Note on the SimSiam objective

1 Notations

We are using similar notations to the SimSiam paper. For a single input image x the model generates two random augmented views $x_1 = \mathcal{T}_1(x)$ and $x_2 = \mathcal{T}_2(x)$. These views are then fed into an encoder $\mathcal{F}_\phi$ parameterized by ϕ

$$z_i = \mathcal{F}_\phi(\mathcal{T}_i(x)),\ i = 1, 2$$

Predictions are produced using a separate predictor network h parameterized by θ

$$p_i = h_\theta(z_i),\ i = 1, 2$$

Finally, the loss is computed as

$$\mathcal{L}_{\text{SimSiam}}(z_1, z_2) = \frac{1}{2}\mathcal{D}(p_1, \text{SG}(z_2)) + \frac{1}{2}\mathcal{D}(p_2, \text{SG}(z_1)) \tag{1}$$

where SG is stop gradient operator and $\mathcal{D}$ is some similarity measure (e.g., cosine similarity or L2 distance).

For the subsequent derivations we assume an input image x to be fixed and transformations $\mathcal{T}_1, \mathcal{T}_2$ are randomly and independently sampled. In this context, view encodings z_1 and z_2 also become random variables.

2 SimSiam and Mutual Information

We are going to show that minimizing $\mathcal{L}_{\text{SimSiam}}$ is equivalent to maximizing the lowerbound on the mutual information between different views encodings z_1, z_2 of the same image x. In other words,

$$\text{E}_{\mathcal{T}_1,\mathcal{T}_2}[\mathcal{L}_{\text{SimSiam}}] \geq \text{constant} - \mathcal{I}(z_1, z_2) \tag{2}$$

which becomes tighter as the predictor h_θ becomes optimal. This makes the training objective very similar to the Contrastive Predictive Coding or InfoNCE (or, SimCLR in a context of images).

First, let $Q(\cdot;\mu)$ be a probabilistic distribution over view encodings parameterized by some μ. Then,

$$\mathcal{D}(h_\theta(z1), z_2) = -\log Q(z_2; \mu = h_\theta(z_1)) \tag{3}$$

for an appropriate choice of Q. For example, when $\mathcal{D}$ is L2 distance then Q is multivariate Gaussian distribution with an identity covariance and mean μ.

Our first observation is that $Q(\cdot;\mu = h_\theta(z_1))$ is trained[1] to approximate $P(z_2 \mid z_1)$, since the objective function is a cross-entropy between these two distributions. More formally, consider the expected loss conditioned on a known z_1

$$\begin{aligned}
\mathrm{E}_{\mathcal{T}_2}\left[\mathcal{D}(h_\theta(z1), z_2) \mid z_1\right] &= \mathrm{E}_{\mathcal{T}_2}\left[-\log Q(z_2;\mu = h_\theta(z_1)) \mid z_1\right] \\
&= \mathrm{E}_{\mathcal{T}_2}\left[-\log Q(z_2;\mu = h_\theta(z_1)) + \log P(z_2|z_1) - \log P(z_2|z_1) \mid z_1\right] \\
&= \mathrm{E}_{\mathcal{T}_2}\left[-\log P(z_2 \mid z_1)\right] + D_{KL}(P(z_2|z_1) \mid\mid Q(z_2;\mu = h_\theta(z_1)) \\
&\geq \mathrm{E}_{\mathcal{T}_2}\left[-\log P(z_2 \mid z_1)\right] \qquad (4)
\end{aligned}$$

where the inequality becomes more tight when $Q(\cdot;\mu = h_\theta(z_1))$ approximates $P(z_2 \mid z_1)$ better, which happens when parameters θ are closer to optimum. This corresponds to the empirical evidence by SimSiam and follow up papers that the model benefits from the predictor h_θ being optimal – for example by making several gradient updates or using higher learning rate just for θ.

If we take expectation with respect to the $\mathcal{T}_1$

$$\begin{aligned}
\mathrm{E}_{\mathcal{T}_1,\mathcal{T}_2}[\mathcal{D}(p_1, z_2)] &\geq \mathrm{E}_{\mathcal{T}_1,\mathcal{T}_2}\left[-\log P(z_2|z_1)\right] \\
&= \mathrm{E}_{\mathcal{T}_1,\mathcal{T}_2}\left[-\log \frac{P(z_1, z_2)}{P(z_1)}\right] \\
&= \mathrm{E}_{\mathcal{T}_1,\mathcal{T}_2}\left[-\log \frac{P(z_1, z_2)}{P(z_1)P(z_2)} - \log P(z_2)\right] \\
&= \mathcal{H}(z_2) - \mathcal{I}(z_1, z_2) \qquad (5)
\end{aligned}$$

Finally, we can subtitute it back to the $\mathcal{L}_{\text{SimSiam}}$ and add SG operations

$$\begin{aligned}
\mathrm{E}_{\mathcal{T}_1,\mathcal{T}_2}\left[\frac{1}{2}\mathcal{D}(p_1, \mathrm{SG}(z_2)) + \frac{1}{2}\mathcal{D}(p_2, \mathrm{SG}(z_1))\right] &\geq \frac{1}{2}\mathcal{H}(\mathrm{SG}(z_1)) + \frac{1}{2}\mathcal{H}(\mathrm{SG}(z_2)) \\
&\quad - \frac{1}{2}\left(\mathcal{I}(z_1, \mathrm{SG}(z_2)) + \mathcal{I}(\mathrm{SG}(z_1), z_2)\right) \\
&= \frac{1}{2}\mathcal{H}(\mathrm{SG}(z_1)) + \frac{1}{2}\mathcal{H}(\mathrm{SG}(z_2)) - \mathcal{I}(z_1, z_2) \qquad (6)
\end{aligned}$$

where we can treat $\mathcal{H}(\mathrm{SG}(z_1))$ and $\mathcal{H}(\mathrm{SG}(z_2))$ as constants because of the SG operations.

[1] meaning that the predictor h_θ is being optimized