Statement of Purpose

1 Overview

I am Tao Shi, an applicant for the PhD program in Electrical Engineering at McGill University. I am currently pursuing a Master of Science degree in Data Science and Information Technology at Tsinghua University. Through this statement of purpose, I aspire to provide a comprehensive discussion of my research interests and experiences, outline my proposed research plans for the PhD program at McGill University, express my enthusiasm for potential collaborations with esteemed faculty members, and articulate my long-term career goal.

2 Research Interests and Experiences

Human languages serve as a fundamental tool for communication, while vision enables us to visually perceive the world around us. Motivated by my fervent passion for understanding language and vision from a computational perspective, my research interests are primarily centered around **the intersection of natural language processing (NLP), computer vision (CV), and multimodal learning**. Specifically, my Master's research has been dedicated to two pivotal vision-language tasks: **emotion recognition in conversations** and **video grounding**.

2.1 Emotion Recognition in Conversations

Emotion recognition in conversations (ERC) is a rapidly evolving task that spans both multimodal NLP and vision-language learning, which aims to accurately classify the emotion of each utterance in a conversational video.

2.1.1 Project 1: MultiEMO

Most existing ERC approaches focus on modeling speaker and contextual information based on the textual modality, while the complementarity of multimodal cues has not been well leveraged. Furthermore, state-of-the-art ERC models have difficulty classifying minority and semantically similar emotion categories. To address these challenges, I proposed a novel attention-based correlation-aware multimodal fusion framework named **MultiEMO**, which effectively integrated multimodal information by capturing cross-modal mapping relationships across textual, audio and visual modalities based on stacked bidirectional multi-head cross-attention layers. Moreover, the difficulty of recognizing minority and semantically hard-to-distinguish emotion classes was alleviated by the proposed sample-weighted focal contrastive (SWFC) loss. Extensive experiments on benchmark ERC datasets demonstrated that MultiEMO consistently outperformed existing state-of-the-art ERC approaches in all emotion categories, with significant improvements in minority and semantically similar emotions. This work, titled MultiEMO: An Attention-Based Correlation-Aware Multimodal Fusion Framework for Emotion Recognition in Conversations, has been **published at the *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)***, with me as the **first author**.

2.1.2 Project 2: SSLCL

Despite MultiEMO has achieved state-of-the-art performances, it suffers from a limitation: due to the underlying class-imbalanced issue with ERC datasets, SWFC loss requires a large batch size to ensure that each training sample has at least one positive pair within the batch, which can be computationally expensive. To overcome this limitation, I proposed an efficient and model-agnostic framework named **supervised sample-label contrastive learning with soft-HGR maximal correlation (SSLCL)**, which eliminated the need for a large batch size and could be seamlessly integrated with existing ERC models without introducing any model-specific assumptions. Specifically, I introduced a novel perspective on utilizing label representations by projecting discrete labels into dense embeddings, and formulated the training objective to maximize the similarity between sample features and their corresponding ground-truth label embeddings, while minimizing the similarity between sample features and label embeddings of disparate classes. Moreover, I innovatively adopted the soft-HGR maximal correlation as a measure of similarity, and effectively leveraged multimodal cues of utterances as data augmentations. Extensive experiments on ERC benchmark datasets demonstrated the compatibility and superiority of SSLCL compared to existing state-of-the-art SCL methods. This work, SSLCL: An Efficient Model-Agnostic Supervised Contrastive Learning Framework for Emotion Recognition in Conversations, with me as the **co-first author**, is currently **under review at the *The 38th Annual AAAI Conference on Artificial Intelligence*** and has received **multiple positive reviews**.

2.2 Visual Grounding

Video grounding (VG) is a fundamental vision-language task, which focuses on retrieving a temporal moment from an untrimmed video that semantically aligns with a sentence query.

2.2.1 Project 3: DiffusionVG

The majority of existing approaches in VG rely on a single-shot or fixed-step prediction strategy, which overlooks the crucial role of a systematic refinement strategy in improving the initial prediction, leading to suboptimal localization of the target moment. To address this issue, I proposed **DiffusionVG**, a novel framework that formulated VG as a conditional generative task using diffusion models, allowing for iterative refinements of predicted spans through the reversed denoising diffusion process. During training, DiffusionVG progressively added noise to the target span with a fixed forward diffusion process, and learned to recover the target span in the reverse diffusion process. During inference, DiffusionVG generated the target span from Gaussian noise inputs through the learned reverse diffusion process, conditioned on video-sentence representations. Extensive experiments on CharadesSTA and ActivityNet Captions benchmarks demonstrate the superiority of DiffusionVG against existing state-of-the-arts. This work, titled Exploring Iterative Refinement with Diffusion Models for Video Grounding, where I am the **co-first author**, is currently **under review at the *The 38th Annual AAAI Conference on Artificial Intelligence*** and has garnered **universally positive feedback from reviewers**.

