Farewell to Mutual Information: Variational Distillation for Cross-Modal Person Re-Identification

1. Appendix

In this section, we introduce and prove the theorems mentioned in the main text of this paper.

A. ON SUFFICIENCY

Consider $x \in \mathbb{X}$ and y as the input data and the label, and let the v be an observation containing the same amount of predictive information regarding y as x does, and let z be the corresponding representation produced by an information bottleneck.

Hypothesis:

(H_1) v is sufficient for y, *i.e.*, $I(v;y) = I(x;y)$

Thesis:

(T_1) $\min I(v;y) - I(z;y) \iff \min H(y|z) - H(y|v)$

(T_2) *reducing* $D_{KL}[p(y|v)||p(y|z)]$ *is consistent with preserving sufficiency of* z *for* y*.*

Proof.

Based on the definition of mutual information [1]:

$$I(v;z) := H(v) - H(v|z), \tag{1}$$

where $H(v)$ denotes Shannon entropy, and $H(v|z)$ is the conditional entropy of z given v [1]. Based on the symmetry of mutual information, we have:

$$I(v;z) = I(z;v), \tag{2}$$

which indicates that the requirement of sufficiency is equivalent to:

$$\begin{aligned} & I(v;y) = I(z;y) \\ \iff & I(y;v) = I(y;z) \\ \iff & H(y) - H(y|v) = H(y) - H(y|z) \\ \iff & -H(y|v) = -H(y|z). \end{aligned} \tag{3}$$

Therefore, we have:

$$\min I(v;y) - I(z;y) \iff \min H(y|z) - H(y|v), \tag{4}$$

which proves (T_1). Based on the definition of conditional entropy, for any continuous variables v, y and z, we have:

$$\begin{aligned} & I(v;y) - I(z;y) = H(y|z) - H(y|v) = \\ & -\int p(z)dz \int p(y|z)\log p(y|z)dy \\ & +\int p(v)dv \int p(y|v)\log p(y|v)dy = \\ & -\iint p(z)p(y|z)\log\left[\frac{p(y|z)}{p(y|v)}p(y|v)\right]dzdy \\ & +\iint p(v)p(y|v)\log\left[\frac{p(y|v)}{p(y|z)}p(y|z)\right]dvdy. \end{aligned} \tag{5}$$

By factorizing the double integrals in Eq. (5) into another two components, we show the following:

$$\iint p(z)p(y|z)\log\left[\frac{p(y|z)}{p(y|v)}p(y|v)\right]dzdy =$$

$$\underbrace{\iint p(z)p(y|z)\log\frac{p(y|z)}{p(y|v)}dzdy}_{\text{term } Z_1} + \underbrace{\iint p(z)p(y|z)\log p(y|v)dzdy}_{\text{term } Z_2}. \tag{6}$$

Conduct similar factorization for the second term in Eq.(5), we have:

$$\iint p(v)p(y|v)\log\left[\frac{p(y|v)}{p(y|z)}p(y|z)\right]dvdy =$$

$$\underbrace{\iint p(v)p(y|v)\log\frac{p(y|v)}{p(y|z)}dvdy}_{\text{term } V_1} + \underbrace{\iint p(v)p(y|v)\log p(y|z)dvdy}_{\text{term } V_2}. \tag{7}$$

Integrate term Z_1 and term V_1 over y:

$$Z_1 = \int p(z) D_{KL}[p(y|z) \| p(y|v)] dz, \tag{8}$$

$$V_1 = \int p(v) D_{KL}[p(y|v) \| p(y|z)] dv, \tag{9}$$

where D_{KL} denotes KL-divergence. Integrate term Z_2 and term V_2 over z and v respectively, we have:

$$Z_2 = \int p(y) \log p(y|v) dy. \tag{10}$$

$$V_2 = \int p(y) \log p(y|z) dy \tag{11}$$

In the view of above, we have the following:

$$\begin{aligned} & I(v;y) - I(z;y) = H(y|z) - H(y|v) = \\ & \int p(v) D_{KL}[p(y|v) \| p(y|z)] dv + \int p(y) \log \left[\frac{p(y|z)}{p(y|v)} \right] dy \\ & - \int p(z) D_{KL}[p(y|z) \| p(y|v)] dz \end{aligned} \tag{12}$$

