Vision-Code Transformer for Screenshot-to-HTML/CSS Generation

Davit Soselia, Khalid Saifullah, Tianyi Zhou
University of Maryland

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Motivation:

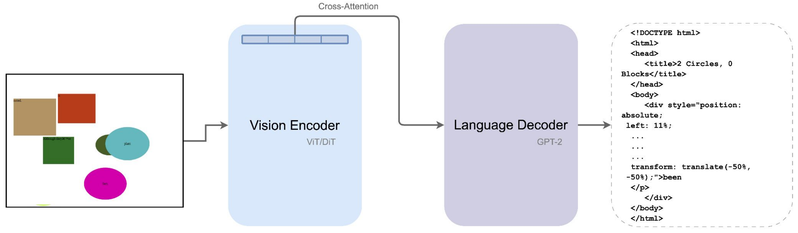
The automation of HTML/CSS code generation from screenshots is a significant challenge with broad applications in website development and design. This task has the potential to streamline the website creation process, reducing the time and effort required by developers.

Problem Description:

Despite its importance, automating code generation from screenshots is a complex task due to the need to accurately interpret visual elements and translate them into corresponding code snippets. Existing methods have limitations in handling this task efficiently and accurately.

Key Contributions:

- We present a strong baseline for the problem, leveraging an Encoder-Decoder architecture, specifically using Vision Transformer (ViT) and Document Image Transformer (DiT) as image encoders.
- To train and evaluate our models, we created a synthetic dataset of 30,000 unique pairs of code and corresponding screenshots.
- We introduce a novel htmlBLEU score for evaluating the quality of the generated code.



Dataset:

To train and evaluate our models, we created a synthetic dataset of 30,000 unique pairs of code and corresponding screenshots. This dataset provides a diverse range of examples for our models to learn from.

Approach:

We use an Encoder-Decoder architecture. We use Vision Transformer (ViT) and Document Image Transformer (DiT) as image encoders. These encoders extract features from the input screenshots, which are then passed to a decoder.. The decoder generates the corresponding HTML/CSS code in a token-by-token manner.

Table 1: Evaluation of code generation from screenshot

Model	ViT-GPT2	DiT-GPT2
Metrics		
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 ↓	19.63 ± 11.59	**12.25** ± 8.83
MSE (Single Channel) ↓	0.15 ± 0.09	**0.07** ± 0.06
Element Counts ↑	**0.97** ± 0.16	**0.97** ± 0.18
Human Evaluation (Normalized)		
Color Fidelity ↑	0.26 ± 0.26	0.61 ± 0.32
Structural Similarity ↑	0.44 ± 0.35	0.62 ± 0.35