On Unifying Deep Generative Models: Supplementary Materials

1 Adversarial Domain Adaptation (ADA)

ADA aims to transfer prediction knowledge learned from a source domain with labeled data to a target domain without labels, by learning domain-invariant features. Let $D_\phi(\boldsymbol{x}) = q_\phi(y|\boldsymbol{x})$ be the domain discriminator. The conventional formulation of ADA is as following:

$$
\begin{aligned}
&\max\nolimits_D \mathcal{L}_D = \mathbb{E}_{\boldsymbol{x}=G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z}|y=1)}\left[\log D(\boldsymbol{x})\right] + \mathbb{E}_{\boldsymbol{x}=G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z}|y=0)}\left[\log(1-D(\boldsymbol{x}))\right], \\
&\max\nolimits_G \mathcal{L}_G = \mathbb{E}_{\boldsymbol{x}=G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z}|y=1)}\left[\log(1-D(\boldsymbol{x}))\right] + \mathbb{E}_{\boldsymbol{x}=G(\boldsymbol{z}),\boldsymbol{z}\sim p(\boldsymbol{z}|y=0)}\left[\log D(\boldsymbol{x})\right].
\end{aligned} \tag{1}
$$

Further add the supervision objective of predicting label t in the source domain with a classifier $u(t|\boldsymbol{x})$:

$$
\max\nolimits_{u,G} \mathcal{L}_{u,G} = \mathbb{E}_{(\boldsymbol{z},t)}\left[\log u(t|G(\boldsymbol{x}))\right]. \tag{2}
$$

We then obtain the conventional formulation of adversarial domain adaptation used or similar in [3, 4, 5, 2].

2 Lemma 1

Proof.

$$
\begin{aligned}
&\mathbb{E}_{p(y)}\left[\mathbb{E}_{p_\theta(\boldsymbol{x}|y)}\left[\log q^r(y|\boldsymbol{x})\right]\right] = \\
&-\mathbb{E}_{p(y)}\left[\mathrm{KL}\left(p_\theta(\boldsymbol{x}|y)\|q^r(\boldsymbol{x}|y)\right) - \mathrm{KL}(p_\theta(\boldsymbol{x}|y)\|p_{\theta_0}(\boldsymbol{x}))\right],
\end{aligned} \tag{3}
$$

where

$$
\begin{aligned}
&\mathbb{E}_{p(y)}\left[\mathrm{KL}(p_\theta(\boldsymbol{x}|y)\|p_{\theta_0}(\boldsymbol{x}))\right] = \\
&\quad p(y=0)\cdot\mathrm{KL}\left(p_\theta(\boldsymbol{x}|y=0)\|\frac{p_{\theta_0}(\boldsymbol{x}|y=0)+p_{\theta_0}(\boldsymbol{x}|y=1)}{2}\right) + \\
&\quad p(y=1)\cdot\mathrm{KL}\left(p_\theta(\boldsymbol{x}|y=1)\|\frac{p_{\theta_0}(\boldsymbol{x}|y=0)+p_{\theta_0}(\boldsymbol{x}|y=1)}{2}\right).
\end{aligned} \tag{4}
$$

Taking derivatives w.r.t $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ we get

$$
\begin{aligned}
&\nabla_\theta \mathbb{E}_{p(y)}\left[\mathrm{KL}(p_\theta(\boldsymbol{x}|y)\|p_{\theta_0}(\boldsymbol{x}))\right]|_{\theta=\theta_0} \\
&= \frac{1}{2}\int_{\boldsymbol{x}} \nabla_\theta p_\theta(\boldsymbol{x}|y=0)\frac{p_{\theta_0}(\boldsymbol{x}|y=0)+p_{\theta_0}(\boldsymbol{x}|y=1)}{2}\,|_{\theta=\theta_0} + \\
&\quad \frac{1}{2}\int_{\boldsymbol{x}} \nabla_\theta p_\theta(\boldsymbol{x}|y=1)\frac{p_{\theta_0}(\boldsymbol{x}|y=0)+p_{\theta_0}(\boldsymbol{x}|y=1)}{2}\,|_{\theta=\theta_0} \\
&= \nabla_\theta JSD(p_\theta(\boldsymbol{x}|y=0)\|p_\theta(\boldsymbol{x}))\,|_{\theta=\theta_0}
\end{aligned} \tag{5}
$$

Taking derivatives of the both sides of Eq.(3) at w.r.t $\boldsymbol{\theta}$ at $\boldsymbol{\theta}_0$ and plugging the last equation of Eq.(5), we obtain the desired results. □

