-IN service, which means users and content providers should be used or not.

3.6 6. What are the notable findings in the evaluation results?

- 36% of Alexa's top 1000 websites and 77% of the top one million websites have not implemented CDNs, which the perfect target for Web-LEGO.(if I understood the paper correctly)
- Lots of similar content. Google Images returned at least 10 smilar images for 92% of tested image URLs, and at least two similar images for 97% of cases.
- 92.5% of evaluation gave positive feedback on Web-LEGO generated webpages. Which means many users are ready to trade content strictness for speed.
- Web-LEGO speeds up load time reducing PLT(Page Load Time) by up to 5.5s and uPLT(User-Perceived Page Load Time) by up to 5s. The median speedup was 7 times greater than even CDN-hosted sites. So Web-LEGO reduced PLT for 80% of tested CDN-less websites.
- The overall accuracy of finding alternative images was reported as 94.2%. 73.9% of original images had at least one identical alternative, and 11.9%

had very similar alternatives. Which means that images have the most potential for replacement.
- The normalized benefit analysis showed that the performance benefits outweigh the cost of extra traffic in broadband wireline and 4G mobile scenarios. It is economically viable for clients.
- Web-LEGO is beneficial with 4G networks, but causes congestion and degrade performance increasing PLT with 3.5G networks.

3.7 7. Do you agree or disagree with the authors' proposal? Please explain your reasoning.

I agree with authors, their research shows that the websites contain very similar content. So there is no need to request from the farther servers, when we can request closer server, reducing PLT. And the fact that it is an OPT-IN function, so the users and providers can choose to use or not to use if the content requires to be strict.

3.8 8. Describe any ideas to enhance the proposed scheme.

I would think more about the security. Maybe replacing the original content with similar ones can be open to different kind of vulnerabilities.