Learning Cycle-Consistent Cooperative Networks via Alternating MCMC Teaching for Unsupervised Cross-Domain Translation

Jianwen Xie[1,*], Zilong Zheng[2,*], Xiaolin Fang[3], Song-Chun Zhu[2,4,5], Ying Nian Wu[2] (*equal contribution)

[1] Baidu Research, [2] University of California, Los Angeles, [3] Massachusetts Institute of Technology, [4] Peking University, [5] Tsinghua University

Problem definition and modeling

(1) Problem

Suppose we have two different data domains, say $\mathcal{X}$ and $\mathcal{Y}$, and data collections from these two different domains $\{x_i, i = 1, ..., n_x\}$ and $\{y_i, i = 1, ..., n_y\}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Let $p_{\text{data}}(x)$ and $p_{\text{data}}(y)$ denote the unknown data distributions for these two domains. Without instance-level correspondence between two collections, we want to learn translators between two domains for cross-domain translation.

(1) Latent variable model as a translator

We first talk about the model for one-way translation, e.g., $\mathcal{Y} \to \mathcal{X}$. We specify a mapping $G_{\mathcal{Y}\to\mathcal{X}}$ that seeks to re-express the image y in domain $\mathcal{Y}$ by the image x in domain $\mathcal{X}$. The representational model is of the following form:

$$y \sim p_{data}(y),$$
$$x = G_{\mathcal{Y}\to\mathcal{X}}(y; \alpha_{\mathcal{X}}) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_D),$$

where $G_{\mathcal{Y}\to\mathcal{X}}(y; \alpha_{\mathcal{X}})$ is an encoder-decoder network whose parameters are $\alpha_{\mathcal{X}}$, and ϵ is a Gaussian residual. We assume σ is given and I_D is the D-dimensional identity matrix. We have conditional distribution : $q(x|y) \sim \mathcal{N}(G_{\mathcal{Y}\to\mathcal{X}}(y; \alpha_{\mathcal{X}}), \sigma^2 I_D)$ and marginal density : $q(x; \alpha_{\mathcal{X}}) = \int p_{\text{data}}(y) q(x|y; \alpha_{\mathcal{X}}) dy$. Learning $\alpha_{\mathcal{X}}$ via maximum likelihood is hard, because it requires the prior $p_{data}(y)$ to be a tractable density.

(2) Energy-based model as a teacher

To avoid the challenge of inferring y from x, we train $G_{\mathcal{Y}\to\mathcal{X}}$ via MCMC teaching by recruiting an energy-based model (EBM) as a teacher. The EBM specifies the distribution of x by

$$p(x; \theta_{\mathcal{X}}) = \frac{1}{Z(\theta_{\mathcal{X}})} \exp\left[f(x; \theta_{\mathcal{X}})\right] p_0(x),$$

where $p_0(x) \propto \exp(-||x||^2/2s^2)$ is the Gaussian reference distribution. The energy function $\mathcal{E}(x; \theta_{\mathcal{X}}) = -f(x; \theta_{\mathcal{X}}) + ||x||^2/s^2$, where f is parameterized by a ConvNet with parameters $\theta_{\mathcal{X}}$. $Z(\theta_{\mathcal{X}}) = \int \exp\left[f(x; \theta_{\mathcal{X}})\right] p_0(x) dx$ is the intractable normalizing constant. In practice, we can learn $\theta_{\mathcal{X}}$ by maximum likelihood estimation with the gradient that can be approximated by

$$\Delta(\theta_{\mathcal{X}}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_{\mathcal{X}}} f(x_i; \theta_{\mathcal{X}}) - \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \frac{\partial}{\partial \theta_{\mathcal{X}}} f(\tilde{x}_i; \theta_{\mathcal{X}}),$$

where $\tilde{x}_i$ are MCMC examples sampled from $p(x; \theta_{\mathcal{X}})$. With an EBM model defined on domain $\mathcal{X}$, we can sample x by MCMC, such as Langevin dynamics, which iterates

$$x_{\tau+1} = x_\tau - \frac{\delta^2}{2} \frac{\partial}{\partial x} \mathcal{E}(x_\tau; \theta_{\mathcal{X}}) + \delta U_\tau,$$

where τ indexes the time step, δ is the step size, and $U_\tau \sim \mathcal{N}(0, I_D)$.

(3) MCMC teaching

(i) Let $\mathcal{M}_{\theta_{\mathcal{X}}}$ be the Markov transition kernel of l steps of Langevin dynamics that samples from $p(x; \theta_{\mathcal{X}})$, and $\mathcal{M}_{\theta_{\mathcal{X}}} q_{\alpha_{\mathcal{X}}}$ be the marginal distribution obtained by running the Markov transition $\mathcal{M}_{\theta_{\mathcal{X}}}$ from distribution $q(x, \alpha_{\mathcal{X}})$. That is we use the translator $q(x, \alpha_{\mathcal{X}})$ to initialize the Langevin dynamics of p.

(ii) The EBM $p(x; \theta_{\mathcal{X}})$ can distill its MCMC algorithm to $q(x; \alpha_{\mathcal{X}})$ through MCMC teaching, which seeks to find α at time t to minimize $\text{KL}(\mathcal{M}_{\theta_{\mathcal{X}}} q_{\alpha_{\mathcal{X}}^{(t)}} | q_{\alpha_{\mathcal{X}}})$. That is, $q(x; \alpha_{\mathcal{X}})$ gets close to $p(x; \theta_{\mathcal{X}})$.

