# Research Seminar Data Science Task 2 — Paper 5: What the DAAM

Daan Brugmans (S1080742)

## I. Summary

The authors present DAAM, an explainability method for text-to-image diffusion models. DAAM is capable of producing 2D heatmaps for every word in a text-to-image prompt by analyzing the cross-attention maps of the diffusion model, visually showing how a word in the prompt has effected which parts of the generated image. The authors use DAAM on the Stable Diffusion 2.0 model to study the effect of text prompt syntax on the generated image and how certain linguistic features negatively impact image generation.

## II. Evidence

- The authors perform experiments on DAAM's capability to produce good segmentation masks. Qualitative comparisons to existing state-of-the-art baselines and subjective judgments by human annotators show DAAM's to generate good explanations for varying parts-of-speech and image semantics.
- The authors perform an experiment on the effect of syntax on text-to-image generation by studying how varying head-dependency structures are generally visualized by DAAM, showing that certain head-dependency structures show one word dominating the other in importance for the generated image, implying that certain syntactic structures influence the text-to-image diffusion process.
- The authors perform experiments on the relationship between semantics of text-to-image prompts and feature entanglement in the generated image, showing that the presence of cohyponyms in text prompts cause worse image generation, and that adjectives in prompts affect the generated image in more than just the noun it belongs to.

## III. Strengths

- The authors provide the reader preliminaries that teach the reader the required knowledge about latent diffusion model architecture needed to understand their paper, which helps bring the reader up to speed.
- When findings and explanations are provided textually, they are often accompanied by a visualization or formalization to help understanding.
- Experiments measuring both quantitative improvements over baselines and perception by humans in practice
- The conducted experiments are both quantitative through measures and qualitative through human judgments, which makes the results both more concrete as well as more intuitive for end users.
- The qualitative measures used fit the experiments very well and seem very relevant.

## IV. Weaknesses

- The results of the experiments for the authors' visuosemantic analyses are purely based on manual human annotation, reducing certainty of interpretability.

## V. Evaluation

I would recommend acceptance. In clear and concise writing, the authors propose a method for the interpretability of a topically relevant model architecture that improves over existing baselines.

## VI. Quality of Writing

The main body of the paper is easy to read. Mathematical definitions and proofs are kept to a minimum in the main body, only frequently used for the preliminaries required to understand the rest of the paper. By placing mathematical definitions, technical details, and auxiliary supplements in appendices, the authors have made their paper, despite what may be considered a challenging topic, an enjoyable read.

## VII. Queries

- Can the presenters offer a measure for the visuosemantic analyses that would not be dependent on human annotation? If yes, which? If no, for what reason?