Image to HTML

Generating HTML from Images with Language Models

Graduate

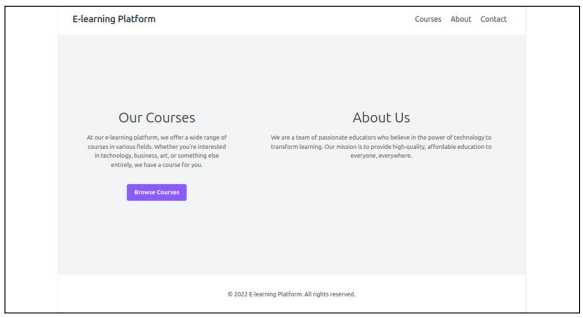
Abinas Kuganathan

Introduction: Designing websites demands both creative design and technical development skills. Modern design tools such as Figma, Adobe XD, and Penpot allow designers to rapidly create and refine user interface concepts without extensive coding expertise. However, transforming these static designs into functional prototypes remains a time consuming process. While Large Language Models (LLMs) have demonstrated strong capabilities in generating text and code, describing complex visual layouts purely through text is often inefficient and unintuitive. Visual inputs, such as design mockups, offer a more direct and expressive means of communication. Vision-Language Models (VLMs), which can process and interpret both textual and visual information, are therefore well-suited for this task.

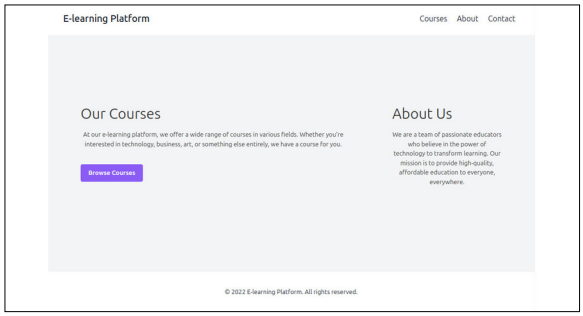
Problem: Current state-of-the-art VLMs are often inaccessible to smaller projects or individual designers who prefer not to send their designs to cloud-based services. Many existing solutions are closed source or require substantial computational resources, limiting their practicality for a wide range of applications. This thesis addresses these limitations by proposing a small image to HTML model, enabling faster and more seamless prototyping of web applications. The proposed architecture integrates a Vision Transformer (ViT) encoder with the decoder of a sequence-to-sequence model, trained to translate visual designs into functional HTML code.

Result: This thesis presents a model that combines google/siglip2-base-patch16-512 as the ViT with the decoder from Salesforce/codet5-base. The input image is split into four quadrants, processed independently, and subsequently merged within the model. The architecture was evaluated on the Design2Code Benchmark using both the provided dataset and a holdout set. The model achieved the following scores on the holdout set: Block: 98.08%, Text: 99.34%, Position: 96.84%, Color: 98.43%, and CLIP: 97.35%, confirming the effectiveness of the architecture. The model has 235 million parameters, which is an order of magnitude smaller than other vision-language models for this task.

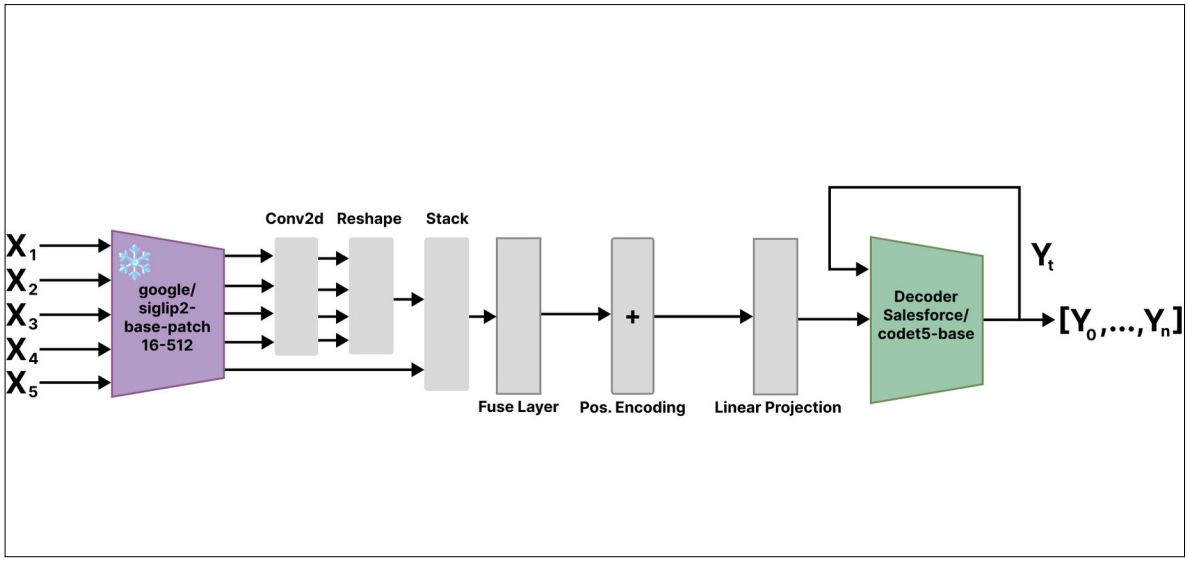
Input Image [Holdout set]
HuggingFaceM4/WebSight



Generated and rendered HTML
Own presentment



Proposed Architecture
Own presentment



Advisor
Prof. Dr. Markus Stolze

Co-Examiner
Dr. Cristiano Malossi, IBM Research Europe, Rüschlikon, Zürich

Subject Area
Data Science, Computer Science