Supplements: Counterfactual Vision and Language Learning

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, Anton van den Hengel
{ehsan.abbasnejad,damien.teney,amin.parvaneh,javen.shi,anton.vandenhengel}@adelaide.edu.au
Australian Institute for Machine Learning & The University of Adelaide, Australia

Theorem 1 (Theorem 4)**.** *Denote* $u^i(\boldsymbol{\theta}) \equiv \ell(f_{\boldsymbol{\theta}}(\boldsymbol{q}_i, \boldsymbol{v}_i), a_i)\omega_i(\boldsymbol{\theta})$, $\overline{u} \equiv \sum_{i=1}^n u^i(\boldsymbol{\theta})/n$, $\hat{\mathbb{V}}(u) \equiv \sum_{i=1}^n \left(u^i(\boldsymbol{\theta}) - \overline{u}\right)^2/(n-1)$ *and* $\mathcal{Q}_\gamma \equiv \log(10 \cdot \epsilon/\gamma)$ *for* $0 < \gamma < 1$ *and* ϵ *the* ϵ*-cover for the function class that predicts the answer. With probability at least* $1-\gamma$ *for* $n \geq 16$ *we have*

$$R(\boldsymbol{\theta}) \leq \hat{R}^M(\boldsymbol{\theta}) + \sqrt{18\hat{\mathbb{V}}(u)\mathcal{Q}_\gamma/n} + 15M\mathcal{Q}_\gamma/(n-1)$$

Proof. Follows the proof in Theorem 6 of [2]. □

The density of the counterfactuals based on the observations, i.e.

$$p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}) = \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p(\boldsymbol{q},\boldsymbol{v})}\left[p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})\right] \tag{5}$$

Proof of Equation 5. We have:

$$\begin{aligned}
p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}) &= \int p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u})p^{\mathrm{do}(I)}(\mathbf{u})d\mathbf{u} \\
&= \int p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u})p(\mathbf{u})d\mathbf{u} \\
&= \int p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u})\left(\int p(\boldsymbol{q}, \boldsymbol{v}, \mathbf{u})\, dp(\boldsymbol{q}, \boldsymbol{v})\right) d\mathbf{u} \\
&= \iint p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u})p(\mathbf{u}|\boldsymbol{q}, \boldsymbol{v})\, dp(\boldsymbol{q}, \boldsymbol{v})\, d\mathbf{u} \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\left[\int p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u})p(\mathbf{u}|\boldsymbol{q}, \boldsymbol{v})\, d\mathbf{u}\right] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\left[p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})\right]
\end{aligned}$$

□

Proof of Lemma (5). We have,

$$\begin{aligned}
p^{\mathrm{do}(I)}(a, \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}) &= p^{\mathrm{do}(I)}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}) \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\left[p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})\right].
\end{aligned}$$

Then using Jensen's inequality we have,

$$\begin{aligned}
&\log(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\left[p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})\right]) \\
&\quad \geq \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\left[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}))\right],
\end{aligned}$$

We have:

$$\begin{aligned}
&\mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\left[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}))\right] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})) \\
&\qquad + \log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}))\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})) \\
&\qquad + \log(\int p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u})p(\mathbf{u}|\boldsymbol{q}, \boldsymbol{v})\, d\mathbf{u})\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})) \\
&\qquad + \log(\int p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u})p(\mathbf{u}|\boldsymbol{q}, \boldsymbol{v})\, d\mathbf{u})\Big]
\end{aligned}$$

which is then lower-bounded as

$$\begin{aligned}
&\geq \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})) \\
&\qquad + \int \log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u}))p(\mathbf{u}|\boldsymbol{q}, \boldsymbol{v})\, d\mathbf{u}\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}))\Big] \\
&\qquad + \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\int \log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u}))p(\mathbf{u}|\boldsymbol{q}, \boldsymbol{v})\, d\mathbf{u}\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}))\Big] \\
&\qquad + \int \log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u}))p(\mathbf{u})\, d\mathbf{u}
\end{aligned}$$

□

Log-density of the joint for the question,image and answer as Eq. (8) in the paper:

$$\begin{aligned}
\log(p^{\mathrm{do}(I)}(a, \tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}})) \geq\ & \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p(\boldsymbol{q},\boldsymbol{v})}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}))\Big] \\
&+ \mathbb{E}_q[\log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}}, \tilde{\boldsymbol{v}}|\mathbf{u}))] \\
&+ H(q) - H_q(p).
\end{aligned} \tag{8}$$

Proof of Equation 8. If we want to use an alternative distri-

bution

$$
\begin{aligned}
&\mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}))\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) \\
&\qquad + \log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}))\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) \\
&\qquad + \log(\int p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u})p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right)d\mathbf{u})\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) \\
&\qquad + \log(\int \frac{p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u})p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right)}{q(\mathbf{u})}q(\mathbf{u})d\mathbf{u})\Big]
\end{aligned}
$$

