Research Collaborations : Academic Year 2019-20

Sai Krishna Rallabandi
Carnegie Mellon University
srallaba@cs.cmu.edu

Abstract

I love collaborations. Almost everything we are proud of accomplishing as species has come out of excellent collaborations. I have been fortunate to be part of some wonderful collaborations myself[12]. To me collaborations is one of the infinity stones[11] I am collecting as a researcher++[10]. In this document, I present the directions in which I am looking for collaborations this Fall and beyond. Of course, you know that I am always open to doing things not on this document as well.

1. Introduction

In this writeup I will outline the research directions from a medium term perspective. I will briefly mention the context and then describe the three(problems in section 2, through 4 including what has been done so far and what is in store. Will end this document with a look ahead. If you think this document is too long, just read this line and close the pdf: *Ping me if you have time. Lets work on something cool - whatever - and have fun along the way.*

1.1. Context

The overarching goal of my research career is to bring technologies closer to life. During PhD - preliminary stage of this career - the goal is more narrow and focused : To bring *language technologies* closer to *humans*. I believe this is an important step for a number of reasons, primary of which has to do with the trajectory of Artificial Intelligence (AI). I believe that AI - latest addition to human toolbox - has the potential to be a horizontal enabling layer and massively benefit life in all forms. However, today it is very far from making progress towards fulfilling this potential. I posit that there are three kinds of applications AI is affecting: (1) *Low risk and Low stake*: These include games such as Go, Chess and applications such as speech recognition and synthesis. (2) *Medium risk and Medium stake:* These include applications such as supply chain management, dialog systems, etc. and (3) *High risk and High stake:* Autonomous driving, Diagnosing diseases and Exoplanet Terraforming. Most depictions of a sufficiently advanced human civilization involve AI as a corner stone technology and perhaps rightfully so. But how can we be certain that AI has reached an acceptable level such that it can be deployed in real life? In my thesis, I argue that any technology choosing to impact lives needs to satisfy a trinity of requirements: (1) **Scalability** (2) **Flexibility** and (3) **Explainability**. Consider electricity as example. It is impossible to imagine electricity having the effect it had on our lives if it could only power light bulbs and not televisions. Equally unlikely is the scenario if we could not explain the failure scenario and debug. Similar arguments can be made for other key enabling technologies such as internet. It has to be noted that AI needs to primarily address and solve these three issues before even being considered subject to interesting (and important) topics like ethics and bias.

I argue that AI is currently in a transition towards the second category - medium risk and medium stake. Language Technologies are at the forefront of this category: AI systems today, sure, can identify objects in images and video, recognize and convey information via speech, translate across multiple languages providing immense benefit to people across a multitude of domains such as industry, government as well as society. But, they are also inherently complex due to the presence of human element and the consequent stochasticity. In addition, Language Technologies have another fascinating characteristic making them special: They are both rich in terms of quantity and diversity of tasks. Hence they are the epitomy of medium risk medium stake applications. Consequently, we can employ such applications as sanity tests in gauging the ability of AI in fulfilling its potential to impact our lives. In the context of language technologies, the three challenges persist: (1) *Scalability*

- These technologies are currently only accessible in a handful number of languages around the planet. In order to have a meaningful impact it is imperative that such technologies need to at least exist in many more languages. (2) *Flexibility* - Although deep learning based systems outperform their shallow learning counterparts in terms of quality, they still pale away in terms of flexibility and controllability. (3) *Interpretibility* - Almost every deep learning system today is akin to a black box: we can neither interpret nor justify predictions by most of these models. In my thesis, I argue that a principled solution to address all these three issues can be obtained by focusing on a single thing - de entanglement of relevant information[1]. Specifically, I argue that we need to incorporate controlled de-entanglement - the ability of AI models to de entangle the relevant factors of variation in the observable data - as first class object to achieve the goal of addressing the three issues simultaneously. In the subsequent sections I will expand on each of these issues.

