Application Materials for

Haekyu Park

Georgia Institute of Technology

haekyu@gatech.edu

Primary Research Area: **Human-Computer Interaction**

Secondary Research Area: AI for Social Good

To whom it may concern,

I confirm that Haekyu Park, a PhD student in my department, passes the eligibility requirements.

Haesun Park
Chair, School of Computational Science and Engineering
Regents' Professor

Student CV

Short (1-page) CV of the student’s advisor

Letters of Recommendation

Research / Dissertation Proposal

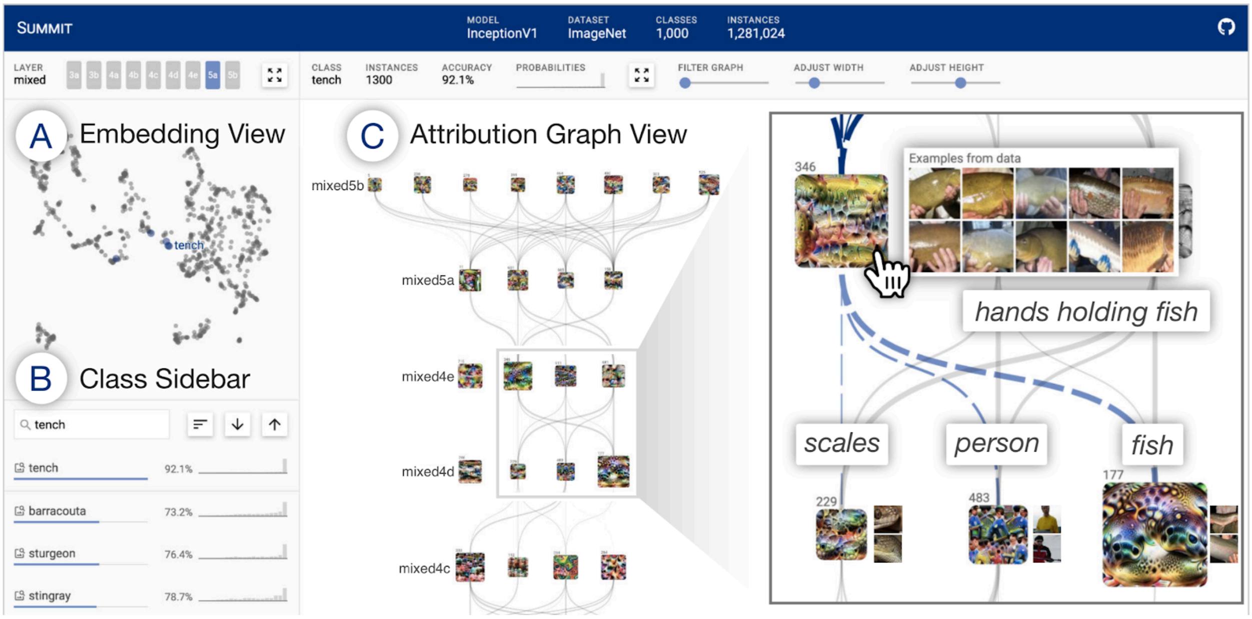
Human-centered AI: Interactive Scalable Interfaces for Equitable and Trustworthy AI

Haekyu Park, Georgia Tech
haekyu@gatech.edu, haekyu.com

Artificial intelligence (AI) technologies, especially Deep Neural Networks (DNNs), are now solving some of the world's hardest problems. Yet, our society still faces fundamental barriers to learning, understanding, and ultimately trusting AI technologies: (1) AI models are often used as "black boxes" and their lack of transparency makes people hesitant to trust and deploy them, (2) when models do not perform satisfactorily, people lack actionable guidance for understanding their errors and how to fix them, (3) despite research advances in state-of-the-art AI, they are increasingly seen as a *walled garden* accessible only to highly specialized experts; this skills gap quickly demotivates aspiring beginners from learning these powerful new technologies.

