UM EECS 542: Advanced Topics in Computer Vision
UM CSE 598: Action and Perception
Homework #2: Multimodal Automated Interpretability Agent
Due: 24 September 2025 11:59pm

1 Overview

This assignment introduces students to the Multimodal Automated Interpretability Agent (MAIA) [1], a framework that approaches AI interpretability as a process of hypothesis-driven investigation. Students will set up and run MAIA using the provided codebase and API access, then analyze the results produced by the framework. Through these tasks, students will gain hands-on experience with automated interpretability methods and develop the ability to critically evaluate the reasoning processes of AI systems. The assignment extends across multiple vision architectures, including ResNet[2], DINO[3], and CLIP[4], providing opportunities to compare interpretability across model families and to reflect on the broader challenges of explaining modern AI systems.

2 Background

MAIA [1] is a framework (fig. 1) that treats the interpretation of AI models as a scientific process of forming and testing hypotheses. At its core, MAIA uses a vision-language model to automatically write and run small Python programs. These programs are designed to investigate and understand the behavior of other AI systems, from a single artificial neuron to the model's final output.

To do this, MAIA employs a versatile toolkit that includes finding the best real-world images that activate a specific part of a model, creating new[5] or edited[6] images to perform targeted "what if" experiments, measuring the model's responses in a quantitative way, and summarizing its findings in natural, human-readable language. MAIA operates in a continuous loop: it proposes a hypothesis about the model's behavior, runs experiments to gather evidence, evaluates the results, and then refines its hypothesis. This method ensures that the explanations it produces are based on systematic evidence rather than simple, one-off descriptions.

The key advantage of MAIA is that it transforms AI interpretability from a passive act of observation into an ac-

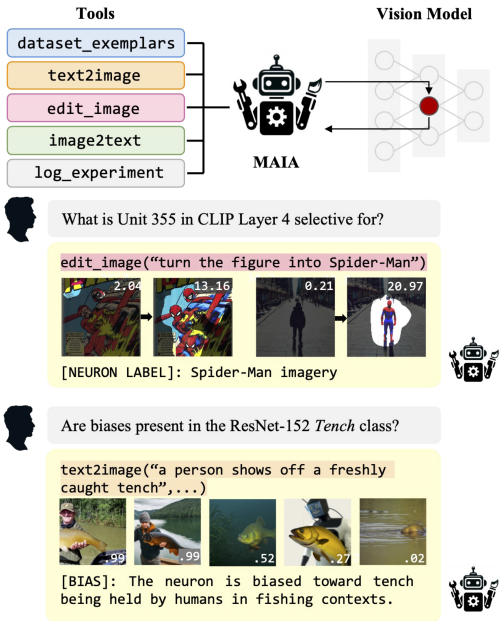

Figure 1: **MAIA framework.** MAIA autonomously conducts experiments on other systems to explain their behavior.

tive investigation. By combining real-world data with carefully controlled synthetic images, MAIA can distinguish between very similar concepts that might confuse a model. This allows it to identify and help correct issues where a model relies on misleading shortcuts or irrelevant features. It can also uncover biases or blind spots in how a model classifies information. This approach is highly scalable, working for both very specific questions—like what a single neuron is designed to detect—and for broad audits of a model's overall performance. Ultimately, the explanations generated by MAIA are designed to be predictive. They help researchers anticipate how a model will respond to new, unseen data and provide clear, actionable insights for debugging, improving safety and fairness, and making targeted edits to the model's architecture.

The framework is also modular, meaning it can be adapted to study a wide range of vision models (such as ResNet[2], DINO[3], CLIP[4]) and can incorporate new tools as they become available. Together, these features make MAIA a rigorous, flexible, and reliable foundation for studying and understanding modern AI systems.

3 Problem Set

The goal of this assignment is to familiarize you with the automated interpretability pipeline and to provide insight into the capabilities of state-of-the-art models, offering a opportunity to critically evaluate their "intelligence" against your own analytical skills.