2. Scalability - Languages, Domains and CodeMixing

A major bottleneck in the progress of many data-intensive language processing tasks such as speech recognition and synthesis is scalability to new languages and domains. Building such technologies for unwritten or under-resourced languages is often not feasible due to lack of annotated data or other expensive resources. I posit that there are three distinct categorizations that pose challenges in terms of scalability: (1) Unwritten languages and low resource scenarios (2) Code switching and other non native speech phenomena and (3) Different domains

2.1. Unwritten Languages

Let us consider building speech technology for unwritten or under-resourced languages. A fundamental resource required to build speech technology stack in such languages is phonetic lexicon: something that translates acoustic input to textual representation. Having such a lexicon - even if noisy and incomplete - can help bootstrap speech recognition and synthesis models which in turn enable other applications such as key word spotting.

2.1.1 Work done so far - Unsupervised Acoustic Unit Discovery

We have employed controlled de-entanglement for unsupervised acoustic unit discovery in the context of our submission to ZeroSpeech Challenge 2019 [13]. We make an observation that articulatory information about speech production presents a discrete set of independent constraints. For instance, manner and place of articulation are two articulatory dimensions characterized by discrete sets(labial vs dental, etc). Based on this, we condition the prior space to conform to articulatory conditions by using a bank of discrete prior distributions.

2.1.2 Work Ideas - Tracking trajectory from lyrics of the latent space to dance of the Decoder

I really have not done justice to visualizing and squeezing everything out of latent space. There are some cool examples like making Alan speak Hindi [2] and generating sarcasm but there is a lot to do here.

2.2. Code switching and other non native Speech phenomena

Code-switching (or mixing) refers to the phenomenon where bilingual speakers alternate between the languages while speaking. It occurs in every multilingual society such as India, Singapore, etc. It is used both to express opinions as well as for personal and group communications. This can go beyond simple borrowing of words from one language in another and is manifested at lexical, phrasal, grammatical and morphological levels. The technology today - from speech processing systems through conversational agents - assume monolingual mode of operation and do not process code-switched content. However, the mixed content is intuitively the most important part in the content. Since the systems are now handling conversations, it becomes important that they handle code-switching.

In simple terms, we need the following:

- Speech Synthesis systems that can synthesize codemixed content.
- Speech Recognition systems that can recognize codemixed content.

[1] I wont bore you with math here. You can read more here about the framework itself https://t.ly/GR7D0

[2] https://t.ly/GR7Rz Okay. That does not sound Alan completely. Fair enough. But Alan is a Scottish speaker

2.2.1 Work done so far - Speech Synthesis of codemixed content

We can do this in multiple ways. First lets look at this problem with applications in mind: While code mixing happens across different scenarios, there are two semi formal scenarios that might make sense to target as first applications: (a) News paper headlines where the content is primarily in native language (say, Hindi) with English words interspersed and (b) Navigation instructions where the content is primarily in English with named entities in the native language. We have handled (a) in [14, 2]. Now let's look at this from the perspective of available data. Speech synthesis typically uses clean recordings from a speaker in a controlled settings. Given that code mixing happens in social scenarios, it is difficult to get speaker data. There will be three scenarios here: (a) When we have data only from one language (b) When we have data from both the languages but monolingual in the language - One records data first in Hindi and later in English (c) When we have data that is truly mixed - YouTube videos with interviews of contemporary stars. We have handled (b) in [14]. Finally let's look at this from the perspective of algorithms. Speech synthesis has at least these three modules: (a) Text processing (b) Acoustic Modeling and (c) Prosody Modeling. The details about these can be found in [14].

