Adversarial Localized Energy Network for Structured Prediction

Pingbo Pan,[1,2*] **Ping Liu,**[2] **Yan Yan,**[3] **Tianbao Yang,**[3] **Yi Yang**[2]
[1]Baidu Research, [2]The ReLER Lab, University of Technology Sydney, [3]University of Iowa
pingbo.pan@student.uts.edu.au, pino.pingliu@gmail.com,
{yan-yan-2, tianbao-yang}@uiowa.edu, yi.yang@uts.edu.au

Abstract

This paper focuses on energy model based structured output prediction. Though inheriting the benefits from energy-based models to handle the sophisticated cases, previous deep energy-based methods suffered from the substantial computation cost introduced by the enormous amounts of gradient steps in the inference process. To boost the efficiency and accuracy of the energy-based models on structured output prediction, we propose a novel method analogous to the adversarial learning framework. Specifically, in our proposed framework, the generator consists of an inference network while the discriminator is comprised of an energy network. The two sub-modules, i.e., the inference network and the energy network, can benefit each other mutually during the whole computation process. On the one hand, our modified inference network can boost the efficiency by predicting good initializations and reducing the searching space for the inference process; On the other hand, inheriting the benefits of the energy network, the energy module in our network can evaluate the quality of the generated output from the inference network and correspondingly provides a resourceful guide to the training of the inference network. In the ideal case, the adversarial learning strategy makes sure the two sub-modules can achieve an equilibrium state after steps. We conduct extensive experiments to verify the effectiveness and efficiency of our proposed method.

Introduction

Structured output prediction, which aims to learn a mapping from an input $\mathbf{x}$ to a complex multivariate output structure $\mathbf{y}$, has been receiving significant attention due to its wide applications. For example, given an image, we predict a set of semantic labels with inter-relations for it, or output a semantic segmentation map for it. Not surprisingly, predicting structured output for a given input is challenging because input data, output structures, and their internal relations all live in a high-dimensional space. To learn the sophisticated relationships between the input and output, a prediction model with excellent expressivity capability as well as computational tractability needs to be learned.

To solve the aforementioned problem, LeCun (LeCun et al. 2006) suggest to define a prediction function that associates energy values to different configurations of output structures and name it as Energy-Based Method (EBM). Specifically, EBM suggests to solve the structured interrelation between input $\mathbf{x}$ and ground-truth labels $\mathbf{y}^*$ by learning parameters $\mathbf{w}$ for an energy function $v(.)$, i.e., $\mathbf{y}^* = argmax_{\mathbf{y}} v(\mathbf{x}, \mathbf{y})$. Comparing to feed-forward methods, the energy based method is able to achieve more accurate results since it can substantially handle complicated non-convex energies, which has been experimentally proved in previous work (Belanger, Yang, and McCallum 2017).

As indicated in the previous work (Belanger and McCallum 2016), the critical component in the learning of a deep energy-based model for structured prediction is *how to find* "the most offending incorrect answer" at each training step. The most offending answer $\tilde{\mathbf{y}}$ is defined as an answer with the highest energy but with an incorrect label, i.e., different from the ground truth $\mathbf{y}^*$:

$$\tilde{\mathbf{y}} = \arg\max_{\mathbf{y} \neq \mathbf{y}^*} v(\mathbf{x}, \mathbf{y}). \tag{1}$$

Some methods (Belanger and McCallum 2016; Gygli, Norouzi, and Angelova 2017) utilize gradient-based inference with zero initialization to find $\tilde{\mathbf{y}}$. The inference process convergences slowly, especially when the energy-based models are extremely complex, *e.g.*, deep neural networks, and $\mathbf{y}$ is high-dimensional. Other methods (Tu and Gimpel 2018) propose to learn an inference network to approximate $\tilde{\mathbf{y}}$. However, learning an optimal inference network at each training step is still computationally expensive.

To improve the process of finding the most offending $\tilde{\mathbf{y}}$, we propose a novel adversarial learning framework. Under the designed framework, the generator network is in the form of an inference network, while the discriminator network is in the form of an energy network. The inference network is learned to approximate $\tilde{\mathbf{y}}$ and updated for several (usually one) gradient steps at each training step. To eliminate the discrepancy between the inference network approximation $\mathbf{y}' = g(\mathbf{x})$ and the true $\tilde{\mathbf{y}}$, we further apply a gradient-based inference to refine $\mathbf{y}'$. The refined $\mathbf{y}'$ is used to train the energy network, i.e., the discriminator.

Thanks to the adversarial learning strategy, our frame-

*Part of this work was done when Pingbo Pan was a research intern with baidu research.

work can make two sub-modules leverage the mutual benefits to boost the final performance. In particular, the inference network is learned to predict outputs that get closer to the optimal outputs as the training continues. These outputs serve as the initialization points for the gradient-based inference, which speeds up both the training and inference of the energy network. On the other hand, the energy network will provide a resourceful guide to the training of the inference network by evaluating the quality of the generated output from the inference network.