My research addresses these significant challenges in AI through a *human-centered* approach, by creating novel interactive visualization tools that enable people to instill trust in well-performing AI models and to fix malfunctioning models. Also, to prevent inequality in acquiring AI proficiency, I have been building education modules that help broad audiences to easily learn and leverage these modern technologies. Through my research in information visualization, machine learning, and data analytics over the past five years [1-11], I have realized that the key to promote trust in AI and broaden education access for AI is to bring humans into the loop. Specifically, my research focuses on these three complementary thrusts.

1. **Scalable visual discovery for trustworthy and interpretable AI:** developing the *first scalable graph representations for understanding DNNs for millions of images,* through novel integration of feature visualization, graph visualization, and graph discovery algorithms (e.g., *Summit* [1]).
2. **Actionable insights to protect and troubleshoot models:** developing scalable interpretation techniques to pinpoint malfunctioning components (i.e., neurons and their connections) in deep learning models and to understand how those defects induce incorrect predictions (e.g., *Bluff* [2], *Massif* [3], *NeuralDivergence* [4], *SkeletonVis* [11]).
3. **Equitable AI access and education opportunities:** designing interactive visual interfaces that broaden AI education access for students and non-experts, inspiring them to learn and apply AI in their domain (e.g., NVIDIA RAPIDS education module; *CNN Explainer* [5], *Argo Lite* [7] used by thousands of learners worldwide).



MODEL InceptionV1
DATASET ImageNet
CLASSES 1,000
INSTANCES 1,281,024

LAYER mixed
3a 3b 4a 4b 4c 4d 4e 5a 5b

CLASS tench
INSTANCES 1300
ACCURACY 92.1%
PROBABILITIES

FILTER GRAPH
ADJUST WIDTH
ADJUST HEIGHT

tench

tench	92.1%
barracouta	73.2%
sturgeon	76.4%
stingray	78.7%

Fig 1. *Summit* [1] helps users scalably summarize and interactively interpret deep neural networks by visualizing *what* features a network detects and *how* they are related. *Summit* can discover surprising associations, as in this “tench” prediction (yellow-brown fish). It reveals that the classification for “tench” heavily depends on recognizing parts of people (e.g., fingers) and their close interactions, instead of typical fish features.

Thrust 1: Scalable Visual Discovery for Trustworthy and Interpretable AI

AI, especially DNNs, are often considered “unintelligible” due to their complex architectures and huge number of parameters. Unfortunately, this lack of interpretability poses significant concerns as DNNs are increasingly used in safety-critical applications and societal challenges. How do we help people interpret DNNs, so they can understand, trust, and confidently deploy such models?

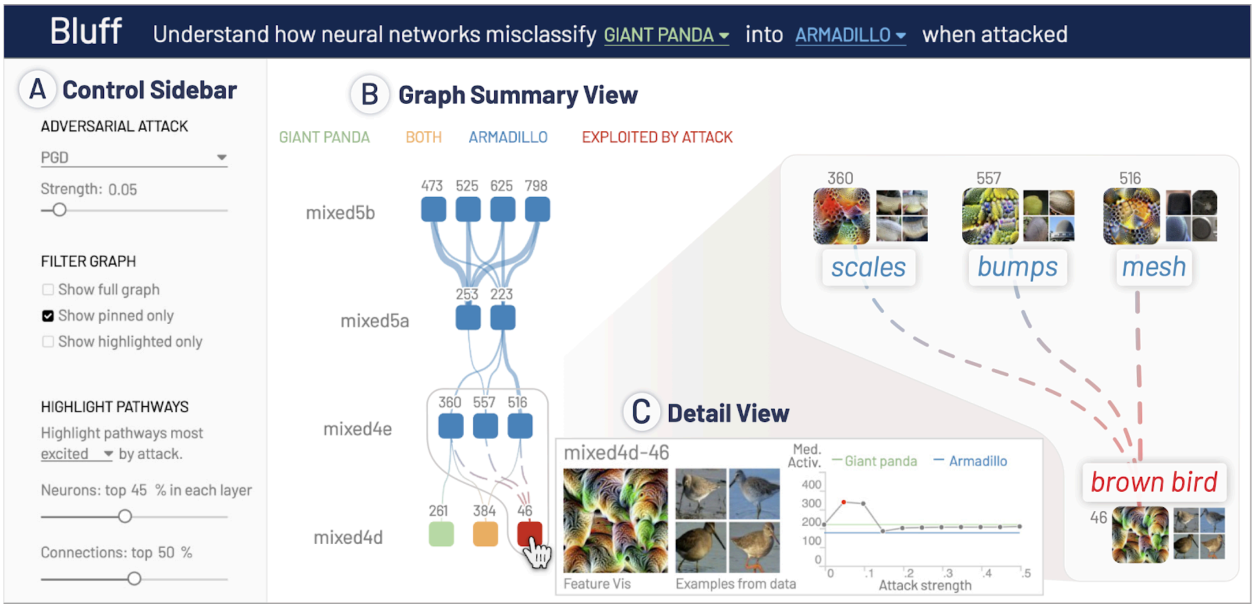
My answer is by examining the **connections among neurons inside** of the models. This idea stemmed from my biology studies as part my early interest in food science and nutrition. I read about the fascinating *OpenWorm* project where a robot's artificial neural network fully "cloned" the brain of the famous *C. elegans* worm by replicating the worm's 302 neurons and their connections [11]. Miraculously, the robot mimicked the worm's actions without further programming (e.g., sensing and bypassing obstacles). Such discovery suggested that neuron connections captured both the essence of knowledge and behaviors. Fascinated by this observation, I became convinced that the crux in enabling interpretable and trustworthy models lies in investigating DNNs' internals (i.e., neurons and their connections) and explaining their operations with effective human-centered designs.