2.2.2 Work Ideas - Speech Recognition and Synthesis of codemixed content

Speech has both continuous as well as discrete priors: The generative process of speech assumes a Gaussian prior distribution which is continuous in nature. However, the language which is also present in the utterance can be approximated to be sampled from a discrete prior distribution. Exact manifestation of this in linguistics can be at different levels: phonemes, words, syllables, sub word units, etc. Based on this insight, I show[8] that incorporating priors help encode language independent information thereby facilitating synthesis of code mixed content. In addition to basing priors on knowledge about characteristics, I believe that it is also possible to base them on discovered patterns. In [15], we have discovered several code switching styles based on [6]. Going forward, I plan to incorporate priors based on this style information to help speech recognition models targeted at decoding code switched speech.

3. Flexibility - Global and Local Control in Deep Generative Models

We humans exhibit explicit global as well as fine grained control over how we deliver information. Such modulated content enables effective communication in various scenarios that we face. The goal is to build AI models that can mimic this behavior.

3.1. Global Control

3.1.1 Image Captioning - What has been done

In the context of image captioning, an interesting observation is that both the involved modalities - textual even though primarily symbolic and visual even though primarily spatial - are characterized by distinct discrete and continuous factors of variation. For instance, distinct objects or entities would intuitively perhaps be better represented by discrete variables, while their spatial location and relationships between them might be represented by continuous variables. Therefore, we split the latent prior space[16] used for approximating the posterior distribution into continuous and discrete counterparts. Pressurizing the model to encode such prior information into the latent space provides us the flexibility to control the generative process by pinging different latent states during inference.

3.2. Work Ideas - Global + Local Control

Consider speech synthesis from given text. Long form text is characterized by rich natural variations in terms of content, persona, speaking style, etc. Therefore the voice talent is expected to reflect these prosodic variations in recordings. However typical generative models of speech end up normalizing such variability due to a multitude of reasons. Consequently, the models end up with bland generation during inference. On the other hand, social media and web 2.0 has enabled an outburst of highly varied data at an unprecedented rate. Therefore, I am interested in building on the solutions proposed in subsection 2 and design techniques that can leverage such resources to build models that allow variability and flexibility. Specifically, I propose to build speech synthesis systems that automatically discover prosodic cues using controlled de-entanglement and accomplish fine grained emphasis at desired level. I have done a version of this already[3]. But there is lots to do here.

[3] https://t.ly/NExyy

4. Justification

Language and vision are inherently composite in nature. For example different questions share substructure viz *Where is the dog?* and *Where is the cat?* Similarly images share abstract concepts and attributes viz *green pillow* and *green light*. Hence it is vital not only to focus on understanding the information present across both these modalities, but also to model the abstract relationships so as to capture the unseen compositions of seen concepts at test time. However, accomplishing this is a deceptively non trivial task and might lead to models learning just surface level associations[1, 3, 5]. Therefore, interpretibility is an important facet while building models targeted at such tasks. I present a case that flexible generative models provide additional information to improve performance in such tasks. Further, I hypothesize that when optimized using either a disjoint learning mechanism or a different divergence function, such models can also act as justifying modules for the task at hand. To ground this argument, I am looking at two example applications that employ flexible models mentioned in the previous subsections: (1) Visual Question Answering System(VQA) that receives additional information in the form of targeted captions. I propose to use Reward Augmented Maximum Likelihood[7] to generate and integrate captions in the framework of Visual Question Answering. High level idea is that tying the reward function to length of the generated caption forces the model to encode most relevant information thereby acting as justification to the selected answer. (2) Based on similar insights, I propose to apply[7] to obtain and integrate speech recognition transcripts in the context of Acoustic Topic Identification System.

5. Misc Projects

I work on a ton of miscellaneous projects. To me this is equivalent to the 20 percent time at Google. Here are a few:

5.1. JUDITH[4]

I am building an assistant inspired by EDITH from spiderman. The current goal of this project is to facilitate building an intelligent agent like Jarvis that can accompany Research. The assistant currently does minimal things like tracking experiment status[5], etc. Over the year I want to include atleast two components:

- (1) *Model Visualization* using Augmented Reality[4] I think this will be massively useful for professional Masters students and undergrads looking at industry.
- (2) *Text2Code*: Generating code based on Natural Language input.

