On Controlled DeEntanglement for Natural Language Processing

Sai Krishna Rallabandi

Language Technologies Institute,
Carnegie Mellon University,
Pittsburgh, PA 15213

`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. In my thesis, I argue that we need to incorporate controlled de-entanglement as first class object to succeed in this 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. I conclude this writeup by a roadmap of experiments that show the applicability of this framework to scalability, flexibility and interpretibility.

1. Motivation of Research

In order to bring technology closer to humans it is important to provide mechanisms for them to interact with the same. Applications like Speech Recognition, Speech Synthesis, Visual Question Answering and Machine Translation to name a few have been shown to be massively useful in this context. Such Language Technologies are both rich in terms of quantity and diversity of tasks as well as inherent complexity due to the presence of human element. Hence they are the epitomy of medium risk medium stake applications forming a bridge between low stake applications such as playing games, spam filtering and high stake applications such as climate control, health related applications. Consequently, we can employ such applications as sanity tests in gauging the ability of AI - the latest addition to human toolbox - in fulfilling its potential to impact our lives. However, there are three broad problems that challenge this promise - (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 this thesis, I argue that employing a slightly different framework of learning - one with controlled de-entanglement of relevant factors of variation as a first class object - has the potential to address all the three issues simultaneously and henceforth rapidly accelerate progress in language technologies.

1.1. Controlled DeEntanglement

Let us consider a typical deep learning architecture such as AlexNet[1]. 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[2] across modalities and tasks. It can be shown that the pooling layer acts as information bottleneck[3] module in such architectures. I point out[4] 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 [4, 5] 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 posit that designing learning paradigms such that we explicitly control de-entanglement of relevant factors of variation while marginalizing the nuisance factors of variation leads to massive improvements. I refer to this as controlled DeEntanglement. Such an approach, I claim, leads to further advantages in the context of both generative processes - generation of novel content and discriminative processes - robustness of such models to noise and attacks.

2. Key Issues identified and/or addressed

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, 7, 8, 9]. Note that although obtaining such degenerate representations is typical, it is not the only manifestation of disentanglement: it also manifests as continuous representations[10] and other abstract phenomena(e.g. grounding). I argue that explicitly controlling what and how much gets de-entangled [7] is better than implicit disentanglement as is followed today[11]. 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 section 1: *scalability* via priors based controlled de-entanglement, *flexibility* via priors, adversarial training and *interpretibility* via alternative formulation of density estimation and objective of divergence.

2.1. Scalability - Discovery of Acoustic Units and Styles

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 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.

2.1.1. 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. Therefore I proposed to employ controlled de-entanglement for unsupervised acoustic unit discovery in the context of our submission to ZeroSpeech Challenge 2019 [12]. 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. Code switching and other non native Speech phenomena

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[4] 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 [13], we have discovered several code switching styles based on [14]. I plan to incorporate priors based on this style information to help speech recognition models targeted at decoding code switched speech.

2.2. Flexibility - Global 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.

2.2.1. Targeted 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[5] 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.

2.2.2. Emphatic Text to Speech

I am interested in investigating approaches to incorporate automatically derivable information from speech into the model architecture for better modeling and controlling prosody. In a sample implementation[1], I have attempted disentangling heuristics about tonal information to accomplish local control. I first quantize fundamental frequency(F_0) - a feature highly correlated with prosody - into multiple bins. I 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 [15]. We show that our approach generates appropriate emphasis at word level and significantly outperforms AuToBI in terms of flexibility.

2.3. Justification

Language is inherently composite in nature. Hence it is vital 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[16, 17, 18]. 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 present two example applications that employ flexible models mentioned in the previous subsections: (1) Visual Question Answering System(VQA) that receivers additional information in the form of targeted captions. I propose to use Reward Augmented Maximum Likelihood[19] to generate and integrate captions in the framework of Visual Question Answering. I show 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[19] to obtain and integrate speech recognition transcripts in the context of Acoustic Topic Identification System.

3. Acknowledgements

---

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

and 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.

4. References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012.

[2] D. Rousseau and S. Tsaftaris, "Data augmentation techniques for deep learning," in *Tutorial Session, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[3] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[4] S. K. Rallabandi and A. Black, "Variational Attention using Articulatory priors for generating code mixed speech using monolingual corpora," in *in proceedings of Interspeech*, 2019.

[5] N. Vyas, S. Rallabandi, L. Morishetti, E. Hovy, and A. 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.

[6] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Advances in Neural Information Processing Systems*, 2018.

[7] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in Beta VAE," *arXiv preprint arXiv:1804.03599*, 2018.

[8] B. Esmaeili, H. Wu, S. Jain, A. Bozkurt, N. Siddharth, B. Paige, D. H. Brooks, J. Dy, and J.-W. van de Meent, "Structured disentangled representations," in *International Conference on Artificial Intelligence and Statistics*, 2018.

[9] A. F. Ansari and H. Soh, "Hyperprior induced unsupervised disentanglement of latent representations," *arXiv preprint arXiv:1809.04497*, 2018.

[10] M. Ravanelli and Y. Bengio, "Interpretable convolutional filters with sincnet," *arXiv preprint arXiv:1811.09725*, 2018.

[11] F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," *arXiv preprint arXiv:1811.12359*, 2018.

[12] S. Rallabandi and A. Black, "Submission from CMU to ZeroSpeech Challenge 2019," 2019.

[13] S. Rallabandi, S. Sitaram, and A. W. Black, "Automatic detection of code-switching style from acoustics," in *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, 2018.

[14] G. A. Guzmán, J. Ricard, J. Serigos, B. E. Bullock, and A. J. Toribio, "Metrics for modeling code-switching across corpora." in *Proceedings of INTERSPEECH*, 2017.

[15] P. Wu, S. Rallabandi, E. Nyberg, and A. Black, "Ordinal triplet loss: Investigating sleepiness detection from speech," in *Interspeech*, 2019.

[16] A. Agrawal, A. Kembhavi, D. Batra, and D. Parikh, "C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset," *arXiv preprint arXiv:1704.08243*, 2017.

[17] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, "Attacking visual language grounding with adversarial examples: A case study on neural image captioning," *arXiv preprint arXiv:1712.02051*, 2017.

[18] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. 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.

[19] M. Norouzi, S. Bengio, N. Jaitly, M. Schuster, Y. Wu, D. Schuurmans *et al.*, "Reward augmented maximum likelihood for neural structured prediction," in *Advances In Neural Information Processing Systems*, 2016.