On Controlled De Entanglement for Language Technologies

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

Abstract

Latest addition to the toolbox of human species is Artificial Intelligence(AI). Thus far, AI has made significant progress in low stake low risk scenarios such as playing Go and we are currently in a transition toward medium stake scenarios such as Visual Dialog. But how can we be certain that AI has reached an acceptable level to be deployed in real life? In this write up, I posit that any technology choosing to impact lives needs to satisfy three things: (1) Scalability (2) Flexibility and (3) Explainability. I argue that a principled solution to address all these three issues can be obtained by focusing on a novel property I refer to as 'De- Entanglement'. In my thesis, I present an approach to incorporate controlled de-entanglement as first class object and argue that this helps in the transition. I present mathematical analysis from information theory to show that employing stochasticity leads to controlled de-entanglement of relevant factors of variation at various levels. Based on this, I highlight results from initial experiments that depict efficacy of the proposed framework with respect to generation of code mixed speech in terms of ***scalability****, emphatic Text to Speech in terms of* ***flexibility*** *and Visual Question Answering in terms of* ***Explainability****.*

1. Introduction

The overarching goal of my research career is to bring technologies closer to life. My favourite technology is Artificial Intelligence(AI) and I seek to answer to the question 'How can we integrate AI into daily life?'. During PhD - preliminary stage of this career - the goal is more narrow and focused: To bring *language technologies* closer to *human life*. Specifically I am interested in answering the question 'What are the scientific challenges to be solved so that language technologies can be seamlessly integrated into everyday life?' 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: It is impossible to imagine electricity having the effect it had on our lives if it could only power light bulbs and not other devices. Similar arguments can be made for other key enabling technologies such as internet. Although most depictions of a sufficiently advanced civilization have AI as a cornerstone technology, it has thus far has not had remarkable success in either of the three mentioned requirements. We need to primarily address and solve these three core technical issues before moving on to interesting (and important) topics like ethics and bias. In my thesis, I present a framework aimed at addressing these three issues in the context of language technologies employing deep learning, perhaps the most promising approach to build AI systems.

I posit that there are three kinds of applications AI can de deployed in: (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 visual dialog, searching / summarizing media content and (3) *High risk and High stake:* Autonomous driving, Diagnosing diseases and Exoplanet Terra forming. 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. Systems today can identify objects in images and video, recognize and convey information via speech, translate across multiple languages. But, as we move on to slightly more reasonably intuitive as well as interesting tasks like visual dialog, the complexity involved becomes deceptively non trivial due to the presence of human element and the consequent stochasticity. Language Technologies are also rich in terms of both quantity and diversity of tasks. Hence they can be considered epitome of medium risk medium stake applications. I therefore posit that we can employ such applications as sanity tests to gauge 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) *Explainability* - 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 property I refer to as 'De-Entanglement'[1]. Specifically, I argue that we need to incorporate controlled de-entanglement - the ability of AI models to isolate the relevant factors of variation in the observable data - as first class object to achieve the goal of addressing the three issues simultaneously.

1.1. Controlled De Entanglement

I propose that designing learning paradigms such that we explicitly control isolation of relevant factors of variation while marginalizing the nuisance factors of variation leads to massive improvements. I refer to this as controlled De-Entanglement. Such an approach, I claim, leads to further advantages in the context of both generative processes: in terms of generation of novel content and discriminative processes: in terms of robustness to noise and attacks. Let us consider a typical deep learning architecture such as AlexNet[10]. It is characterized by a series of convolutional layers (feature extraction module) followed by a pooling layer and a SoftMax layer(classification module). Note that while I mention AlexNet as an example, this abstraction can be extended to most sequence to sequence architectures with encoder as feature extraction module and decoder as the classification module[19] across modalities and tasks. It can be shown that the pooling layer acts as information bottleneck[20] module in such architectures. I point out[13] that in case of conventional Seq2Seq architectures deployed today, attention plays the role of information bottleneck module regulating the amount of information being utilized by the decoder. In [13, 21] I show that this module controls optimization in encoder decoder models leading to (1) Disentanglement of Causal Factors of variation in the data distribution (2) Marginalization of nuisance factors of variation from the input distribution. In case of models that employ stochasticity, two more effects can be observed: (a) Posterior collapse or Degeneration due to powerful decoders and (b) Loss of output fidelity due to finite capacity decoders. In current architectures, marginalization and disentanglement are realized implicitly and often lead to (a) and (b) when deployed in practise.

I argue that explicitly controlling what and how much gets de-entangled [3] is better than implicit disentanglement as is followed today[11]. The most popular approach to obtain disentanglement in neural models is by employing stochastic random variables. This approach provides flexibility to jointly train the latent representations as well as the downstream network. It has been observed that the latent representations resemble disentangled representations under certain conditions [6, 3, 7, 2]. Note that although obtaining such degenerate representations is considered typical, it is not the only manifestation of disentanglement: it also manifests as continuous representations[18] and other abstract phenomena(e.g. grounding). I identify four ways to computationally control de-entanglement in encoder decoder models, by employing (1) suitable priors (2) additional adversarial or multi task objectives (3) an alternative formulation of probability density estimation and (4) a different objective of divergence. In my thesis, I present experimental findings from various tasks to show that the proposed framework has the ability to address all the three problems mentioned in above: *scalability* via priors based controlled de-entanglement, *flexibility* via priors, adversarial training and *interpretibility* via alternative formulation of density estimation and objective of divergence.