5.2. Detection of interesting paralinguistic phenomena from speech

- (1) Intent Recognition[9].
- (2) Sleepiness Level Detection[17].

6. Look Ahead

These are some of the additional projects I am looking at:

- *Local Control in Image Captioning*: Adding POS tags to words as additional features during training. Goal is to have explicit control over how the caption gets generated. This is an extension of qF0 based Focused TTS to text. While we can only play with style(focus) in TTS, caption generation seems a richer task giving us freedom to modify content too.
- *Acoustic Question Answering:* Applying variational inference to identify the modes of variation, helping answer the questions.
- *AutoDL*[6]*:* Building algorithms that effortlessly scale.
- *Galactic Phonest:* Building resources for low/zero resource languages
- *Review Papers:* Writing reviews of concepts, topics and paradigms. Summarizing involves a lot of skill . I think this will be useful for 1st year PhD students and MLTs looking to convert to PhD.

[4]https://www.cs.cmu.edu/ srallaba/ProjectAssistCore
[5]Check samples: http://www.cs.cmu.edu/ srallaba/ProjectDeveloperCut/
[6]https://autodl.chalearn.org/

References

[1] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*, 2017.

[2] Khyathi Raghavi Chandu, Sai Krishna Rallabandi, Sunayana Sitaram, and Alan W Black. Speech synthesis for mixed-language navigation instructions. *Proc. Interspeech 2017*, 2017.

[3] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. *arXiv preprint arXiv:1712.02051*, 2017.

[4] Google. Augmented Reality core. Google IO2019. URL: `https://developers.google.com/ar/`.

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. Metrics for modeling code-switching across corpora. In *Proceedings of INTERSPEECH*, 2017.

[7] Mohammad Norouzi, Samy Bengio, Navdeep Jaitly, Mike Schuster, Yonghui Wu, Dale Schuurmans, et al. Reward augmented maximum likelihood for neural structured prediction. In *Advances In Neural Information Processing Systems*, 2016.

[8] Sai Krishna Rallabandi and Alan Black. Variational Attention using Articulatory priors for generating code mixed speech using monolingual corpora. In *in proceedings of Interspeech*, 2019.

[9] Sai Krishna Rallabandi, Carla Viegas, Bhavya Karki, Eric Nyberg, and Alan Black. Investigating utterance level representations for detecting intent from acoustics. *Interspeech 2018*, 2018.

[10] SaiKrishna Rallabandi. On researcher++. 2018. URL: `https://www.quora.com/q/mntpepgrzwykelgg/On-Researcher`.

[11] SaiKrishna Rallabandi. Project Cupcake. 2018-2021. URL: `http://www.cs.cmu.edu/~srallaba/ProjectCupcake/`.

[12] SaiKrishna Rallabandi. VQA Retro Analysis. Fall 2018. URL: `http://www.cs.cmu.edu/~srallaba/VQA/retro_analysis.html`.

[13] SaiKrishna Rallabandi and Alan Black. Submission from CMU to ZeroSpeech Challenge 2019. 2019.

[14] SaiKrishna Rallabandi and Alan W Black. On building mixed lingual speech synthesis systems. *Proc. Interspeech 2017*, 2017.

[15] SaiKrishna Rallabandi, Sunayana Sitaram, and Alan W Black. Automatic detection of code-switching style from acoustics. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, 2018.

[16] Nidhi Vyas, SaiKrishna Rallabandi, Lalitesh Morishetti, Eduard Hovy, and Alan W Black. Learning disentangled representation in latent stochastic models: A case study with image captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[17] Rallabandi Sai Krishna Wu, Peter, Eric Nyberg, and Alan Black. Ordinal triplet loss: Investigating sleepiness detection from speech. *Interspeech 2019*, 2019.