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

3 Lemme 2

Proof. For the reconstruction term:

$$\begin{aligned}
&\mathbb{E}_{p_{\theta_0}(\boldsymbol{x})}\left[\mathbb{E}_{q_\eta(\boldsymbol{z}|\boldsymbol{x},y)q_*^r(y|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},y)\right]\right] \\
&= \frac{1}{2}\mathbb{E}_{p_{\theta_0}(\boldsymbol{x}|y=1)}\left[\mathbb{E}_{q_\eta(\boldsymbol{z}|\boldsymbol{x},y=0),y=0\sim q_*^r(y|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},y=0)\right]\right] \\
&+ \frac{1}{2}\mathbb{E}_{p_{\theta_0}(\boldsymbol{x}|y=0)}\left[\mathbb{E}_{q_\eta(\boldsymbol{z}|\boldsymbol{x},y=1),y=1\sim q_*^r(y|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z},y=1)\right]\right] \\
&= \frac{1}{2}\mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mathbb{E}_{\tilde{q}_\eta(\boldsymbol{z}|\boldsymbol{x})}\left[\log \tilde{p}_\theta(\boldsymbol{x}|\boldsymbol{z})\right]\right] + const,
\end{aligned} \tag{6}$$

where $y=0\sim q_*^r(y|\boldsymbol{x})$ means $q_*^r(y|\boldsymbol{x})$ predicts $y=0$ with probability 1. Note that both $q_\eta(\boldsymbol{z}|\boldsymbol{x},y=1)$ and $p_\theta(\boldsymbol{x}|\boldsymbol{z},y=1)$ are constant distributions without free parameters to learn; $q_\eta(\boldsymbol{z}|\boldsymbol{x},y=0)=\tilde{q}_\eta(\boldsymbol{z}|\boldsymbol{x})$, and $p_\theta(\boldsymbol{x}|\boldsymbol{z},y=0)=\tilde{p}_\theta(\boldsymbol{x}|\boldsymbol{z})$.

For the KL prior regularization term:

$$\begin{aligned}
&\mathbb{E}_{p_{\theta_0}(\boldsymbol{x})}\left[\mathrm{KL}(q_\eta(\boldsymbol{z}|\boldsymbol{x},y)q_*^r(y|\boldsymbol{x})\|p(\boldsymbol{z}|y)p(y))\right] \\
&= \mathbb{E}_{p_{\theta_0}(\boldsymbol{x})}\left[\int q_*^r(y|\boldsymbol{x})\mathrm{KL}\left(q_\eta(\boldsymbol{z}|\boldsymbol{x},y)\|p(\boldsymbol{z}|y)\right)dy + \mathrm{KL}\left(q_*^r(y|\boldsymbol{x})\|p(y)\right)\right] \\
&= \frac{1}{2}\mathbb{E}_{p_{\theta_0}(\boldsymbol{x}|y=1)}\left[\mathrm{KL}\left(q_\eta(\boldsymbol{z}|\boldsymbol{x},y=0)\|p(\boldsymbol{z}|y=0)\right)+const\right] + \frac{1}{2}\mathbb{E}_{p_{\theta_0}(\boldsymbol{x}|y=1)}\left[const\right] \\
&= \frac{1}{2}\mathbb{E}_{p_{data}(\boldsymbol{x})}\left[\mathrm{KL}(\tilde{q}_\eta(\boldsymbol{z}|\boldsymbol{x})\|\tilde{p}(\boldsymbol{z}))\right].
\end{aligned} \tag{7}$$

Combining Eq.(6) and Eq.(7) we recover the conventional VAE objective in Eq.(7) in the paper. □

4 Importance Weighted GANs (IWGAN)

From Eq.(4) in the paper, we can view GANs as maximizing a lower bound of the "marginal log-likelihood":

$$\begin{aligned}
\log q(y) &= \log\int p_\theta(\boldsymbol{x}|y)\frac{q^r(y|\boldsymbol{x})p_{\theta_0}(\boldsymbol{x})}{p_\theta(\boldsymbol{x}|y)}d\boldsymbol{x} \\
&\geq \int p_\theta(\boldsymbol{x}|y)\log\frac{q^r(y|\boldsymbol{x})p_{\theta_0}(\boldsymbol{x})}{p_\theta(\boldsymbol{x}|y)}d\boldsymbol{x} \\
&= -\mathrm{KL}(p_\theta(\boldsymbol{x}|y)\|q^r(\boldsymbol{x}|y)) + const.
\end{aligned} \tag{8}$$

We can apply the same importance weighting method as in IWAE [1] to derive a tighter bound.

$$\begin{aligned}
\log q(y) &= \log\mathbb{E}\left[\frac{1}{k}\sum_{i=1}^k\frac{q^r(y|\boldsymbol{x}_i)p_{\theta_0}(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i|y)}\right] \\
&\geq \mathbb{E}\left[\log\frac{1}{k}\sum_{i=1}^k\frac{q^r(y|\boldsymbol{x}_i)p_{\theta_0}(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i|y)}\right] \\
&= \mathbb{E}\left[\log\frac{1}{k}\sum_{i=1}^k w_i\right] \\
&:= \mathcal{L}_k(y)
\end{aligned} \tag{9}$$

where we have denoted $w_i=\frac{q^r(y|\boldsymbol{x}_i)p_{\theta_0}(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i|y)}$. We recover the lower bound of Eq.(8) when setting $k=1$.

