Supplementary Material: Variational Sequential Labelers for Semi-Supervised Learning

Mingda Chen **Qingming Tang** **Karen Livescu** **Kevin Gimpel**
Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA
{mchen,qmtang,klivescu,kgimpel}@ttic.edu

Algorithm 1 Prior Update Algorithm

Given the number of epochs E, the training data $\{x, l\}^i$, an update frequency f and loss function $L(x^j, l^j) = -\log p_\theta(x^j|z^j) + KL(q_\phi(z^j)\|p_j)$ where p_j means jth element in the list p and p_θ, q_ϕ are the generative model and inference model respectively
Initialize a count list C as zeros for all data
Initialize a prior list p for all data as standard Gaussian for Gaussian random variable
for t in range(E) **do**
  for Sample an instance x^j, l^j from the data **do**
    $\phi = \phi - \nabla_\phi L(x^j, l^j)$
    $C_j = C_j + 1$
    if $C_j > f$ **then**
      $p_j = q_\phi(z^j)$
      $C_j = 0$
    end if
  end for
end for

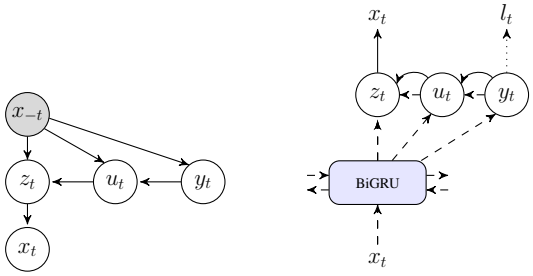

Figure 1: Variational sequential labeler with three latent variables (VSL-GGG-Hier). The left plot shows the graphical model and the right plot shows how we perform inference and learning, showing inference models (in dashed lines), generative models (in solid lines), and classification losses (in dotted lines). Training seeks to reconstruct x_t and predict the label l_t.

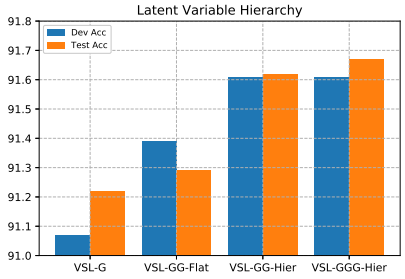

Figure 2: Latent structure vs. dev/test accuracies on Twitter dataset.

A Hyperparameter Tuning

The character embedding dimension is tuned over the set {10, 50, 100, 200}. The BiGRU hidden size is tuned over {100, 200, 300, 400, 500}. The latent variable dimension is tuned over {10, 50, 100, 200}. The trade-off hyper-parameter for the unlabeled loss is tuned over the range from 0.1 to 1 in increments of 0.1. The prior update frequency is tuned over {1,2,3,4,5,10}. The output variance is tuned over {0, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5}. We did not tune the KL annealing rate and the weight of the KL term, which are fixed at 1e-4 and 1 respectively. All the latent variables are parametrized by a single layer feedforward neural network. $p_\theta(x_t|\ z_t)$ is always a three layer feedforward neural network with hidden size of 100 and has ReLU as activation function.

B More Latent Variables

Figure 2 summarizes the differences in accuracy among models with different hierarchical structures on the Twitter dataset. There is a trend that a deeper latent hierarchy can help the latent variables to better capture the data distribution.