Based on the non-negativity of KL-divergence, Eq. (12) is upper bounded by:

$$\int p(v) D_{KL}[p(y|v) \| p(y|z)] dv + \int p(y) \log \left[\frac{p(y|z)}{p(y|v)} \right] dy. \tag{13}$$

Equivalently, we have the upper bound as:

$$\begin{aligned} & \mathbb{E}_{v \sim E_\theta(v|x)} \mathbb{E}_{z \sim E_\phi(z|v)} [D_{KL}[p(y|v) \| p(y|z)]] \\ & + \mathbb{E}_{v \sim E_\theta(v|x)} \mathbb{E}_{z \sim E_\phi(z|v)} \left[\log \left[\frac{p(y|z)}{p(y|v)} \right] \right], \end{aligned} \tag{14}$$

where θ, ϕ denote the parameters of the encoder and the information bottleneck. Therefore, the objective of preserving sufficiency of z for y can be formalized as:

$$\min_{\theta,\phi} \mathbb{E}_{v \sim E_\theta(v|x)} \mathbb{E}_{z \sim E_\phi(z|v)} \left[D_{KL}[\mathbb{P}_v || \mathbb{P}_z] + \log \left[\frac{\mathbb{P}_z}{\mathbb{P}_v} \right] \right], \tag{15}$$

in which $\mathbb{P}_z = p(y|z)$ and $\mathbb{P}_v = p(y|v)$ denote the predicted distributions of the representation and observation.

Clearly, the objective of preserving sufficiency is equivalent to minimize the discrepancy between the predicted distributions of v and z. Notice that this can be achieved by minimizing $D_{KL}(\mathbb{P}_v || \mathbb{P}_z)$, which can explicitly approximate $p(y|z)$ to $p(y|v)$ and implicitly reduce the second term in Eq.(15) in the same time. At the extreme, the representation z retrieves all label information contained in the sufficient observation v, indicating that z is sufficient for y as well. Formally, we have:

$$\lim_{\mathbb{P}_z \to \mathbb{P}_v} D_{KL}[\mathbb{P}_v || \mathbb{P}_z] + \int p(y) \log \left[\frac{\mathbb{P}_v}{\mathbb{P}_z} \right] dy = 0 \tag{16}$$

Based on Eq. (12) , we show the following:

$$\lim_{\mathbb{P}_z \to \mathbb{P}_v} I(v;y) - I(z;y) = \lim_{\mathbb{P}_z \to \mathbb{P}_v} H(y|v) - H(y|z) = 0 \tag{17}$$

which reveals that minimizing $D_{KL}[\mathbb{P}_v || \mathbb{P}_z]$ is consistent with the objective of preserving sufficiency of the representation. Thus (T_2) holds.

B. ON CONSISTENCY

Consider v_1, v_2 as two sufficient observations of the same objective x from different viewpoints or modals, and let y be the label. Let z_1, z_2 be the corresponding representations obtained from an information bottleneck.

Hypothesis:

(H_1) both v_1, v_2 are sufficient for y

(H_2) z_1, z_2 are in the same distribution

Thesis:

(T_1) minimizing $D_{KL}[p(y|v_2) \| p(y|z_1)]$ is consistent with the objective of eliminating task-irrelevant information encoded in $I(z_1; v_2)$, and is able to preserve those predictive and view-consistent information, vice versa for $D_{KL}[p(y|v_1) \| p(y|z_2)]$ and $I(z_2; v_1)$

(T_2) minimizing $D_{JS}[p(y|z_1) \| p(y|z_2)]$ is consistent with the objective of elimination of view-specific information for both z_1 and z_2

(T_3) performing VCD and VML can promote view-consistency between z_1 and z_2

Proofs.

By factorizing the mutual information between the data observation v_1 and its representation z_1, we have:

$$I(v_1; z_1) = I(v_1; z_1 | v_2) + I(z_1; v_2), \tag{18}$$

where $I(z_1; v_2)$ and $I(v_1; z_1 | v_2)$ denote the view-consistent and view-specific information, respectively.