1.1. Scalable Interpretation via Novel Graph Representations for DNNs

I have been developing novel visual representations that serve as *smart microscopes*, helping users inspect a model's internals and how they lead to the model's decisions. For example, in Fig. 1 our *Summit* [1] system visualizes and explains how DNNs recognize an object (e.g., tench, a yellow-brown fish) through the novel *attribution graph*, which summarizes what concepts (e.g., hands holding fish) are detected by neurons and how those features interact through the neural connections (e.g., a neuron detecting fish and another neuron detecting person fire a neuron for hands holding fish). *Summit* scales to large data, such as the ImageNet dataset with 1.2M images, and leverages neural network feature visualization and dataset examples to help users distill large, complex neural network models into combat, interactive visualizations. It was published at *IEEE TVCG'20*, the top data visualization journal.

1.2. Model Incubator: Nurturing Infant DNNs — A Stitch in Time Saves Nine

Based on our techniques to interpret how neural connections store the knowledge, I plan to build a **model incubator** that helps machine learning developers make sense of how neural connections evolve during training. The model incubator will monitor the training process and caution the experts about potential undesirable model developments. For example, in our surprising tench discovery, model incubator could alert the developers that the model is learning features (e.g., human faces) that are anomalous with respect to the desired (fish) characteristics. My tool could help machine learning developers build more robust model, more efficiently, and importantly allow them discover model deficiencies and correct them in the early stage of the training.



Bluff Understand how neural networks misclassify GIANT PANDA into ARMADILLO when attacked

ADVERSARIAL ATTACK: PGD

Strength: 0.05

FILTER GRAPH

□ Show full graph
☑ Show pinned only
□ Show highlighted only

HIGHLIGHT PATHWAYS

Highlight pathways most excited by attack.

Neurons: top 45 % in each layer

Connections: top 50 %

Fig 2. Bluff [2] visualizes how adversarial attacks penetrate a DNN to induce incorrect results. Here, a user inspects why a DNN misclassifies adversarial giant panda images, crafted by the *Projected Gradient Descent* (PGD) attack, as armadillo. PGD successfully perturbed pixels to induce the "brown bird" feature, an appearance more likely shared by an armadillo (brown, roundish body) than a panda, activating features that contribute to the armadillo (mis)classification (e.g., "scales," "bumps," "mesh"). Altogether, these neurons and their connections form adversarial pathways that overwhelm the benign panda pathways, leading to the ultimate incorrect prediction.