We name the proposed method Adversarial Localized Energy Network (**ALEN**). Our approach inherits the advantages of EBM and can handle complicated energy spaces as EBM does. Compared to previous EBM based methods (Gygli, Norouzi, and Angelova 2017; Belanger, Yang, and McCallum 2017; Tu and Gimpel 2018), there are critical differences between the proposed framework and them: (1) (Gygli, Norouzi, and Angelova 2017) requires to sample massive training samples in the entire output space, while in our framework, the energy network leverages an inference network to provide training samples which largely reduces the training complexity; (2) our method takes an inference network as a generator and conducts the learning process in an adversarial learning paradigm, while (Belanger, Yang, and McCallum 2017) fails to achieve this; (3) We adopt a gradient-based inference to refine further the "inaccurate" approximations predicted by the inference network, while (Tu and Gimpel 2018) simply utilize the "inaccurate" approximations to train the energy network.

Although based on the adversarial learning framework, our framework has key differences comparing with the classical adversarial learning framework: (1) the generator (inference network) in our framework does not generate images but make a coarse structured output prediction; (2) the discriminator (energy network) in our framework does not treat the generator outputs as negative samples, instead, inheriting the benefits of EBM, it evaluates the quality of the outputs from the generator by calculating a real score in a soft way. Comparing to the hard "real/fake" prediction, the estimated quality score can provide more fine-grained information to facilitate the generator to provide predictions closer to the optimal configurations. In this way, the generator and discriminator in our framework can take advantage of more vigorous information to facilitate the training process.

We conduct experiments on different problems for the validation, including multi-label classification, binary image segmentation, and 3-class face segmentation tasks. The experimental results indicate that our proposed method can not only refine the final results to a higher stage even with a smaller input resolution, but also improve the convergence in the training and inference stages.

Preliminary

Energy-based models.

Different from feed-forward models, energy-based models (LeCun et al. 2006) take an input $\mathbf{x}$ and its corresponding label $\mathbf{y}$ as inputs, and predict an energy value. This energy value measures the compatibility of the label $\mathbf{y}$ corresponding to the input $\mathbf{x}$. A large score indicates the high coherence between $\mathbf{y}$ and the ground truth label $\mathbf{y}^*$.

To simplify the discussion without losing the generality, we begin with the generalized perceptron loss for energy-based models $v(\mathbf{x}, \mathbf{y}; \theta_v)$ proposed in (LeCun et al. 2006).

$$L(\theta_v) = \sum_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}} \left[\max_{\mathbf{y}\in\mathcal{Y}} v(\mathbf{x}, \mathbf{y}; \theta_v) - v(\mathbf{x}, \mathbf{y}^*; \theta_v) \right], \quad (2)$$

with the input $\mathbf{x}$ and ground-truth label $\mathbf{y}^*$ from training set $\mathcal{D}$.

When learning an energy-based model with this loss function, it is critical to find the most offending incorrect answers $\tilde{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} v(\mathbf{x}, \mathbf{y}; \theta_v)$ at each training step. Many existing methods (Belanger and McCallum 2016; Gygli, Norouzi, and Angelova 2017) adopt gradient-based inference with *zero initialization* to find $\tilde{\mathbf{y}}$. However, as indicated in previous works (Belanger and McCallum 2016), when energy-based models are deep neural networks, and $\mathbf{y}$ is high-dimensional, the inference process usually costs many steps to converges.

Inference network.

An alternative way of finding the most offending $\tilde{\mathbf{y}}$ is introducing an inference network $g(\mathbf{x}; \theta_g)$ to make an approximation (Tu and Gimpel 2018). After introducing the inference network $g(\mathbf{x}; \theta_g)$, the original objective function 2 is transformed into the following objective function:

$$\max_{\theta_v} \min_{\theta_g} \sum_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}} \left[v(\mathbf{x}, \mathbf{y}^*; \theta_v) - v(\mathbf{x}, g(\mathbf{x}; \theta_g); \theta_v) \right]. \quad (3)$$

The most frequently used strategy to solve this objective function is to alternatively learn $v(\mathbf{x}, \mathbf{y}; \theta_v)$ and $g(\mathbf{x}; \theta_g)$ at each training step. More specifically, fixing $g(\mathbf{x}; \theta_g)$ and optimize $v(\mathbf{x}, \mathbf{y}; \theta_v)$, and vice verse. In an ideal case, the training process stops when an equilibrium state is achieved.

Theoretically speaking, different energy-based models correspond to different inference networks for achieving the results with high accuracy. During the training process, the parameters of the energy-based models keep changing, which correspondingly needs thousands of gradient steps to find an optimal inference network to match. All of these make the whole computation process too expensive.

Adversarial localized energy network.

