

Learning to Augment: Non-Autoregressive Translation

Justin Chiu
Cornell Tech
jtc257@cornell.edu

November 8, 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 augmented data is more amenable to a particular model class, while hopefully preserving the properties we care about.

A concrete example of a setup where maximum likelihood may not be desirable is non-autoregressive translation, which we focus on in this note. In translation, the goal is to translate a source sentence from one language into a sentence in a different target language. The success of translation is measured by the BLEU score between our generated translation and the true reference target sentence. As the BLEU score is not a function of distributional closeness to the true data, translation systems often discard distributional information and output a single translation of a source sentence. However, the training objective is still often maximum likelihood, a function of the model distribution.

In the case of non-autoregressive translation, this results in undesirable behaviour. As non-autoregressive models cannot model output dependencies, resulting models trained via maximum likelihood using the true distribution often have pathological errors due to misspecification, and therefore poor BLEU scores. We aim to resolve this via data augmentation.

Prior work shows that distillation is a successful data augmentation approach. We instead rely on neural editor models [Guu et al., 2017], which are a trainable alternative to data augmentation [Akyürek et al., 2021].

2 Problem Setup

Source sentence x , target sentence y . True translation distribution $p(y \mid x)$.

28 We propose to, given a family of student models $q_\theta(y \mid x)$, learn an edit model $q_\phi(\hat{y} \mid y, x)$
 29 whose conditional distribution over \hat{y} is easier for the student model q_θ to learn. Learning an inter-
 30 mediate distribution will allow the student model to focus on only modeling the important aspects
 31 of the true distribution.

To accomplish this, we propose to solve the following optimization problem

$$\operatorname{argmin}_{\phi} D_{\text{KL}} \left[p(y \mid x) \parallel \sum_y q_\phi(\hat{y} \mid y, x) p(y \mid x) \right] + \min_{\theta} D_{\text{KL}} [q_\phi(\hat{y} \mid y, x) \parallel q_\theta(\hat{y} \mid x)]. \quad (1)$$

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 1 as the outer problem, and the second term as the inner problem:

$$\operatorname{argmin}_{\theta} D_{\text{KL}} [q_\phi(\hat{y} \mid y, x) \parallel q_\theta(\hat{y} \mid x)]. \quad (2)$$

The first term of the outer problem, the KL divergence between the true distribution and the marginal posterior of the edit distribution, serves as a regularizer and encourages the marginal distribution of the edit distribution combined with the data distribution to be close to the original data distribution. As this setting may be both too restrictive as well as not example-specific, we also consider the following variation:

$$\operatorname{argmin}_{\phi} \sum_y \Omega(y, q_\phi(\hat{y} \mid y, x)) + \min_{\theta} D_{\text{KL}} [q_\phi(\hat{y} \mid y, x) \parallel q_\theta(\hat{y} \mid x)], \quad (3)$$

32 allowing us to perform instance-level regularization rather than distributional.

33 **3 Method 1: The Exact Setting**

34 We first make the simplifying assumption that we can solve the inner problem exactly. In order to
 35 then solve the full outer problem, we will differentiate through the solution of the inner problem. In
 36 full generality, this would require an application of the implicit function theorem, or approximations
 37 through sampling or incomplete optimization.

38 **3.1 Experiment: Length-2 Binary Strings + Non-autoregressive Student**

39 Our first experiment will determine whether training an edit model is possible in a very simple
 40 setting, where we can compute the solution to the inner problem (training the student model given
 41 an edit model) in closed form, and subsequently differentiate through that procedure.