Thrust 2: Actionable Insights to Protect and Troubleshoot Models

2.1. Understand Why and How Models Fail

As AI technologies expand to almost every aspect of our lives, a model's malfunctions have literally become a matter of life and death, especially when AI is used in safety-critical applications such as self-driving cars and data-driven health care. However, due to the black box nature of AI, it is difficult to pinpoint where and why models malfunction. For example, when a DNN is harmed by an *adversarial attack*, it is hard to pinpoint the parts of the model exploited by the attack, let alone to understand how such exploitations lead to incorrect outcomes.

To identify and figure out where and why models would fail, I focus on comparing how the neurons and their connections respond differently to normal inputs and abnormal inputs. For example, to understand DNNs' vulnerability to adversarial attacks, our *Bluff* [2] system compares the *attribution graph* of benign inputs and adversarially perturbed inputs, to find where a DNN model is exploited by the attack and what impact the exploitation has on the final prediction across multiple attack strengths. As seen in Fig 2, *Bluff* interactively interprets why adversarial panda images, perturbed by Projected Gradient Descent (PGD) attack, are misclassified as armadillo. It reveals that the adversarial panda images suspiciously activate a neuron that detects brown birds whose appearance are similar to an armadillo than a panda (e.g., both brown bird and armadillo have roundish brown bodies). PGD exploits such a neuron as a stepping stone to abnormally stimulate neurons that are important for armadillo classification, such as “scales”, “bumps”, and “mesh”. *Bluff* was published at *IEEE VIS’20*, the top data visualization conference. Our *Massif* [3] system also interactively interprets adversarial attacks on DNNs, by fractionating the neurons and their connections based on their vulnerability and visualizing the fractionated groups with different positions and colors. It was published at *CHI’20*, the top human computer interaction conference. Our *NeuralDivergence* [4] system visually summarizes and compares neural activation distributions of pairs of adversarial attacked and benign images. It was published at PacificVis’19.

2.2. Neural Model Surgery: Troubleshooting Vulnerabilities

Pinpointing where a model does not work well and interpreting how the problem leads to undesirable outcomes are useful for revealing the root cause of the problem. However, detecting a problem does not mean solving it, as there could be a myriad of possible paths to address it, some more effective and efficient than others. In my ongoing research, I am developing interactive visual systems that help users easily fix and improve DNNs, while helping them make sense of why their actions would improve the models, and how effective they are.

Specifically, I am building on my experience with *Bluff* to investigate the novel idea of **neural model surgery**, where users would interactively edit neural connections to enhance DNNs’ robustness to adversarial attacks, such as inhibiting or removing vulnerable connections from the network; in real time, the users would observe how the *attribution graph* evolves and how the predictions improve. Such real time feedback could enable multiple new ways to protect a model: (1) users will incrementally develop new insights into constructing stronger defenses; and (2) a model could learn from such feedback to automatically fortify its neural connections.

Thrust 3: Equitable AI Access and Education Opportunities

3.1. Visual Tools to Equalize AI Education Opportunities

State-of-the-art AI is rapidly expanding across domains, much faster than the learning speed of those who wish to learn and apply AI. Accordingly, AI is increasingly seen as a *walled garden* accessible only to expert scientists and engineers; enormous skills gaps quickly demotivate aspiring students and practitioners from learning about these powerful modern technologies.

To promote equitable education for AI, I have been developing interactive visual interfaces to help people more easily get started when learning about AI. For example, our *CNN Explainer* [5] system enables deep learning beginners to visually examine how Convolutional Neural Networks (CNNs) transform input images into classification and interactively learn about their underlying mathematical operations. It was published at *IEEE TVCG'21*. It successfully attracted an **overwhelming amount of attention**, receiving almost 5000 GitHub stars and thousands of “likes” on social media.

3.2. Broadening AI Access for Diverse Domains and Users

In Summer 2019 and 2020, as a research intern at NVIDIA’s RAPIDS team, I leveraged my visualization and system skillset to build a visual analytics tool that enabled people to run graph algorithms interactively, and to apply the techniques to various domains such as cybersecurity. Working closely with my mentors Brad Rees, Joe Eaton, and Bartley Richardson, I developed an interactive visual tool that helps cyber analysts detect, visualize, and examine suspicious events in time-evolving cyber networks by using graph algorithms such as personalized PageRank. My work [6] was presented at NVIDIA’s Tutorial in *ACM KDD* conference, the top data mining conference.