Furthermore, by using the chain rule of mutual information, which subdivides $I(z_1; v_2)$ into two components (proofs could be found in [2]), we have:

$$I(z_1; v_2) = I(v_2; z_1 | y) + I(z_1; y) \tag{19}$$

combining with Eq. (18), we show the following:

$$I(z_1; v_1) = \underbrace{I(v_1; z_1 | v_2)}_{\text{view-specific}} + \underbrace{I(v_2; z_1 | y)}_{\text{superfluous}} + \underbrace{I(z_1; y)}_{\text{predictive}}, \tag{20}$$

Based on Appendix A, reducing $D_{KL}[\mathbb{P}_{v_2}||\mathbb{P}_{z_1}]$, where $\mathbb{P}_{z_1} = p(y|z_1), \mathbb{P}_{v_2} = p(y|v_2)$, can minimize $I(v_2; z_1|y)$ and maximize $I(z_1; y)$ in the same time, thus we conclude that (T_1) holds.

Considering that $z_1, z_2 \in \mathbb{Z}$, $I(v_1; z_1|v_2)$ can be expressed as:

$$\begin{aligned}
&I(v_1; z_1|v_2) = \mathbb{E}_{v_1,v_2\sim E_\theta(v|x)}\mathbb{E}_{z_1,z_2\sim E_\phi(z|v)}\left[\log\frac{p(z_1|v_1)}{p(z_1|v_2)}\right] \\
&= \mathbb{E}_{v_1,v_2\sim E_\theta(v|x)}\mathbb{E}_{z_1,z_2\sim E_\phi(z|v)}\left[\log\frac{p(z_1|v_1)p(z_2|v_2)}{p(z_2|v_2)p(z_1|v_2)}\right] \\
&= D_{KL}[p(z_1|v_1)||p(z_2|v_2)] - D_{KL}[p(z_2|v_1)||p(z_2|v_2)] \\
&\leq D_{KL}[p(z_1|v_1)||p(z_2|v_2)]. \qquad (21)
\end{aligned}$$

Notice this bound is tight whenever z_1 and z_2 produce consistent encodings [2], which can be assured by the proposed VCD and is visualized in the main body of this paper. On the other hand, since y is constant with respect to the parameters to be optimized, we utilize Eq. (22) to approximate Eq. (21):

$$\mathbb{E}_{v_1,v_2\sim E_\theta(v|x)}\mathbb{E}_{z_1,z_2\sim E_\phi(z|v)}\left[D_{KL}[\mathbb{P}_{z_1}||\mathbb{P}_{z_2}]\right], \qquad (22)$$

in which $\mathbb{P}_{z_1} = p(y|z_1)$ and $\mathbb{P}_{z_2} = p(y|z_2)$ denote the predicted distributions. Based on the above analysis, we conclude that $I(v_1; z_1|v_2)$ can be minimized by reducing $D_{KL}[\mathbb{P}_{z_1}||\mathbb{P}_{z_2}]$. Similarly, we introduce the following objective to minimize $I(v_2; z_2|v_1)$.

$$\mathbb{E}_{v_1,v_2\sim E_\theta(v|x)}\mathbb{E}_{z_1,z_2\sim E_\phi(z|v)}\left[D_{KL}[\mathbb{P}_{z_2}||\mathbb{P}_{z_1}]\right], \qquad (23)$$

For simplicity, we apply Eq. (24) to eliminate the view-specific information for both z_1 and z_2.

$$\min_{\theta,\phi}\mathbb{E}_{v_1,v_2\sim E_\theta(v|x)}\mathbb{E}_{z_1,z_2\sim E_\phi(z|v)}\left[D_{JS}[\mathbb{P}_{z_1}||\mathbb{P}_{z_2}]\right], \quad (24)$$

where D_{JS} denotes the Jensen-Shannon divergence. Thus (T_2) holds.

Finally, according to [2], $I(z_1; y) = I(v_1 v_2; y)$ when the following hypotheses stand: z_1 is a representation of v_1 and $I(y; z_1|v_1 v_2) = 0$, both v_1 and v_2 are sufficient for y, z_1 is sufficient for v_2. As a consequence of data processing inequality, the amount of information encoded in z_1 cannot be more than the joint observation, *i.e.* $I(y; z_1|v_1 v_2) \equiv 0$. Since sufficiency of v_1 and v_2 for y is consistent with the given task, it is widely adopted as an established assumption. Notably, sufficiency of z_1 for v_2 can be achieved by preserving view-consistent information while simultaneously eliminating the view-specific details, which correspond to the proposed VCD and VML, respectively. Therefore, (T_3) holds.

References

[1] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. Mutual information neural estimation. In *ICML*, 2018. 1

[2] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020. 2, 3