To maximize the importance weighted lower bound, we compute the gradient:

$$\begin{aligned}
\nabla_\theta\mathcal{L}_k(y) = \nabla_\theta\mathbb{E}_{\boldsymbol{x}_1,\dots,\boldsymbol{x}_k}\left[\log\frac{1}{k}\sum_{i=1}^k w_i\right] &= \mathbb{E}_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_k}\left[\nabla_\theta\log\frac{1}{k}\sum_{i=1}^k w(y,\boldsymbol{x}(\boldsymbol{z}_i,\boldsymbol{\theta}))\right] \\
&= \mathbb{E}_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_k}\left[\sum_{i=1}^k\widetilde{w_i}\nabla_\theta\log w(y,\boldsymbol{x}(\boldsymbol{z}_i,\boldsymbol{\theta}))\right],
\end{aligned} \tag{10}$$

where $\widetilde{w_i} = w_i / \sum_{i=1}^k w_i$ are the normalized importance weights. We expand the weight at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

$$w_i|_{\theta=\theta_0} = \frac{q^r(y|\boldsymbol{x}_i)p_{\theta_0}(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i|y)} = q^r(y|\boldsymbol{x}_i)\frac{\frac{1}{2}p_{\theta_0}(\boldsymbol{x}_i|y=0) + \frac{1}{2}p_{\theta_0}(\boldsymbol{x}_i|y=1)}{p_{\theta_0}(\boldsymbol{x}_i|y)}|_{\theta=\theta_0}. \tag{11}$$

The ratio of $p_{\theta_0}(\boldsymbol{x}_i|y=0)$ and $p_{\theta_0}(\boldsymbol{x}_i|y=1)$ is intractable. Using the Bayes' rule and approximating with the discriminator distribution, we have

$$\frac{p(\boldsymbol{x}|y=0)}{p(\boldsymbol{x}|y=1)} = \frac{p(y=0|\boldsymbol{x})p(y=1)}{p(y=1|\boldsymbol{x})p(y=0)} \approx \frac{q(y=0|\boldsymbol{x})}{q(y=1|\boldsymbol{x})}. \tag{12}$$

Plug Eq.(12) into the above we have

$$w_i|_{\theta=\theta_0} \approx \frac{q^r(y|\boldsymbol{x}_i)}{q(y|\boldsymbol{x}_i)}. \tag{13}$$

In Eq.(10), the derivative $\nabla_\theta \log w_i$ is

$$\nabla_\theta \log w(y, \boldsymbol{x}(\boldsymbol{z}_i, \boldsymbol{\theta})) = \nabla_\theta \log q^r(y|\boldsymbol{x}(\boldsymbol{z}_i, \boldsymbol{\theta})) + \nabla_\theta \log \frac{p_{\theta_0}(\boldsymbol{x}_i)}{p_\theta(\boldsymbol{x}_i|y)}. \tag{14}$$

Similar to GAN, we omit the second term on the RHS of the equation. Therefore, the resulting update rule of $p_\theta(\boldsymbol{x}|y)$ is

$$\nabla_\theta \mathcal{L}_k(y) = \mathbb{E}_{\boldsymbol{z}_1,\dots,\boldsymbol{z}_k}\left[\sum_{i=1}^k \frac{q^r(y|\boldsymbol{x}_i)}{q(y|\boldsymbol{x}_i)}\nabla_\theta \log q^r(y|\boldsymbol{x}(\boldsymbol{z}_i, \boldsymbol{\theta}))\right] \tag{15}$$

5 Experimental Results of SVAE

Table 1 shows the results.

	1%	10%
SVAE	0.9412±.0039	0.9768±.0009
AASVAE	**0.9425±.0045**	**0.9797±.0010**

Table 1: Classification accuracy of semi-supervised VAEs and the adversary activated extension on the MNIST test set, with varying size of real labeled training examples.

References

[1] Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[2] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*, 2016.

[3] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[4] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu. Variational recurrent adversarial deep domain adaptation. In *ICLR*, 2017.

[5] L. Qin, Z. Zhang, H. Zhao, Z. Hu, and E. P. Xing. Adversarial connective-exploiting networks for implicit discourse relation classification. *arXiv preprint arXiv:1704.00217*, 2017.