We argue that it is not necessary to achieve a "perfect" inference network at the cost of thousands of gradient steps for each training step. On the contrary, a "coarse" inference network has already been able to provide sufficient information for the further process. Based on this, we propose to optimize the inference network for only one gradient step at each training step. Since the inference network is adjusted for only one step, the computation cost is minimized to a great extent. To reduce the discrepancy between the ground truth and predicted answers $\mathbf{y}' = g(\mathbf{x}; \theta_g)$ generated by the "coarse" inference network, we apply a gradient-based inference to refine $\mathbf{y}'$:

$$\mathbf{y}^{(t+1)} = \mathcal{P}_{\mathcal{Y}} \left(\mathbf{y}^{(t)} + \eta \frac{\partial}{\partial \mathbf{y}} v(\mathbf{x}, \mathbf{y}^{(t)}; \theta_v) \right), \quad (4)$$

with $\mathbf{y}^{(0)} = \mathbf{y}'$. Here t denotes the t-th refinement step, and $\mathcal{P}_{\mathcal{Y}}$ denotes an operator that projects the predicted outputs back to the feasible set of solutions. Note that the inference process aims to refine $\mathbf{y}'$ rather than learning the energy network. In our observations, this refinement process converges to an answer approximated to the most offending answer, i.e., $\tilde{\mathbf{y}}$, in several refinement steps. We name this approximated answer $\mathbf{y}''$. Only a few refinement steps are needed in each training step. We propose to learn $v(\mathbf{x}, \mathbf{y}; \theta_v)$ with $\mathbf{y}''$ and rewrite Equation 3 as:

$$\max_{\theta_v} \min_{\theta_g} \sum_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}} [v(\mathbf{x}, \mathbf{y}^*; \theta_v) - v(\mathbf{x}, g(\mathbf{x}; \theta_g); \theta_v) - v(\mathbf{x}, \mathbf{y}''; \theta_v)]. \tag{5}$$

Note that the above equation is similar to the objective function of Wasserstein GAN (WGAN) (Arjovsky, Chintala, and Bottou 2017). The difference is that we aim to learn an energy-based model to capture the training data distribution, while WGAN seeks to learn a vanilla generative model to achieve the same goal.

Least square objective function. In the experiments, we notice that learning energy-based models with Equation 5 is unstable. Inspired by the success of Least Square GAN (LSGAN) (Mao et al. 2016), we propose a least-squares objective function:

$$L_{adv}(\theta_v) = \sum_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}} \left[\frac{1}{2} \left(v(\mathbf{x}, \mathbf{y}^*; \theta_v) - a \right)^2 \right. \tag{6a}$$

$$+ \frac{\lambda_v}{2} \left(v(\mathbf{x}, g(\mathbf{x}; \theta_g); \theta_v) - b \right)^2 \tag{6b}$$

$$\left. + \frac{\beta_v}{2} \left(v(\mathbf{x}, \mathbf{y}''; \theta_v) - c \right)^2 \right], \tag{6c}$$

$$L_{adv}(\theta_g) = \sum_{\mathbf{x}\sim\mathcal{D}} \left[\frac{1}{2} \left(v(\mathbf{x}, g(\mathbf{x}; \theta_g); \theta_v) - d \right)^2 \right], \tag{7}$$

with trade-off parameters λ_v and β_v. We set $a = d = 1$. We set b and c to the oracle value functions proposed in deep value network (DVN) (Gygli, Norouzi, and Angelova 2017): $b = h(g(\mathbf{x}; \theta_g), \mathbf{y}^*)$ and $c = h(\mathbf{y}'', \mathbf{y}^*)$. The formulation of function h will be illustrated in the following parts.

In our experiments, the oracle value function h can be F_1 metrics, which are defined on $(\mathbf{y}, \mathbf{y}^*) \in \{1, 0\}^M \times \{0, 1\}^M$,

$$h(\mathbf{y}, \mathbf{y}^*) = \frac{2(\mathbf{y} \cap \mathbf{y}^*)}{(\mathbf{y} \cap \mathbf{y}^*) + (\mathbf{y} \cup \mathbf{y}^*)}. \tag{8}$$

Here $\mathbf{y} \cap \mathbf{y}^*$ denotes the number of dimension i where both y_i and y_i^* are active and $\mathbf{y} \cup \mathbf{y}^*$ denotes the number of dimensions where at least one of y_i and y_i^* is active. y_i and y_i^* denote the i-th variable of $\mathbf{y}$ and $\mathbf{y}^*$. To apply $h(\mathbf{y}, \mathbf{y}^*)$ to the continuous output $\mathbf{y}$, the notions of intersection and union are extended by using element-wise min and max operators,

$$\mathbf{y} \cap \mathbf{y}^* = \sum_{i=1}^{M} \min(y_i, y_i^*), \tag{9}$$

Algorithm 1 ALEN training

Input: training data $\mathcal{C}$

1: **while** not converged **do**
2: $(\mathbf{x}, \mathbf{y}^*) \sim \mathcal{C}$
3: calculate $\mathbf{y}''$ according to Equation 4.
4: update θ_v according to Equation 6.
5: update θ_g according to Equation 11.
6: **end while**

