Studying the Effect of Generalized Entropy Regularization on Question Answering Tasks

Anya Trivedi (aht324), Vishal Kumar(vk2161), Mahima Gaur(mg6827)

Defining the Problem A common problem faced by probabilistic Natural Language Generation (NLG) models is their proneness to produce "peaky" distributions, or distributions that place most of their probability mass on a few candidates. This is a sign of overfitting, and such models are known to have low entropy. An intuitive way to then combat such overfitting would be to train models that have higher entropy, and thus generalize better. In their seminal paper, Meister et al. [1] use two common regularizers to introduce entropy based regularizers- label smoothening and confidence penalties. Label smoothening is a technique to smooth hard target distributions by using a weighted average between hard targets of a data set and uniform distribution over labels [2]. What this effectively does is allow some probabilitiy mass to be "reserved" for unseen words, improving the generalization power of a model. By bringing the approximate emperical distribution of a model close to the uniform distribution, label smoothening can be thought of as a method of increasing entropy- recall how Shannon's theory states that Uniform Distribution is a distribution of maximum entropy. In a similar manner, Confidence penalties also aim to penalize low entropy models. Arguing that these can be though of as regularization techniques, [1] aims to study whether model performance could be explained by its resulting entropy, and introduced a family of Entropy regularizers, of which label smoothening and confidence penalties are specific case.

While the authors of [1] study the effect of "entropy regularizers" on two common NLG tasks- Machine Translation and Abstract Summarization, we aim to study the effect of this family of regularization techniques on another common NLG task- Question Answering. NLG tasks often lead to sparse distributions (for example, we would want the model to assign very low probability to grammatically incorrect statements), and [1] argues that label smoothening does not allow for sparse distributions, and recommend using other regularizers in its place. We aim to study if that approach would improve model performance on Question-Answering Systems. A further expansion to this scope could be to compare this improvement using simple data augmentation techniques.

Approach We aim to use SQuAD 2.0 [3], a reading comprehension dataset, consisting of questions and answers derived from a set of Wikipedia articles. We aim to study the effect of various Entropy-based Regularizers (including label smoothening and confidence penalty) on pre-trained Language models for question answering. We aim to use a transformer based Language Model - BERT for this task. This model is based on bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers, which makes it the state of art in many NLP tasks such as Question-Answering (QA) [4]. BERT uses 2 unsupervised tasks for pre-training - Mask Language Model (MLM) which enables learning the context around a word and Next Sentence prediction (NSP) which trains a binary classifier to predict if one sentence is next to the other. This step improves accuracy on tasks such as QA which are dependent on the relationship between two sentences. We aim to use the code released with the original work of [1], which is built on top of **fairseq**[5], a sequence modeling toolkit. By using fairseq, we can train custom models for translation, summarization, language modeling and other text generation tasks. We follow the same framework as the original paper-using the fairseq criterion (plug-in). This plug-in provides a way to compute the loss given a model and a batch of data which will allows us to add entropy regularization to the loss calculation.

Evaluation Following the approach of [1], evaluation can be done using BLEU scores[6], a common scoring method for NLG tasks, generated using the SacreBLEU framework [7]. We aim to analyse the relationship between the entropy of the model and the corresponding BLEU scores. Further, an interesting experiment would be to see whether the different GERs are significant towards the derived BLEU scores.

References

[1] Clara Meister, Elizabeth Salesky, and Ryan Cotterell, 2020. *Generalized entropy regularization or: There's nothing special about label smoothing*

[2] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2015. *Rethinking the inception architecture for computer vision*

[3] Pranav Rajpurkar, Robin Jia, Percy Liang. 2018. *Know What You Don't Know: Unanswerable Questions for SQuAD*

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*

[5] *FAIRSEQ-A Fast Extensible Toolkit for Sequence Modeling,* `https://github.com/pytorch/fairseq`

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*

[7] Matt Post. 2018. *A call for clarity in reporting BLEU scores.*