(4) The proposed models

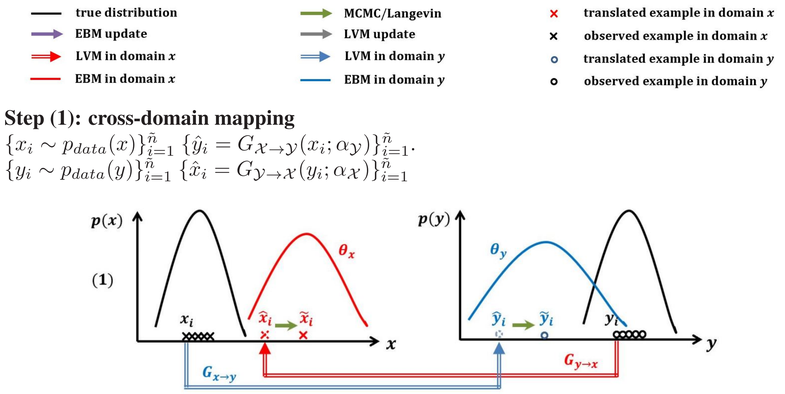
We propose the Cycle-consistent cooperative network (`CycleCoopNets`), to simultaneously learn and align two translator-teacher pairs, i.e.,

$$\mathcal{Y} \to \mathcal{X} : \{p(x; \theta_{\mathcal{X}}), G_{\mathcal{Y}\to\mathcal{X}}(y; \alpha_{\mathcal{X}})\},$$
$$\mathcal{X} \to \mathcal{Y} : \{p(y; \theta_{\mathcal{Y}}), G_{\mathcal{X}\to\mathcal{Y}}(x; \alpha_{\mathcal{Y}})\},$$

where each pair of models is trained via MCMC teaching to form a one-way translation. We align them by enforcing mutual invertibility, i.e.,

$$x_i = G_{\mathcal{Y}\to\mathcal{X}}(G_{\mathcal{X}\to\mathcal{Y}}(x_i; \alpha_{\mathcal{Y}}); \alpha_{\mathcal{X}}),$$
$$y_i = G_{\mathcal{X}\to\mathcal{Y}}(G_{\mathcal{Y}\to\mathcal{X}}(y_i; \alpha_{\mathcal{X}}); \alpha_{\mathcal{Y}}).$$

Learning



Experiment 1: Object transfiguration and season transfer

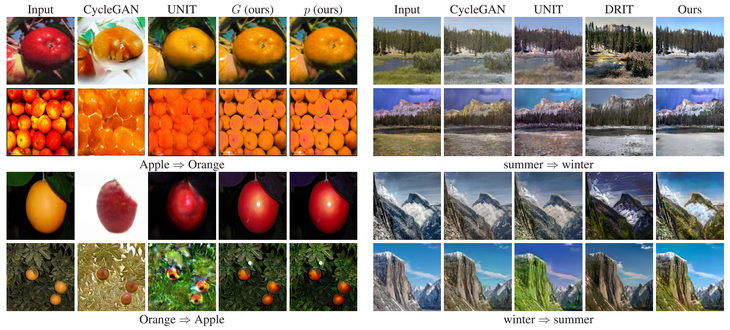


Figure 1: (Left) Object transfiguration. (Right) Season transfer.

Experiment 2: Translation between photo image and semantic label image

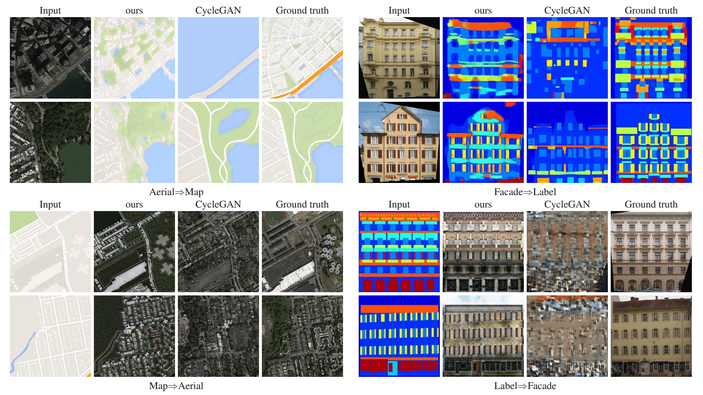


Figure 2: Translation between photo image and semantic label image

Table 1: Quantitative evaluation of unsupervised image-to-image translation by PSNR.

Dataset	Model	↑ L⇒R	↑ R⇒L
Aerial⇔Map	CycleGAN	21.59	12.67
	AlignFlow(mle)	19.47	13.60
	AlignFlow(adv)	20.16	**15.17**
	Ours	**22.29**	14.50
Facade⇔Label	CycleGAN	6.68	7.61
	AlignFlow(mle)	6.47	8.26
	AlignFlow(adv)	7.74	11.74
	Ours	**9.34**	**11.93**

Experiment 3: Style transfer

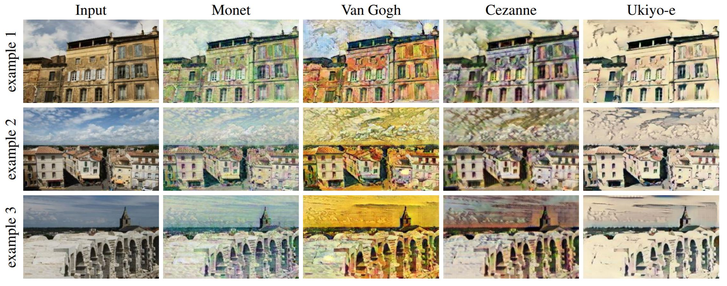


Figure 3: Collection style transfer from photo realistic images to artistic styles

Experiment 4: Image sequence translation

Figure 4: (a) We translate Barack Obama's facial motion to Donald Trump. (b) We translate from the blooming of a violet flower to a yellow flower. (c) We translate the blooming of a purple flower to a red flower.