Output:energy network weight θ_v and inference network weight θ_g

$$\mathbf{y} \cup \mathbf{y}^* = \sum_{i=1}^{M} \max(y_i, y_i^*). \tag{10}$$

Improvement training for inference networks with θ_g. We notice that adding a task-specific surrogate loss $L_{task}(\theta_g) = \text{loss}(g(\mathbf{x}; \theta_g), \mathbf{y}^*)$ can improve the training of inference networks with parameters θ_g. For example, the surrogate loss can be a cross-entropy loss in image segmentation tasks. This loss can make the inference network predicts answers, i.e., $\mathbf{y}'$, and its refined counterpart, i.e., $\mathbf{y}''$, more close to the ground-truth labels. The energy-based model is always learned to capture the data distribution of a local domain around the ground-truth labels. The training of the energy-based model can be accelerated. To this end, we translate Equation. 7 into a following formulation:

$$L(\theta_g) = L_{task}(\theta_g) \tag{11a}$$

$$+\lambda_g L_{adv}(\theta_g), \tag{11b}$$

with a trade-off parameter λ_g.

Optimization. We utilize the Adam optimizer (Kingma and Ba 2014) with the momentum term $\beta_1 = 0.5$ to train the inference network and the energy-based model. At each training iteration, we generate $\mathbf{y}''$ and update the parameters of the inference network and the energy-based model according to Equation. 11 and Equation. 6. The process is summarized in Algorithm 1.

Gradient-based Inference. After the energy-based model is learned, we also use a gradient-based inference to find a structural output with a high energy value. Different from (Gygli, Norouzi, and Angelova 2017), we use the output structure predicted by the inference network as the initialization of the gradient-based inference. However, it is observed in experiments that the learned EBM tends to give zero gradients during the gradient-based inference. One reason is that the predicted output structure of the inference network is already close to the optimal structure. In order to further improve the predicted output structure of the inference network by using the gradient-based inference and overcome the zero-gradient issue, we use a normalized gra-

dient method, $i.e.$,

$$\mathbf{y}^{(t+1)} = \mathcal{P}_{\mathcal{Y}}\left(\mathbf{y}^{(t)} + \eta \frac{\left(\frac{\partial}{\partial \mathbf{y}} v(\mathbf{x}, \mathbf{y}^{(t)}; \theta_v)\right)}{\left|\left|\frac{\partial}{\partial \mathbf{y}} v(\mathbf{x}, \mathbf{y}^{(t)}; \theta_v)\right|\right|}\right). \quad (12)$$

In the experiments, we find that the above equation is beneficial for both training and inference processes.

Discussion

The goal of our proposed learning framework is to improve the efficiency and accuracy of structured output prediction. We will explain why our framework is helpful based on previous studies of optimization and learning theory.

Provide better initialization From many optimization studies for both convex and non-convex optimization, it is well-known that the convergence of gradient-based methods depends on the quality of initialization (Mohri, Rostamizadeh, and Talwalkar 2012). A standard analysis of the gradient descent (GD) method for a smooth function shows that in order to find a stationary solution $\widehat{x}$ such that $\|\nabla f(\widehat{x})\|_2 \leq \epsilon$, the GD needs at most $\frac{2L(f(x_0)-f_*)}{\epsilon^2}$ iterations, where L is the smoothness constant of $f(x)$. It can be seen that the convergence speed of the GD depends linearly on the distance between the initial solution and the stationary point in terms of the function value. In our case, the proposed inference network predicts initializations close to the ground-truth labels for the inference process.

Reduce the size of data space The sample complexity of learning a hypothesis depends on the size of the data space. This is implied by many learning theories (Mohri, Rostamizadeh, and Talwalkar 2012). The proposed framework learns an energy network $v(\mathbf{x}, \mathbf{y}; \theta_v)$ via a least square loss. The framework can be understood from a regression setting with $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ as input feature and $h(\mathbf{y}, \mathbf{y}^*)$ as the target output. Consider a classical learning theory result for learning a linear model $\phi(\cdot) \in \mathcal{H} = \{\phi : \mathbf{z} \to \mathbf{w}^\top \mathbf{z} : \|\mathbf{w}\| \leq B\}$. In particular, the excess risk of empirical risk minimizer is given by

$$\begin{aligned} &\mathbb{E}_{\mathbf{z}}[\ell(\hat{\phi}(\mathbf{z}), v)] - \min_{\phi \in \mathcal{H}} \mathbb{E}_{\mathbf{z}}[\ell(\phi(\mathbf{z}), v)] \\ &\qquad \leq C_1 G \mathcal{R}_n(\mathcal{H}) + C_2 \sqrt{\frac{1}{n}}, \end{aligned} \quad (13)$$

where C_1 and C_2 are some constants or parameters that depend on desired confidence score, G is the Lipchitz constant of the loss function $\ell(\cdot, \cdot)$ with respect to the first variable, and $\mathcal{R}_n(\mathcal{H})$ is the Rademacher complexity of a function class $\mathcal{H}$. For a linear model class, $\mathcal{R}_n(\mathcal{H}) \propto RB/\sqrt{n}$, where $R = \max_{\mathbf{z}} \|\mathbf{z}\|$ is the measure of the size of the data space. Although the above reasoning is not exact for learning the energy network for structured output prediction, it suffices to illustrate our point, $i.e.$, the larger the input data space, the more samples are needed for learning an accurate model. The proposed framework reduces the input space of the energy network to a local region around the ground-truth labels.