2. Scalability - Languages, Domains and CodeMixing

A major bottleneck in the progress of many cornerstone language processing tasks 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 am interested in two distinct categorizations that pose challenges in terms of scalability: (1) Unwritten languages and low resource scenarios and (2) Code switching and other non native speech phenomena. I have worked on both these areas in the context of speech processing thus far. I found speech specifically interesting since it 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.

2.1. 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 multilingual societies such as India, Singapore, etc. and is used both to express opinions as well as

[1] I More about the framework itself https://t.ly/RjxLD

for personal and group communications. This phenomena 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 effectively process code-switched content. However, the mixed segment is usually the important part in the content. Since the systems are now handling conversations, it becomes important that they handle code-switching.

2.1.1 Work done so far - Speech Synthesis of codemixed content and Categorization of Switching Styles

While code mixing happens across different scenarios, there are two semi formal scenarios that perhaps naturally render themselves as first target applications in the context of TTS: (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. I have investigated several approaches for these scenarios in terms of acoustic and prosodic models in [16, 4]. Building mixed lingual systems is also interesting from the perspective of available data: Voice building process for monolingual scenario typically uses clean recordings from a speaker in a controlled settings. Given that code mixing happens in social scenarios, it is difficult to get clean and natural speaker data. I posit that 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. I have investigated approaches to handle (a) in [13] and (b) in [16]. I show[13] 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 [17], we have discovered several code switching styles based on [9]. Going forward, I plan to incorporate priors based on this style information to help speech recognition models targeted at decoding code switched speech.

2.2. Work done so far - Unsupervised Discovery of Acoustic Units for 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. We have employed controlled de-entanglement for unsupervised acoustic unit discovery in the context of our submission to ZeroSpeech Challenge 2019 [15]. 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. This submission shows that incorporating priors to explicitly de-entangle relevant factors of variation results in learning effective units compared to models with implicit disentanglement.

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. This enables us communicate more effectively in a conversation. The goal is to build AI models that can mimic this behavior. I have thus far worked on image captioning in the context of global control and emphatic text to speech in the context of fine grained control.

3.1. Work done so far - Global Control

3.1.1 Image Captioning

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[21] 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 done so far - Local Control

3.2.1 Word Level Emphasis in TTS

I am interested in investigating approaches to incorporate automatically derivable information from speech into the model architecture for better modeling and controlling prosody. In [14][2], I present an algorithm aimed at disentangling heuristics about tonal information to accomplish local control. We first quantize fundamental frequency(F_0) - a feature highly correlated with prosody - into multiple bins. In other words, continuous F_0 values were mapped to their discrete bin indices, resulting in ordinal F_0 values for each utterance. We then incorporate the resultant discretized ordinal information at the phone level into our speech generation architecture in two different scenarios: (1) Explicit labeling scenario where we input the quantized F_0 values alongside the phoneme embeddings as input to the encoder and (2) Implicit incorporation as an additional task to predict these quantized F_0s at the output of encoder. The model was optimized using ordinal triplet divergence [22]. We show that our approach generates appropriate emphasis at word level and significantly outperforms AuToBI in terms of flexibility.

4. Explainability - 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, 5, 8]. 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[12] to generate and integrate captions in the framework of Visual Question Answering. I hypothesize 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[3]. (2) Based on similar insights, I propose to apply[12] to obtain and integrate speech recognition transcripts in the context of Acoustic Topic Identification System.

5. Conclusion

I argue that any technology choosing to impact lives needs to satisfy three things: (1) Scalability (2) Flexibility and (3) Explainability. In my thesis, I posit that principled solution to address all these three issues in the context of language technologies can be obtained by focusing on a novel property I refer to as 'De-Entanglement'. I show that building systems by incorporating such explicit isolation of relevant factors of variation in the data leads to massive improvements.

6. Acknowledgements

I sincerely thank Alan W Black for his invaluable guidance towards my dissertation. I thank student volunteers for taking part in the various subjective evaluations. I am very grateful to my collaborators for providing me an opportunity to work with them and the reviewers for useful feedback at various stages in my research career. I am extremely thankful to ISCA and various sponsors for organizing conferences/workshops that have been immensely useful to me.

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] Abdul Fatir Ansari and Harold Soh. Hyperprior induced unsupervised disentanglement of latent representations. *arXiv preprint arXiv:1809.04497*, 2018.

[3] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in Beta VAE. *arXiv preprint arXiv:1804.03599*, 2018.

[2] Samples can be found here: https://t.ly/NExyy

[3] We have initial results from this approach documented here: https://t.ly/xY1LY

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

[5] 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.

[6] Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

[7] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[8] 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.

[9] 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.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 2012.

[11] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

[12] 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.

[13] 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.

[14] SaiKrishna Rallabandi and Alan Black. EDITH: Emphasis by Disentangling Tonal Heuristics. In *under submission, AAAI*, 2019.

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

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

[17] 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.

[18] Mirco Ravanelli and Yoshua Bengio. Interpretable convolutional filters with sincnet. *arXiv preprint arXiv:1811.09725*, 2018.

[19] David Rousseau and Sotirios Tsaftaris. Data augmentation techniques for deep learning. In *Tutorial Session, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[20] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[21] 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.

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