This assignment does not involve model training. Instead, your primary objective is to successfully execute MAIA using the provided codebase and API access. Once the framework is running, you will be responsible for thoroughly reviewing the experimental results, examining the correctness of the agent's logic, identifying any misleading sections, and writing a comprehensive summary of the entire experimental process. Below are the specific tasks you should complete; implementation details and relevant links are in section 4.

1. **(20 points) MAIA Implementation:** Begin with the resources in section 4 and ensure you can run MAIA end to end in your environment. In addition to following the repository's instructions, configure your own personal Hugging Face access token and a Gemini API key, unless you choose equivalent alternatives.

   In your submission, describe your setup process, note any difficulties you encountered and how you resolved them. You may use generative AI for assistance only at this stage; if you do, describe how you used it.

2. **(50 points) Result Validation on ResNet:** For this task, you will use your configured MAIA environment to interpret **Unit 122 of Layer 4 in the ResNet-152 model**, using the weights provided in the MAIA repository. To validate your implementation, the expected result is the identification of the neuron as a "bow tie detector" or an equivalent concept, though an alternative interpretation will be accepted if your agent's examination process is thoroughly documented and justified.

   In all cases, provide a comprehensive summary and analysis of the entire reasoning process, and critically evaluate the correctness of the agent's findings. Your report must also include exemplar images from the first iteration for the specified neuron (see fig. 2) and a screenshot of MAIA 's final description section (see fig. 3). In addition to the report, submit the generated experiment log named `exp_resnet152_layer4_unit122.html`.

Figure 2: Exemplar images for Unit 122 of Layer 4 in Resnet-152.

MAIA

```
[DESCRIPTION]: The neuron is selectively activated by the distinct, structured forms of formal neckwear.
This includes the symmetric loops and central knot of bow ties, the intricate, tightly folded knots
of neckties (such as Windsor or Four-in-Hand), and the prominent, folded fabric structure of cravats/ascot ties.
Activation is strong for these items, whether worn on a person (typically with a collared shirt)
or clearly visible in isolation. Unstructured neckwear or plain collars produce significantly lower activation.

[LABEL 1]: Structured knots of formal neckties
[LABEL 2]: Bow shapes of neckwear
```

Figure 3: Description from MAIA for Unit 122 of Layer 4 in Resnet-152.

3. **(25 points) Extension to DINO:** Run interpretation on any neuron of your choice within the DINO-ViTS8 model, using the weights provided in the repository. **Please avoid choosing Unit 80 of Layer 9** as it is provided as an example. For the report and the experiment log, follow the format in the Resnet's example.

4. **(25 points) Extention to CLIP:** Apply the same process in the DINO task for the CLIP-RN50 model. **Please avoid choosing Unit 487 of Layer 3**.

4 Setup

1. Clone and follow the installation steps in MAIA's repository: https://github.com/multimodal-interpretability/maia.

2. Set a personal token for Hugging Face: https://huggingface.co/.

3. Set a personal token for the Gemini API. The `gemini-2.5-flash` model should be sufficient to serve the MAIA module: https://aistudio.google.com/.

4. (Optional) Replace the `utils/call_agent.py` file with the version provided in this assignment for Gemini adaptation. Feel free to code your own version.

5 Submission Instruction

1. This assignment is to be completed individually.

2. Submissions should be made through Gradescope and Canvas. Please make sure you submit to the right course session you are enrolled in. **Please upload the PDF to Gradescope and the ZIP file to Canvas**:

   a) A **PDF file** for the **report**: At the top of the first page, include your name, student ID, and the date of submission. Each problem should begin on a separate page, resulting in a total of four pages.

   If you believe there may be an error in your code, provide a written statement in the PDF explaining what might be wrong and how it affected your results. For any functionality/module you were unable to implement, you may include pseudocode and/or expected results. Any additional explanations are optional and, if included, should be presented as a fifth page at the end of the report.

   b) A **ZIP file** for the **experiment logs**: Submit a ZIP file containing a folder with all your experiment logs in HTML format. If you have completed all the problems, the folder should include three files in total.

References

[1] Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent, 2025.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[5] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[6] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. Instructdiffusion: A generalist modeling interface for vision tasks, 2023.