To further close the AI skills gap, I am actively engaged in disseminating AI knowledge through **free, open-source courseware**. As a teaching assistant for the large 250-student Data Visual Analytics course at Georgia Tech (CSE 6242), I lead-developed a brand new education module that introduced students to state-of-the-art GPU-accelerated AI techniques, by leveraging my in-depth GPU and RAPIDS knowledge. I was excited to be **invited as a speaker of NVIDIA’s GTC’20** (GPU Technology Conference) to share my valuable education efforts with educators and practitioners.

Human-Centered AI: Research Contributions and Impact

In summary, my research advances the frontier of AI through human-centered approaches, contributing novel methods and tools that promote human understanding of machine learning, accelerate their development, broaden their education access, and increase people's trust in their results. My works have made novel research contributions and significant impacts to society.

- All my recent systems for promoting AI interpretation have been open-sourced, including *Summit* [1], *Bluff* [2], *Massif* [3], *CNN Explainer* [5], *Argo Lite* [7]. *CNN Explainer*, in particular, attracted an overwhelming amount of interest from learners worldwide (e.g., 5000 GitHub stars and thousands of "likes" on social media).
- Our *CNN Explainer* [5] system for learning CNNs has been adopted by a rapidly increasing number of courses and institutions around the world, including Georgia Tech (Deep Learning, Intro to Perception and Robotics), University of Wisconsin-Madison (Intro to AI), and University of Kyoto (Bioengineering).
- My RAPIDS education module on GPU-accelerated data science was the first of its kind, is publicly available, and is used by thousands of students.
- Our *Argo Lite* [7] graph visualization system is now used by over 2000 students every year.

References

[1] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Chau. “Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations.” IEEE VIS, 2019.

[2] Nilaksh Das*, Haekyu Park*, Zijie J. Wang, Fred Hohman, Robert Firstman, Emily Rogers, and Duen Horng Chau. “Bluff: Interactively Deciphering Adversarial Attacks on Deep Neural Networks.” IEEE VIS, 2020. (*Equal contribution)

[3] Nilaksh Das*, Haekyu Park*, Zijie J. Wang, Fred Hohman, Robert Firstman, Emily Rogers, and Duen Horng Chau. “Massif: Interactive Interpretation of Adversarial Attacks on Deep Learning.” CHI, 2020. (*Equal contribution)

[4] Haekyu Park, Fred Hohman, and Duen Horng Chau. “NeuralDivergence: Exploring and Understanding Neural Networks by Comparing Activation Distributions.” IEEE PacificVis, 2019.

[5] Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Chau. “CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization.” IEEE VIS, 2020.

[6] Rachel Allen, Haekyu Park, Bartley Richardson. “RAPIDS and Cybersecurity: A Network Use Case”, ACM SIGKDD NVIDIA Tutorial, 2019.

[7] Siwei Li, Zhiyan Zhou, Anish Upadhayay, Omar Shaikh, Scott Freitas, Haekyu Park, Zijie Jay Wang, Susanta Routray, Matthew Hull, and Duen Horng Chau, “Argo Lite: Open-Source Interactive Graph Exploration and Visualization in Browsers.” ACM CIKM, 2020.

[8] Nilaksh Das, Siwei Li, Chanil Jeon, Jinho Jung, Shang-Tse Chen, Carter Yagemann, Evan Downing, Haekyu Park, Evan Yang, Li Chen, Michael Kounavis, Ravi Sahita, David Durham, Scott Buck, Duen Horng Chau, Taesoo Kim, and Wenke Lee, “MLsploit: A Framework for Interactive Experimentation with Adversarial Machine Learning Research.” ACM SIGKDD Project, 2019.

[9] Junghwan Kim, Haekyu Park, Ji-Eun Lee, and U Kang, “SiDE: Representation Learning in Signed Directed Networks.” WWW, 2018.