Related Work

Generative Adversarial Network. Our ALEN is similar to the Generative Adversarial Network (GAN) framework to some extent. The GAN is a framework for training generative models, and its ability to generate high-quality images has been shown in (Goodfellow et al. 2014; Radford, Metz, and Chintala 2015; Denton et al. 2015; Isola et al. 2016; Park et al. 2018; Song et al. 2018). The GAN framework consists of a generator network G and a discriminator network D. G is trained to capture the data distribution, while D is trained to distinguish samples generated by G from the training data. The difference between GAN and our framework is that our energy network is not to differentiate the input-output pairs generated by the inference network from the real input-output pairs. Instead, the generated pairs are used to accelerate the training of the energy network.

Deep structured prediction. Recent years have witnessed the rising interest in applying neural networks to the structured prediction (Zheng et al. 2015; Chen et al. 2015; Song et al. 2016). (Amos, Xu, and Kolter 2016) proposed to add constraints to the neural network parameters such that the output of the neural network is a convex function of (some of) the inputs. Their constraints may impose a strong restriction on the expressivity of the neural network.

Partly inspired by the energy-based learning framework proposed by (LeCun et al. 2006), (Belanger and McCallum 2016) introduced Structured Prediction Energy Network (SPEN). SPEN relies on a max-margin surrogate objective to ensure that the neural network predicts the lowest energy value for the ground-truth label. (Belanger, Yang, and McCallum 2017) improved SPEN by proposing an end-to-end version of SPEN, which directly back-propagates through a computation graph that unrolls gradient-based energy minimization.

Inspired by reinforcement learning, (Gygli, Norouzi, and Angelova 2017) proposed a novel energy-based network to critique different output configurations directly. The key to train their network is to generate proper samples that cover the space of the output, but it becomes hard when the output has an enormous number of variables. We solve this problem by introducing an inference network, which is treated as a generator, to collaborate with the energy network. The introduced inference network aims to provide appropriate samples to the energy module, and gradually assist the energy network in covering a local neighborhood around the ground-truth label for each input.

Experiments

To make a fair comparison, we follow the same protocol and tasks like (Gygli, Norouzi, and Angelova 2017), in which three different tasks are tested, including multi-label classification, binary image segmentation, and 3-class face segmentation. Firstly we compare the accuracies from our framework with previous works, and then we analyze the convergence speed between ALEN and DVN, which is the closest to our work. Comparing to DVN, ALEN learns an inference network to provide better initializations, which reduces the searching space and speeds up the inference pro-

cess. Based on the comparison results, we justify that our proposed framework not only achieves better prediction results but also converges faster than (Gygli, Norouzi, and Angelova 2017). Our implementation is based on Tensorflow (Abadi et al. 2016). We also implement the proposed method with PaddlePaddle and achieve similar performance. We find the best hyperparameters for all the algorithms via grid search.

Accuracy Improvement

Multi-label Classification We use standard benchmarks of this task, namely Bibtex and Bookmarks, introduced by (Katakis, Tsoumakas, and Vlahavas 2008). The learned models are evaluated on the testing set of those two datasets to report the F_1 scores. To make a fair comparison with (Gygli, Norouzi, and Angelova 2017), we choose the same energy network architecture, the same optimizer (Adam) with individually-tuned learning rates. A two-layer neural network (Belanger and McCallum 2016) is utilized as our inference network.

We compare the prediction performance of the proposed framework and standard baselines including logistic regression (Lin et al. 2014), a two-layer neural network with a cross-entropy loss (Belanger and McCallum 2016), SPEN (Belanger and McCallum 2016), PRLR (Lin et al. 2014), and DVN (Gygli, Norouzi, and Angelova 2017) on multi-label classification in Table 1. As illustrated in the table, although utilizing the same architecture, the ALEN still outperforms the DVN on both Bibtex and Bookmarks datasets. More than that, ALEN outperforms the SPEN by a large margin (4.2% on Bibtex, 3.9% on Bookmarks). Not surprisingly, our framework significantly improves over feed-forward models: the logistic regression, the two-layer neural network, and the PRLP, which are lacking the ability to learn complex correlations among variables in the output structures.

We implemented an adversarial energy network (AEN) which learns the energy network only with "inaccurate" inference network approximations. In other words, AEN learns the energy network only with the first and the second terms of Equation 6. Our method outperforms AEN and SPEN(InfNet) (Tu and Gimpel 2018) on both datasets. It shows that applying a gradient-based inference to refine "inaccurate" inference network approximation, as indicated by Equation. 4, helps to improve the energy network performance.

