RF Candidate Interest Form

This form collects additional information about you to make it easier for us to determine the best way to evaluate your application and match you with appropriate projects.

Contact Information

We will use your contact information only for the purposes of communicating with you regarding your RF application.

1. Name *

   Please enter your full name as it appears in your application.

   Divyam Goel

2. Email Address *

   Please enter the email address you used to submit your application.

   divyamgoel3@gmail.com

Undergrad Information

If you have or are working towards a dual degree, just enter the information for your entire program here.

3. What is your undergrad major? (e.g., Computer Science, Mathematics) *

Electronics and Communications Engineering

4. Which institution did you or will you receive your undergrad degree from? *

Indian Institute of Technology (IIT) Roorkee

5. Which year did you or will you receive your undergrad degree? *

2022

6. What is your undergrad CGPA normalized to a scale of 10 as of June 2021? If your university provides a GPA on a different scale or a percentage, just rescale them to 10. *

9.16

7. Do you have or are working toward a dual degree? *

- ○ Yes
- ◉ No

Masters Information

Skip this section if you do not have a separate masters degree. If you are have a dual degree, then fill out the information for your entire program in the previous section.

8. Do you have or are you working towards a masters degree? *

- ○ Yes
- ◉ No

Areas of Interest

Please choose our top three areas of interest (pick at least one).

9. First area of interest *

Natural Language Processing

10. Second area of interest

AI for Programming

11. Third area of interest

Applied Machine Learning

Statement of Purpose

Please briefly (one or two paragraphs) answer the following questions.

12. Briefly summarize one of the projects (or multiple closely related projects) that you have worked on. Describe your process of thinking through the steps you took in working on the project.

In the summer of 2021, I worked on a project in AI for programming as a Mitacs Globalink Research Intern advised by Prof. Fatemeh H. Fard at The University of British Columbia. I started by reading the publications for all the state-of-the-art models in the field and trying to figure out their limitations. I was particularly fascinated by Microsoft's

field and trying to figure out their limitations. I was particularly fascinated by Microsoft's CodeBERT. CodeBERT was a first attempt at creating a large pre-trained language model for source code, it's main limitation is that the study lacks a mechanism to effectively and efficiently adapt the model to new programming languages. This is problematic as most end users lack the computational resources to run the expensive pre-training procedure. My teammate (a fellow intern) and I hence proposed an empirical study on the bimodality of adapter modules to facilitate efficient cross-modal transfer from large pre-trained neural language models to other language modalities, i.e. source code.

I decided to study adapter modules for this transfer for two reasons - 1) they are non-destructive in nature, i.e., they do not alter the original language model; and 2) they provide a computationally efficient way to incrementally transfer knowledge. To compare the syntactic and semantic capabilities of our model with CodeBERT we evaluated it on the cloze-tests provided in the CodeXGLUE benchmark. CodeXGLUE consisted of two cloze-tests a) cloze test min/max (CT-mm) and b) cloze test all (CT-all). While CT-mm has been developed by experts for an analysis of semantic capabilities of language models, CT-all was collected using a data driven approach and is only described as a generalized extension of CT-mm. Unsatisfied by this discussion, I conducted an experiment involving the generation and analysis of the abstract syntax trees for the code samples of all six programming languages in the dataset. I then extracted all the named entities, known as code entities, for the entire CT-all vocabulary using their labels from ASTs. Using this experiment I discovered that the masked words in CT-all were all "identifiers" which indicated that this test studies the syntactic capabilities of the language models. We also tested MODE-X on the downstream task of Code Clone Detection (CCD). Interestingly, we were unable to find a python specific CCD dataset. To

13. Why do you want to apply for MSRI RF program?

My main motivation behind applying for the MSRI Research Fellowship program is to extend my repertoire as a researcher in the field of Machine Learning and AI and in doing so, qualify myself for a graduate (PhD) program at one of the top universities. Participation in this fellowship program, in my view, is exactly what I am prepared for. I have spent tremendous time and energy trying to hone my potential in this field and now I am afraid to waste it. I am excited to learn in the company of experts and firmly believe that MSRI offers a unique blend of a fast paced innovative environment, support and motivation to excel. These qualities make the prospect of this program even more exciting. Another important reason behind my application is that I hope to extend my view of working in a multicultural research and teaching environment to in-person

14. What, if any, are your career goals for the future? Describe a problem that

you feel needs attention, and what do you see as the broad directions to approaching it.

My short term goal is to pursue a PhD, in embodied AI or a related field, at one of the top schools (globally) to bolster my long term goal of using AI to build products that have a significant impact on our community. I am interested in the application of machine learning, natural language processing and computer vision for the development of embodied agents. Simply put, an agent needs to see and communicate for it to be fully autonomous or interactive in nature. Despite the tremendous progress in embodied AI research, the biggest obstacle continues to be the fragmentation of feature representations to learn specific tasks/environments. In the future, I want to pursue the idea of unified task agnostic representation learning for embodied semantics. Essentially, I want to draw parallels with how humans learn to transfer skills learned in a given environment on a particular task to multiple settings and tasks (transfer learning in AI). Additionally, embodied agents today are trained using imitation learning (specifically behavioural cloning) or reinforcement learning. Both these algorithms fall prey to a phenomenon called "causal confusion" which is a huge obstacle in the pursuit of fully autonomous agents. In the future, I would like to explore neural continuous differential

Optional Offline Coding Assessment

For quite a few projects, the role of the RF will include software development. To enable you to better express your coding skills, we have designed an offline coding assessment. You can read about the assessment and submit your response here: https://forms.office.com/r/pcF0ttnmm8.

For projects that do not require software development, this assessment is will not play any role. For all other projects, this assessment will simply help us better evaluate your coding skills in an offline setting. However, this assessment is completely optional. Your application will be fully considered regardless of whether you submit this assessment or not.

15. Are you planning on submitting this assessment? If so, we will send you a reminder email before the application deadline.

- (•) Yes
- ( ) No