We restrict our attention to distributions over target sentences that are binary strings of length 2 without conditioning on a source sentence. In particular, our true distribution takes the form

$$y = b_1 \cdot 00 + b_2 \cdot 01 + b_3 \cdot 10 + b_4 \cdot 11,$$

where $b \sim \text{Cat}(\lambda)$, represented as a one-hot sample (meaning only one of b_i can be 1, while the rest must be 0). As we are interested in disambiguation, we first focus on the case where $b_1, b_4 \approx 0$, and $b_2 < b_3$. The edit model then takes the form

$$q_\phi(\hat{y} | y) = [\phi]_{\hat{y}, y},$$

where $\hat{y}, y \in \{0, 1\}^2 = \mathcal{Y}$ and $\phi \in [0, 1]^{\mathcal{Y} \times \mathcal{Y}}$. The student is nonautoregressive, and has the form

$$q_\theta(\hat{y}) = \prod_t q_\theta(\hat{y}_t) = [\theta]_{1, \hat{y}_1} \cdot [\theta]_{2, \hat{y}_2},$$

42 where $\theta \in [0, 1]^{2 \times \mathcal{Y}}$.

In this setting, we can exactly compute the solution of the inner problem,

$$\underset{\theta}{\operatorname{argmin}} \mathbb{E}_y [D_{\text{KL}} [q_\phi(\hat{y} | y) || q_\theta(\hat{y})]] = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_y \left[\sum_{\hat{y}} q_\phi(\hat{y} | y) (\log q_\phi(\hat{y} | y) - \log q_\theta(\hat{y})) \right],$$

and differentiate through it. The minimizer is given by

$$\begin{aligned} [\theta]_{1,0} &= \sum_y \sum_{\hat{y}} p(y) q_\phi(\hat{y} | y) 1(\hat{y}_1 = 0) \\ [\theta]_{2,0} &= \sum_y \sum_{\hat{y}} p(y) q_\phi(\hat{y} | y) 1(\hat{y}_2 = 0) \\ [\theta]_{t,1} &= 1 - [\theta]_{t,0}, \end{aligned}$$

43 which is differentiable wrt ϕ .

44 **Results** We find that the editor model is able to disambiguate the true data and produces a new
45 data distribution with no dependencies across time, resulting in a distribution that is easier for the
46 student to match. We see the results of training the editor model with the following conditions:

- 47 • $b = [0.001, 0.45, 0.54, 0.001]$
- 48 • Editor is initialized to uniform everywhere except the transitions from $[0,1]$ and $[1,0]$, where
49 transitions to $[0,0]$ and $[1,1]$ are given low mass in order to speed up training
- 50 • We recompute the parameters of the student by solving the inner problem at every iteration
- 51 • We solve the outer problem via gradient descent (1k iterations, learning rate of 1e-3)

52 We see in Figure 1 (right) that the editor learns to bias the data further towards $[1, 0]$. This results
53 in the student model being very certain, as seen in Figure 2 (right). In this very simple setting, the
54 bilevel optimization formulation successfully increases the asymmetry of the data.

55 However, in settings where the true data distribution is less skewed, the objective does not allow
56 the edit model to deviate too far. Figures 3 and 4 show the results when $b = [1e-3, 0.5, 0.5, 1-e3]$.
57 In this setting, both the data distribution and edit distribution are symmetric, resulting in the student
58 remaining at a standstill. This is because any movement away from the true distribution would incur
59 too large a penalty in the first term of the objective, $D_{\text{KL}} [p(y) || \sum_y q(\hat{y} | y)p(y)]$.

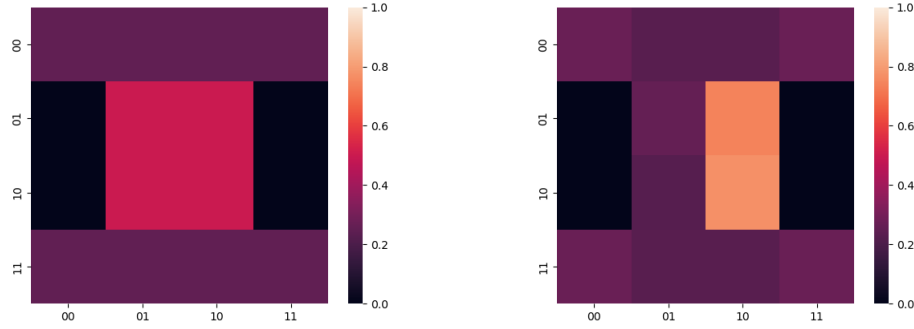


Figure 1: The emission probabilities for the editor model at the start of training (left) and end of training (right).

