Learned Data Augmentation and Non-autoregressive Translation

Justin Chiu
Cornell Tech
jtc257@cornell.edu

October 28, 2021

Abstract

Abstract

1 Introduction

Learned models of data are often misspecified. When the goal of modeling is not density estimation, but some alternative objective, this misspecification may lead to undesirable behaviour under the maximum likelihood objective. In this note, we consider learned data augmentation techniques to edit each data point so that the data as a whole is more amenable to learning for a particular model.

2 Problem Setup

Given data consisting of (x, y) pairs, our goal is to learn a model $q_\theta(y \mid x)$ such that it maximizes the functional

$$F(q) = \mathbb{E}_{p(x,y)} \left[\operatorname*{argmax}_{\hat{y}} d(q_\theta(\hat{y} \mid x), y) \right], \tag{1}$$

where $p(x, y)$ is the data distribution and d is some measure of correctness between our prediction $\hat{y}$ and the true output y.

A concrete example of this is translation, where x is a source sentence (for example, German), y is a target sentence (for example, English), and d is the BLEU score between our generated translation and the true reference target sentence. Our goal is to, given a family of student models $q_\theta(y \mid x)$, learn an edit model $q_\phi(\hat{y} \mid y, x)$ whose conditional distribution over $\hat{y}$ is easier for the student model q_θ to learn. Learning an intermediate distribution will allow the student model to focus on modeling the important aspects of the true distribution while ignoring others, the goal of minimizing the true objective in Equation 1.

To accomplish this, we propose to solve the following

$$\operatorname*{argmin}_{\phi} D_{\mathrm{KL}}\left[p(y \mid x) || q_{\phi}(\hat{y} \mid y, x)\right] + \min_{\theta} D_{\mathrm{KL}}\left[q_{\phi}(\hat{y} \mid y, x) || q_{\theta}(\hat{y} \mid x)\right]. \tag{2}$$

This is a bilevel optimization problem, where we want to find the edit model that is able to balance faithfulness to the true data (the first term) as well as learnability of the student model (the second term). We refer to Equation 2 as the outer problem, and the second term as the inner problem:

$$\operatorname*{argmin}_{\theta} D_{\mathrm{KL}}\left[q_{\phi}(\hat{y} \mid y, x) || q_{\theta}(\hat{y} \mid x)\right]. \tag{3}$$

3 Method 1: Exact Setting

We first make the simplifying assumption that we can solve the inner problem exactly. In order to solve the full outer problem, we will differentiate through the optima of the inner problem using the implicit function theorem.

References