We make comparisons with the other three methods as ablation studies. One is named "InfNet baseline", which is implemented by training an inference network by minimizing a task-specific surrogate loss, $i.e.$, the first term of Equation 11. Another one is "ALEN no L_{task}" where ALEN is learned without the task-specific surrogate loss. The last one is referred to as "InfNet (ALEN)" representing the results predicted by the inference network of the ALEN. In this setting, the inference network and the energy-based model are trained adversarially, but in the inference stage, only the inference network is utilized to make the structured prediction. As shown in Table 1, ALEN outperforms all of the three baseline methods significantly, which demonstrates the superiority of the adversarial learning strategy (ALEN vs. InfNet baseline), task-specific surrogate loss (ALEN vs. ALEN no L_{task}), and the energy network utilization in inference stages (ALEN vs. InfNet(ALEN)).

Method	Eq 6 a	Eq 6 b	Eq 6 c	Eq 11 a	Eq 11 b	Bibtex	Bookmarks
Logistic Regression						37.2	30.7
Two layer Neural Network						38.9	33.8
SPEN						42.2	34.4
SPEN(InfNet)						42.2	37.6
PRLR						44.2	34.9
DVN						44.7	37.1
InfNet baseline				✓		38.9	32.8
AEN	✓	✓		✓	✓	44.2	34.6
ALEN no L_{task}	✓	✓	✓		✓	45.3	37.1
InfNet (ALEN)	✓	✓	✓	✓	✓	42.8	37.2
ALEN	✓	✓	✓	✓	✓	**46.4**	**38.3**

Table 1: The comparison of F_1 scores between ALEN and other state-of-the-art methods on Bibtex and Bookmarks datasets. The "InfNet baseline" is implemented by training an inference network by minimizing a task-specifc surrogate loss, *i.e.*, the first term of Equation 11. "AEN" is a baseline model that learns the energy network only with "inaccurate" inference network approximation. "ALEN no L_{task}" denotes that ALEN is learned without the task-specific surrogate loss. The "InfNet (ALEN)" represents the results predicted by the inference network of the ALEN.

3-class Face Segmentation We utilize the Labeled Faces in the Wild (LFW) dataset (Huang et al. 2007) to evaluate our framework on 3-class face segmentation. This dataset contains more than $13,000$ images, in which $2,927$ images are annotated for face segmentation. The annotations provide superpixel-level labels, which consist of three classes: face, hair, and background. Since our method generates pixel-level labels, we map pixel-level labels to superpixel-level labels by using the most frequent labels in a superpixel as the superpixel's label following (Tsogkas et al. 2015; Gygli, Norouzi, and Angelova 2017). We follow the same training, validation, and testing splits proposed in (Kae et al. 2013; Tsogkas et al. 2015; Gygli, Norouzi, and Angelova 2017) and utilize the same network architecture and data augmentation strategy as them.

We report the comparison results in Table 2. Thanks to the adversarial learning strategy in our method, when given the same input size of 32×32, our framework outperforms DVN (Gygli, Norouzi, and Angelova 2017) by a large marge (4.03%). More than that, given low-resolution input, the performance of our method is still comparable or better than the previous state-of-the-art methods, which need high-resolution input (250×250). The qualitative results of our approach are shown on the LFW dataset in Figure 1. As shown in Figure 1, our method can generate high-quality hair and face segmentation masks that are close to the ground-truth labels except for the mustache, which is tiny in low-resolution images and therefore quite hard to predict.

Input Size		Method	SP Acc. %
	32×32	DVN	92.44
		FCN baseline	95.36
		InfNet (ALEN)	95.87
		ALEN	**96.47**
	250×250	CRF	93.23
		GLOC	93.23
		DNN	96.54
		DNN+CRF+SBM	**96.97**

Table 2: The comparison of superpixel accuracy (SP Acc) between our framework and other state-of-the-art methods on the LFW dataset. "InfNet (ALEN)" represents the results predicted by the inference network of our ALEN.

Binary Image Segmentation Following the work of DVN (Gygli, Norouzi, and Angelova 2017), we utilize the Weizmann horses dataset (Borenstein and Ullman 2004) to compare the performance on binary image segmentation. It is a commonly used dataset for binary image segmentation, which consists of 328 left-oriented horse images and corresponding binary segmentation masks. In (Gygli, Norouzi, and Angelova 2017; Li, Tarlow, and Zemel 2013), all images and segmentation masks are resized to 32×32. At this low resolution, the segmentation becomes challenging and requires models to capture strong priors of the horse shape, since some thin parts of the horse-like legs, tails are almost invisible in the images. We follow the experimental protocol of (Li, Tarlow, and Zemel 2013) to split the Weizmann horses dataset and report results on the same testing set.