[10] Haekyu Park, Jinhong Jung, and U Kang, “A comparative Study of Matrix Factorization and Random Walk with Restart in Recommender Systems.” IEEE BigData, 2017.

[11] Haekyu Park, Zijie Jay Wang, Nilaksh Das, Anindya S. Paul, Pruthvi Perumalla, Zhiyan Zhou, and Duen Horng Chau, “SkeletonVis: Interactive Visualization for Understanding Adversarial Attacks on Human Action Recognition Models.”, Under review of AAAI 2021 Demo.

[12] Szigeti, B., Gleeson, P., Vella, M., Khayrulin, S., Palyanov, A., Hokanson, J., ... & Larson, S. (2014). OpenWorm: an open-science approach to modeling Caenorhabditis elegans. *Frontiers in computational neuroscience*, *8*, 137.

Essay:
Research Impact

Essay: Research Impact

Looking back at my undergraduate years, it has been the most essential time of my life when I deeply understood myself and made a philosophy I live by. I live to be happy, and I feel happy when I engage with others and make a positive impact on them.

During my undergraduate, I have taught more than twenty people how to code. Although their backgrounds were different, all of my students had one thing in common: they were eager to learn modern technologies such as artificial intelligence (AI), but held the misconception that AI is “too high tech” for them to learn despite their strong potential to succeed. Luckily, they actually learned the techniques very fast, once they gained the initial understanding of fundamentals. I truly enjoyed the moments whenever they realized that they actually can learn.

My personal interests in helping my students learn the technologies naturally evolve into my research interests to help broader audience understand, learn, and leverage the modern technologies such as Deep Neural Networks (DNNs). Specifically, I am executing my vision by:

(1) Developing scalable visual discovery for trustworthy and interpretable AI
(2) Enabling actionable protection and troubleshooting AI models by discovering and understanding model defects
(3) Equalizing AI access and education opportunities by designing effecting visual interfaces

During my PhD, I have been developing interactive visual data analytics for helping broader range of people understand, learn, trust, and ultimately leverage the AI technologies. My tools have been contributing to the society in many ways; for example, they have been adopted by a rapidly increasing number of courses and institutions around the world, while helping thousands of beginners and practitioners more easily learn and interpret AI models. They were published at top tier visualization and human computer interaction conferences such as IEEE VIS and CHI.

My long term goal is to build AI ecosystem, where people from diverse backgrounds, skill sets, and domains would work harmoniously to invent next AI-powered innovations. I believe my research for promoting interpretable, trustworthy, and easily learnable AI will enrich such AI ecosystem.

Essay:
Leadership Experience

Essay: Leadership Experience

I believe effective leaders build an environment for team members to succeed. The well-organized environment will help the members do not lose their power to proceed, even if they are under difficult circumstances.

As of 2020, I am a representative of Korean graduate students in College of Computing at Georgia Tech, leading about 50 students. Usually, the representatives host several social events every semester, where the students enjoy a variety of outdoor activities, have dinners together, and develop their social connections.

Unfortunately this year, due to the unprecedented COVID-19 pandemic, every social events had to be canceled. This means that students would miss their opportunities to expand their social networks. More serious than that is “new students” would not even open up any opportunity to start building their network in a new culture, let alone expanding the network. I took this problem as seriously as I know how important joining an existing network is for the beginners. I remember when I first started my PhD. I received a lot of helpful advice that shaped my career, from preparing for internship interviews, to learning new culture in the new country. I really wanted to pay it forward to the new students who entered the school during these tough times.

As a leader, I though a lot about how can I turn the lemon into a lemonade: what should I do to minimize their loss of opportunities to build social connections and maximize their benefits in this very limited circumstance? My idea was to transform the disadvantages that we are limited to only virtual meetings to the advantages that virtual meetings allow people in everywhere to meet at the same time. Specifically, I created a new type of meeting where not only current students but also alumni in different countries can socialize together. The meeting encouraged alumni to mentor the students. Both the students and alumni really liked this meeting, and it cheered me up to host more mentorship events. There will be about ten during my term. I look forward to helping the students more through the remaining events.

Transcript of current academic records

Transcript of previous academic records