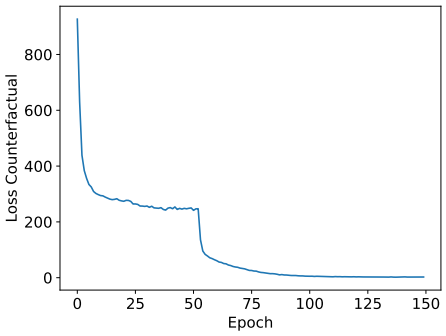
which is then lower-bounded as

$$
\begin{aligned}
&\geq \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) \\
&\qquad + \int \log(\frac{p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u})p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right)}{q(\mathbf{u})})q(\mathbf{u})d\mathbf{u}\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) \\
&\qquad + \int \log(\frac{p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u})p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right)}{q(\mathbf{u})})q(\mathbf{u})d\mathbf{u}\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) \\
&\qquad + \int (\log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u})p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right)) \\
&\qquad\qquad - \log(q(\mathbf{u})))q(\mathbf{u})d\mathbf{u}\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) \\
&\quad + \int (\log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u})p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right))q(\mathbf{u})d\mathbf{u}\Big] + H(q) \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) \\
&\qquad\qquad + \int \log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u}))q(\mathbf{u})d\mathbf{u} \\
&\qquad\qquad + \int \log(p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right))q(\mathbf{u})d\mathbf{u}\Big] + H(q) \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) + H(q) \\
&\qquad + \mathbb{E}_q[\log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u}))] + \mathbb{E}_q[\log(p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right))]\Big] \\
&= \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) + \mathbb{E}_q[\log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u}))]\Big] \\
&\qquad + H(q) + \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\mathbb{E}_q[\log(p\left(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}\right))] \\
&\geq \mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}\Big[\log(p^{\mathrm{do}(I)|\boldsymbol{q},\boldsymbol{v}}(a|\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}})) + \mathbb{E}_q[\log(p^{\mathrm{do}(I)}(\tilde{\boldsymbol{q}},\tilde{\boldsymbol{v}}|\mathbf{u}))]\Big] \\
&\qquad + H(q) - H_q(p)
\end{aligned}
$$

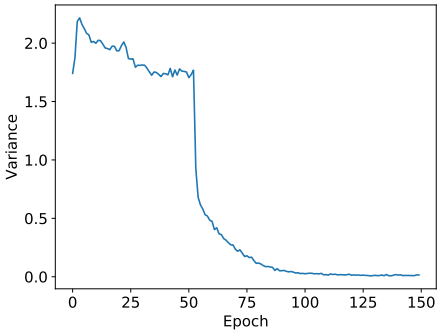
We have

$$
\begin{aligned}
\text{-}\mathbb{E}_q\mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}[\log(\mathbf{p}(\mathbf{u}|\boldsymbol{q},\boldsymbol{v}))] \;&\geq\; \text{-}\mathbb{E}_q\log(\mathbb{E}_{(\boldsymbol{q},\boldsymbol{v})\sim p}[\mathbf{p}(\mathbf{u}|\boldsymbol{q},\boldsymbol{v})] \\
&= -\mathbb{E}_q\log([p\left(\mathbf{u}\right)] = H_q(p)
\end{aligned}
$$

□



(a) Loss ResNet18

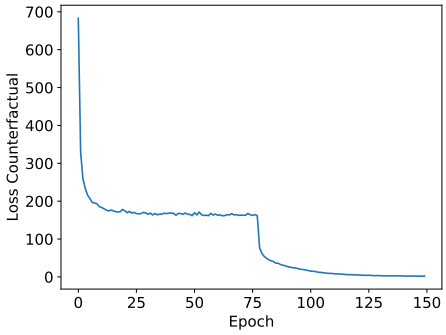


(b) Loss Variance ResNet18

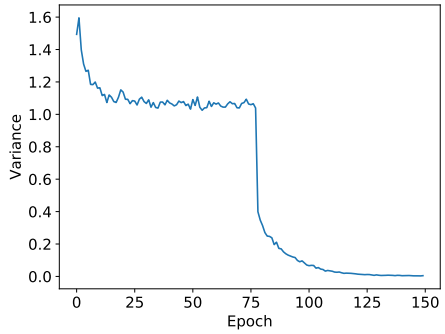
Figure 1: CIFAR10 results

1. Implementation details

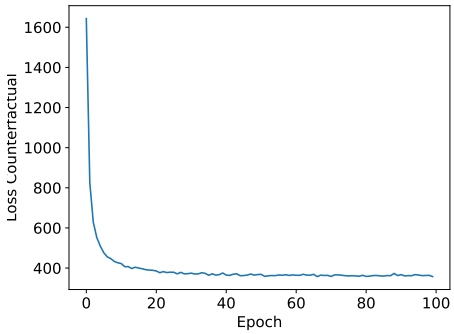
We implemented our approach on top of the original UpDn system [1]. The base system utilizes a Faster R-CNN head in conjunction with a ResNet-101 base network as the object detection module. For the VQA v2 experiment we utilize the ResNet-152 for detection. The detection head is pre-trained on the Visual Genome dataset. UpDn takes the final detection outputs and performs non-maximum suppression (NMS) for each object category using an IoU threshold of 0.7. Then, the convolutional features for the top 36 objects are extracted for each image as the visual features. For question embedding, we perform standard text pre-processing and tokenization. In particular, questions are first converted to lower case and then trimmed to a maximum of 14 words, and the words that appear less than 5 times are replaced with an "<unk>" token. We use GloVe embeddings and subsequently GRU for VQA-CP and LSTM for VQA v2A to sequentially process the word vectors and produce a sentential representation for the pre-processed question.



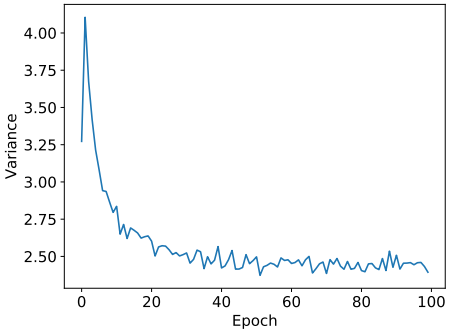
(a) Loss PreResNet18-CIFAR10



(b) Loss Variance PreResNet18-CIFAR10



(c) Loss PreResNet18-CIFAR100



(d) Loss Variance PreResNet18-CIFAR100

Figure 2: CIFAR results

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.