3 Proposed Research Plans

Building upon my extensive research experiences in vision-language tasks, I am dedicated to continuing my work on visual-language learning during my PhD program, with a primary goal of contributing towards **multilingual, multitasking, and trustworthy vision-language learning**.

3.1 Research Plan 1: Towards Multilingual Vision-Language Learning

Despite the abundance of languages spoken worldwide, most existing vision-language datasets are English-based, which significantly hinders the generalizability of vision-language models (VLM) to non-English, low-resource languages. To overcome this limitation, recent studies have created large-scale multilingual video-text corpus, such as Multi-HowTo100M. Nevertheless, there still exists unsolved challenges: (1) **Noisy pretraining.** Existing large-scale multilingual vision-language corpus are collected based on user-generated videos, which contain a considerable amount of noisy and incomplete information, resulting in inferior cross-lingual generalizability of pretrained multilingual VLMs; (2) **Challenging finetuning.** The majority of downstream video-language tasks have no annotations for non-English languages, making it difficult for efficient task-specific finetuning.

To address the challenge of noisy pretraining, my aim is to achieve noise-robust pretraining through two key perspectives: (1) Implementing efficient and reliable data cleansing and sample selection strategies; (2) Developing effective architectural designs and regularization techniques to minimize the negative impact of noisy pairs.

In order to tackle the challenge of task-specific finetuning, my future research will concentrate on two specific aspects. Firstly, I plan to create fine-grained multilingual datasets for commonly-used downstream vision-language tasks. **Currently, I am working on constructing a multilingual multi-label long-context dataset for ERC**. Secondly, I intend to explore zero-shot cross-lingual transfer learning algorithms to facilitate effective cross-lingual learning in the absence of labeled data.

3.2 Research Plan 2: Towards Multitasking Vision-Language Learning

Video-text pretraining (VTP) has shown promising performances on various downstream tasks, including video retrieval, video question answering, and video captioning. However, **the majority of existing VTP methods are limited to retrieval-based downstream tasks**, while the transfer potential of VTP to localization-based tasks, such as VG and temporal action segmentation, remains largely underexplored. To bridge this gap, my future research aims to achieve multitasking VTP that is compatible with both retrieval-based and localization-based downstream tasks.

3.3 Research Plan 3: Towards Trustworthy Vision-Language Learning

The ability to reason over vision and language in a logical and consistent way is essential for building trustworthy VLMs. While existing VLMs have shown impressive performances in reasoning implicit sociocultural backgrounds associated with an image, such as its times and location, they fail to demonstrate strong reasoning capabilities in two key aspects: (1) **Compositonal reasoning.** Existing VLMs struggle at accurately answering subquestions that are parts of the main problem; (2) **Visual reasoning.** Even the best performing state-of-the-art VLMs, such as BLIP-2 and OFA-Large, exhibit inadequate visual reasoning performance and lack reasoning consistency. Inspired by the huge success of instruction tuning in large language models (LLM), I aim to enhance VLMs' capacities in compositional reasoning and visual reasoning through instruction-based learning techniques in my future research endeavors, contributing towards more trustworthy vision-language learning.

4 Collaboration with Faculty Members

Considering my research interests and proposed plans to contribute towards a **multilingual, multitasking, and trustworthy vision-language learning** during my PhD program, I am particularly interested in collaborating with **Professor Chengzhi Mao.** Professor Mao's distinguished work in computer vision, robust machine learning, and foundation models closely align with my interdisciplinary research plans. Therefore, I am confident that his guidance and mentorship will significantly enrich my academic journey, empowering me to make substantial contributions to the domain of video-language learning.

5 Long-Term Career Goal

After successfully completing my PhD degree at McGill University, my plan is to advance further in my academic journey by perusing a postdoctoral research position. Driven by dedication and self-motivation, my ultimate aspiration is to become an exceptional and influential researcher within the machine learning community. With a clear vision for my future, I am resolute in my determination to join the esteemed Department of Computer Science at McGill University as a PhD student.