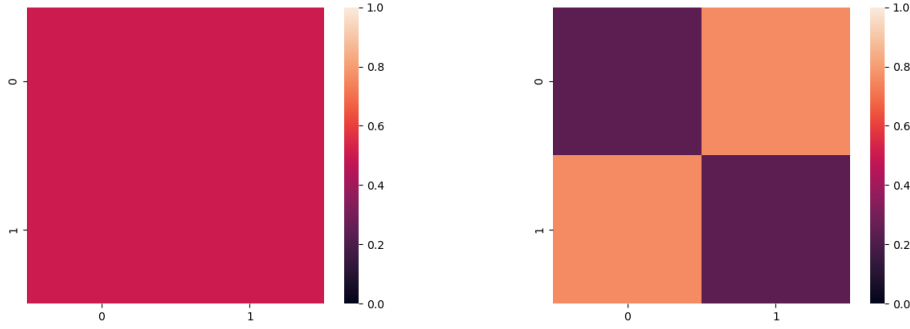


Figure 2: The emission probabilities for the student model at the start of training (left) and end of training (right).

60 Decoding: work back into problem setup later

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

$$F(q) = \mathbb{E}_{p(x,y)} \left[\operatorname{argmax}_{\hat{y}} d(q_\theta(\hat{y} \mid x), y) \right], \quad (4)$$

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

63 References

64 E. Akyürek, A. F. Akyürek, and J. Andreas. Learning to recombine and resample data for compo-
 65 sitional generalization. In *International Conference on Learning Representations*, 2021. URL

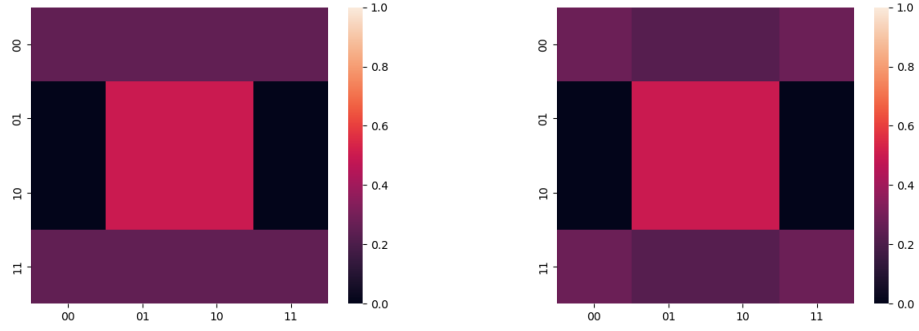


Figure 3: The emission probabilities for the editor model at the start of training (left) and end of training (right) with an even true data distribution.

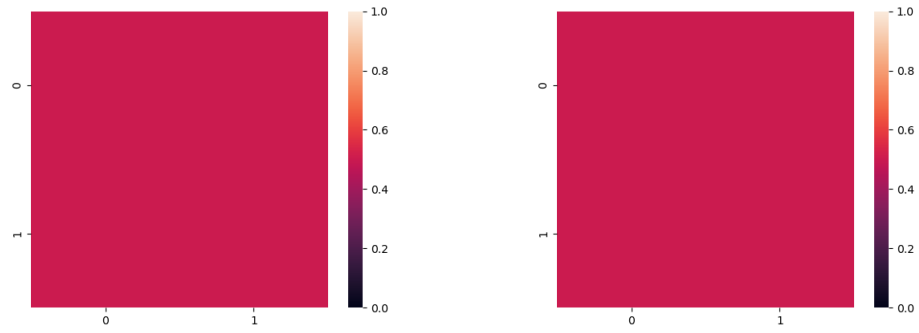


Figure 4: The emission probabilities for the student model at the start of training (left) and end of training (right) with an even true data distribution.

66 <https://openreview.net/forum?id=PS3IMnScugk>.

67 K. Guu, T. B. Hashimoto, Y. Oren, and P. Liang. Generating sentences by editing prototypes. *CoRR*,
 68 abs/1709.08878, 2017. URL <http://arxiv.org/abs/1709.08878>.