Learning to Generate HTML From Images With Actor-Critic

Josh Davis, Zlatko-Salko Lagumdzija, Shelvin Pauly, Davit Soselia, Michel Temgoua, Benjamin Wu
University of Maryland, College Park

Abstract

In recent years, Language Models (LMs) have demonstrated remarkable ability in generating coherent text and text samples. Models such as Codex have shown promise in aiding software engineers in their daily work, while attention-based models have achieved success in various other downstream tasks, such as image classification, segmentation, and annotation. Among them, some landmark works are Vision Transformer (ViT), SWIN, and other architectures, with some models specifically targeting document-related tasks, such as the Document Image Transformer (DiT).

This project explores the application of actor-critic fine-tuning to train a model that generates front-end code capable of reproducing an input image. The research builds upon existing literature on RL-based code generation from text specifications established in CodeRL and extends the idea to image or multimodal contexts. The model's performance is evaluated using the modified CodeBLEU metric to attempt to better approximate the visual similarity of the output that would result in from running the generated code, as well as visual similarity metrics.

Methodology

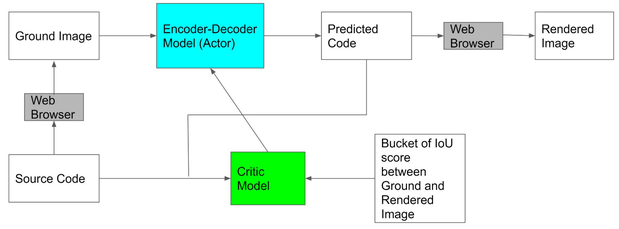
Datasets:

1. **Synthetically generated examples of HTML code to generate combinations of various simple shapes.**
2. Examples of simple website login forms scraped from open-source projects on GitHub.
3. Simple HTML and CSS examples scraped for the website W3Schools.

HTMLBLEU: Derived from CodeBLEU, HTMLBLEU is a similarity measure between generated and ground truth code. It is based on a combination of matching the HTML DOM trees of the two codes, the CSS position, size, and color, BLEU, and a weighted BLEU on HTML and CSS keywords.

Methodology

The Actor-Critic model fine-tunes the baseline. The goal is to encode Visual similarity information.
The Critic is trained to assess IoU scores of source and predicted code renderings and its Intermediate returns serves as weighting for the update function. Specifically,



$$\nabla_\theta \mathcal{L}_{rl}(\theta) \approx -\left(r\left(W^s\right) - r\left(W^b\right)\right) \sum_t \hat{q}_\phi\left(w_t^s\right) \nabla_\theta \log p_\theta\left(w_t^s \mid w_{1:t-1}^s, D\right)$$

Loss is used to improve the actor (Encoder-Decoder) model after the baseline pretraining.

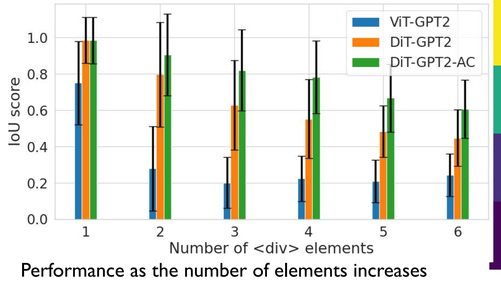
Results

	ViT-GPT2	DiT-GPT2	CNN-BART-E2	CNN-BART-E40	**DIT-GPT2 (AC)**
mean IoU ↑	0.31	0.64	0.05	0.03	**0.78**
mean MSE ↓	19.63	12.25	14.5	14.5	**9.0**
mean MSE (Colorblind) ↓	0.11	0.07	0.13	0.10	**0.04**

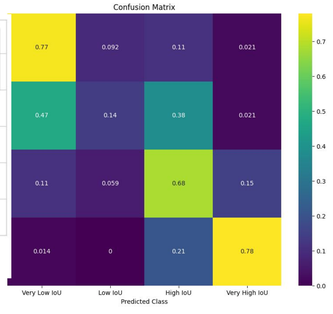
Table 5: Comparison of rendered images from the predicted code with the original. The best-performing model by each metric is highlighted in bold.

In-depth ViT/DiT Comparison

	ViT-GPT2	DiT+GPT2
BLEU ↑	0.65 ± 0.08	0.74 ± 0.09
htmlBLEU ↑	0.62 ± 0.13	0.69 ± 0.14
IoU ↑	0.31 ± 0.25	0.64 ± 0.27
MSE (Single Channel) ↓	19.63 ± 11.59	12.25 ± 8.83
MSE (Colorblind) ↓	0.11 ± 0.06	0.07 ± 0.06
Element Counts ↑	0.97 ± 0.16	0.97 ± 0.18

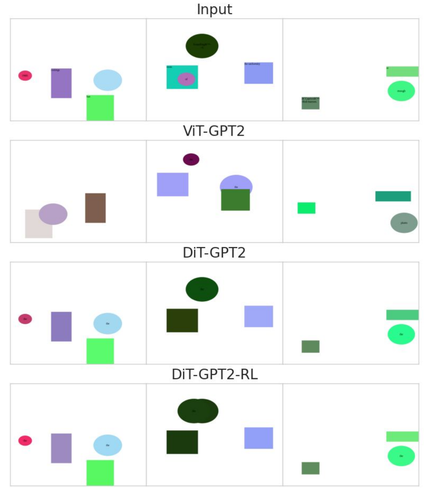


Performance as the number of elements increases



Confusion matrix for Critic Model

Results



Sample outputs from the best models

Conclusion and Future Work

We demonstrate a methodology for generating HTML code from images using a vision-to-text encoder-decoder model fine-tuned with Actor-Critic reinforcement learning. We find that our fine-tuned version of DiT-GPT2 outperforms the baselines on the IoU and MSE scores. Performance worses as more div elements are added.
Future work includes incorporating the web-scraped datasets we previously proposed and adding a human evaluation performance metric from surveys of the team. We can also explore additional training methodologies for the critic and expand the functionality of the HTMLBLEU.