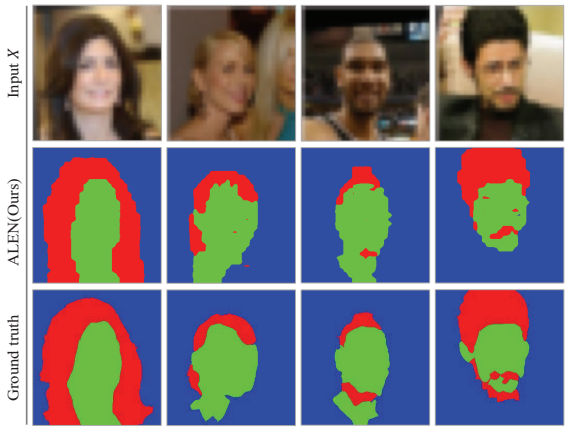


Figure 1: Qualitative results on the LFW dataset. Our method can generate high-quality hair and face segmentation masks that are close to the ground-truth labels. For the mustache, it is tiny on low-resolution images and therefore it is hard to predict.

To achieve better performance on this dataset, we implement the energy network as a fully convolutional network (FCN) rather than a classifier. Our energy networks map $(\mathbf{x}, \mathbf{y})$ to score matrices. We view image segmentation as pixel-level multi-label classification. Our inference network is also implemented as an FCN. Both FCNs consist of three 5×5 convolutional layers and two deconvolution layers. We follow the same DVN optimization schedule in (Gygli, Norouzi, and Angelova 2017) to train the energy network in our framework, which is used as a baseline and referred as to the "DVN baseline".

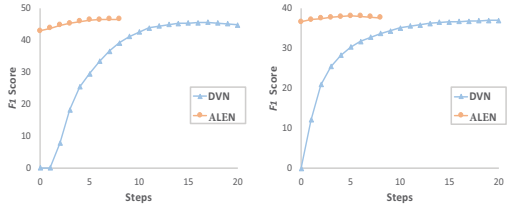
As commonly done in previous works, we report the Mean IOU as well as the Global IOU over the whole testing set on the Weizmann horses dataset in Table 3. A higher IOU score means a more accurate segmentation result. As illustrated in Table 3, using an energy network to refine the predictions of an inference network improves the performance by 4.2% on the Mean IOU metric and 4.1% on the Global IOU metric (ALEN vs. InfNet(ALEN)). On both metrics, our framework outperforms previous state-of-the-art methods, including MMBM2 (Yang, Safar, and Yang 2014), MMBM2+GC (Yang, Safar, and Yang 2014) and Shape NN (Safar and Yang 2015), which take high-resolution images as their input or utilize a deeper segmentation network((Safar and Yang 2015)). It can be observed that our adversarial learning framework can greatly improve the performance of the inference network. It improves the Mean IOU from 78.56% to 81.3% and the Global IOU from 78.7% to 81.3%.

Input size	Method	Mean IOU %	Global IOU %
32^2	CHOPPS	69.9	-
	DVN	84.1	84.0
	FCN baseline	78.56	78.7
	DVN baseline	84.7	84.3
	InfNet (ALEN)	81.3	81.3
	ALEN	**85.5**	**85.4**
128^2	MMBM2	-	72.1
	MMBM2 + GC	-	75.8
	Shape NN	-	**83.5**

Table 3: The comparison of IOU between our ALEN and other state-of-the-art methods on the Weizmann horses dataset. "DVN baseline" is implemented by using the framework of (Gygli, Norouzi, and Angelova 2017) to train our energy network. "InfNet (ALEN)" represents the results predicted by the inference network of our ALEN.

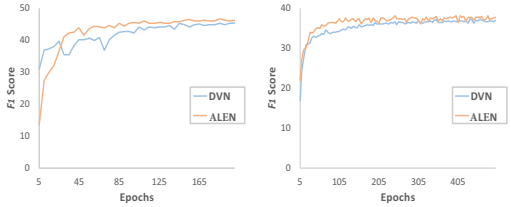
The qualitative results on the Weizmann horses dataset are shown in Figure 2. Without using the energy network in the inference stage, only using the inference network for segmentation (the third row) shows poor performances when segmenting thin parts like legs, and generates single-connected segmentation masks. After introducing the energy network via adversarial learning framework, our proposed method refines the inference network output by filling the missing part ($e.g.$, Figure 2, second and third row, far left images), generating legs to connect disconnected parts ($e.g.$, Figure 2, second and third row, first and second images from the right).

Figure 2: Qualitative results on the Weizmann 32×32 dataset.



(a) Bibtex (b) Bookmarks

Figure 3: The comparison of inference between the DVN and our ALEN.



(a) Bibtex (b) Bookmarks

Figure 4: The comparison of the training speed between the DVN and our ALEN.

Convergence Comparison on Multi-label Classification

To show that our framework can accelerate the training and inference of energy networks, we compare the proposed framework and DVN on the multi-label classification task. F_1 scores calculated at each inference step are reported in Figure 3. The curves show that the inference network of the ALEN estimates a good output initialization, and the ALEN converges faster than the DVN on both datasets. The gradient ascent optimizer finds the optimal output of the ALEN within 6 steps, while it takes 18 steps for the DVN. These results provide the empirical evidence that the inference network in our framework provides close-to-optimal initialization and therefore accelerates the convergence of the gradient-based inference.

We also compare the training process between the proposed framework and the DVN in Figure 4. The models are evaluated on the testing set every 5 epochs during the training process. From the figure we can find that: (1) our method outperforms the DVN at most times; (2) compared with the DVN, our method tends to converge faster. These results indicate that learning an energy network in a local neighborhood of the optimal output configurations promotes the performance of the energy network and accelerates the training.

Conclusion

This paper proposes a new adversarial learning framework to solve the structured output prediction. Comparing with the previous work (Gygli, Norouzi, and Angelova 2017), our proposed method can not only improve the final performance but also speed up the training and inference process. In this framework, an inference network is learned to generate outputs close to the optimal output configurations. An energy network can be learned faster, and the inference can be accelerated by using the outputs of the inference network. We jointly train the inference network as well as the energy network in a systematic way so that they can leverage the mutual benefit. Our method is applied to multi-label classification and image segmentation. The experimental results indicate that the training and inference of our ALEN are faster than the DVN, and the ALEN outperforms the DVN and achieves state-of-the-art results on these tasks.

As the future work, we will explore different ways to generate training tuples and different functions $l_g(*,*)$ for the GAN loss. We will apply our method to state-of-the-art deep neural networks to solve challenging realistic problems.

References

Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*.

Amos, B.; Xu, L.; and Kolter, J. Z. 2016. Input convex neural networks. *arXiv:1609.07152*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Belanger, D., and McCallum, A. 2016. Structured prediction energy networks. In *ICML*, 983–992.

Belanger, D.; Yang, B.; and McCallum, A. 2017. End-to-end learning for structured prediction energy networks. *arXiv:1703.05667*.

Borenstein, E., and Ullman, S. 2004. Learning to segment. *ECCV* 315–328.

Chen, L.-C.; Schwing, A.; Yuille, A.; and Urtasun, R. 2015. Learning deep structured models. In *ICML*, 1785–1794.

Denton, E. L.; Chintala, S.; Fergus, R.; et al. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, 1486–1494.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

Gygli, M.; Norouzi, M.; and Angelova, A. 2017. Deep value networks learn to evaluate and iteratively refine structured outputs. In *ICML*.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2016. Image-to-image translation with conditional adversarial networks. *arXiv:1611.07004*.

Kae, A.; Sohn, K.; Lee, H.; and Learned-Miller, E. 2013. Augmenting crfs with boltzmann machine shape priors for image labeling. In *CVPR*, 2019–2026.

Katakis, I.; Tsoumakas, G.; and Vlahavas, I. 2008. Multilabel text classification for automated tag suggestion. *ECML PKDD discovery challenge* 75.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. A tutorial on energy-based learning. *Predicting structured data* 1:0.

Li, Y.; Tarlow, D.; and Zemel, R. 2013. Exploring compositional high order pattern potentials for structured output learning. In *CVPR*, 49–56.

Lin, X. V.; Singh, S.; He, L.; Taskar, B.; and Zettlemoyer, L. 2014. Multi-label learning with posterior regularization. *NIPS Workshop on Modern Machine Learning and Natural Language Processing*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y. K.; and Wang, Z. 2016. Multi-class generative adversarial networks with the L2 loss function. *arXiv:1611.04076*.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of Machine Learning*. The MIT Press.

Park, D. K.; Yoo, S.; Bahng, H.; Choo, J.; and Park, N. 2018. Megan: mixture of experts of generative adversarial networks for multimodal image generation. *IJCAI*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*.

Safar, S., and Yang, M.-H. 2015. Learning shape priors for object segmentation via neural networks. In *ICIP*, 1835–1839. IEEE.

Song, Y.; Schwing, A.; Urtasun, R.; et al. 2016. Training deep neural networks via direct loss minimization. In *ICML*, 2169–2177.

Song, J.; Zhang, J.; Gao, L.; Liu, X.; and Shen, H. T. 2018. Dual conditional gans for face aging and rejuvenation. In *IJCAI*, 899–905.

Tsogkas, S.; Kokkinos, I.; Papandreou, G.; and Vedaldi, A. 2015. Deep learning for semantic part segmentation with high-level guidance. *arXiv:1505.02438*.

Tu, L., and Gimpel, K. 2018. Learning approximate inference networks for structured prediction. *ICLR*.

Yang, J.; Safar, S.; and Yang, M.-H. 2014. Max-margin boltzmann machines for object segmentation. In *CVPR*, 320–327.

Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; and Torr, P. H. 2015. Conditional random fields as recurrent neural networks. In *ICCV*, 1529–1537.