VIBI: Explaining a Black-Box using Deep Variational Information Bottleneck Approach

Seojin Bang, Pengtao Xie, Heewook Lee, Wei Wu, and Eric Xing

Kurt Willis

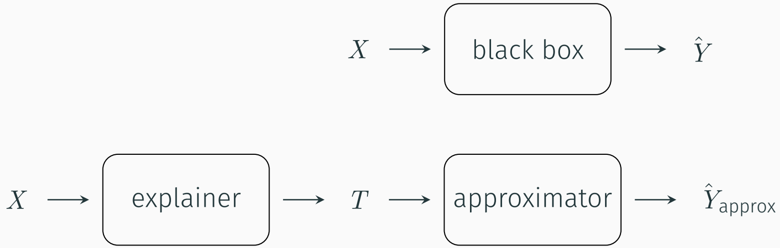
July 20, 2021

Overview

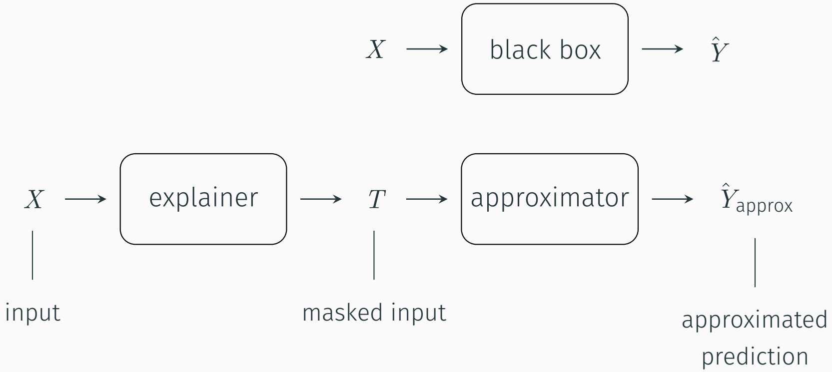
1. Introduction to VIBI
2. Information Bottleneck & Variational Bound
3. Gumbel Softmax
4. Results

Introduction to VIBI

VIBI


X
black box
$\hat{Y}$
X
explainer
T
approximator
$\hat{Y}_{\text{approx}}$

VIBI



X
black box
$\hat{Y}$
X
explainer
T
approximator
$\hat{Y}_{\text{approx}}$
input
masked input
approximated prediction

Notation and Conventions

s : scalar

$\mathbf{x}$: vector

X : random variable

$$p(\mathbf{x}) = p_X(X = \mathbf{x})$$

$$p(\mathbf{x}, \mathbf{y}) = p_{X,Y}(X = \mathbf{x},\, Y = \mathbf{y})$$

$$p(\mathbf{x} \,|\, \mathbf{y}) = p_{X|\,Y}(X = \mathbf{x} \,|\, Y = \mathbf{y}) = \frac{p_{X,Y}(X = \mathbf{x},\, Y = \mathbf{y})}{p_Y(Y = \mathbf{y})}$$

Markov-Chain Assumption

Markov-chain assumption:

$$Y \longleftrightarrow X \longleftrightarrow T$$

- Y is the *discrete* RV over the **class labels**, $\mathbf{y} \in \mathcal{Y} = \{1, \ldots, 10\}$.

Markov-Chain Assumption

Markov-chain assumption:

$$Y \longleftrightarrow X \longleftrightarrow T$$

- Y is the *discrete* RV over the **class labels**, $\mathbf{y} \in \mathcal{Y} = \{1, \ldots, 10\}$.
- X is the *continuous* RV over the **image space**, $\mathbf{x} \in \mathcal{X} = [0, 1]^d$

Markov-Chain Assumption

Markov-chain assumption:

$$Y \longleftrightarrow X \longleftrightarrow T$$

- Y is the *discrete* RV over the **class labels**, $\mathbf{y} \in \mathcal{Y} = \{1, \ldots, 10\}$.
- X is the *continuous* RV over the **image space**, $\mathbf{x} \in \mathcal{X} = [0, 1]^d$
- T is the RV over $\mathcal{T} = \mathcal{X}$

Markov-Chain Assumption

Markov-chain assumption:

$$Y \longleftrightarrow X \longleftrightarrow T$$

- Y is the *discrete* RV over the **class labels**, $\mathbf{y} \in \mathcal{Y} = \{1, \ldots, 10\}$.
- X is the *continuous* RV over the **image space**, $\mathbf{x} \in \mathcal{X} = [0, 1]^d$
- T is the RV over $\mathcal{T} = \mathcal{X}$

d is the image dimension ($28 \times 28 = 784$ in the case of the MNIST dataset).

Markov-Chain Assumption

Markov-chain assumption:

$$Y \longleftrightarrow X \longleftrightarrow T$$

The Markov-chain assumption leads to the joint-distribution

$$p(\mathbf{x}, \mathbf{y}, \mathbf{t}) = p(\mathbf{y} \mid \mathbf{t}, \mathbf{x}) p(\mathbf{t}, \mathbf{x})$$

Markov-Chain Assumption

Markov-chain assumption:

$$Y \longleftrightarrow X \longleftrightarrow T$$

The Markov-chain assumption leads to the joint-distribution

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{t}) &= p(\mathbf{y} \mid \mathbf{t}, \mathbf{x}) p(\mathbf{t}, \mathbf{x}) \\ &= p(\mathbf{y} \mid \mathbf{t}, \mathbf{x}) p(\mathbf{t} \mid \mathbf{x}) p(\mathbf{x}) \end{aligned}$$

Markov-Chain Assumption

Markov-chain assumption:

$$Y \longleftrightarrow X \longleftrightarrow T$$

The Markov-chain assumption leads to the joint-distribution

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{t}) &= p(\mathbf{y} \mid \mathbf{t}, \mathbf{x}) p(\mathbf{t}, \mathbf{x}) \\ &= p(\mathbf{y} \mid \mathbf{t}, \mathbf{x}) p(\mathbf{t} \mid \mathbf{x}) p(\mathbf{x}) \\ &= p(\mathbf{y} \mid \mathbf{t}) p(\mathbf{t} \mid \mathbf{x}) p(\mathbf{x}) \end{aligned}$$

Markov-Chain Assumption

The explainer network models a probability distribution $p(\mathbf{z} \,|\, \mathbf{x})$ over cognitive chunks. Z is drawn as a k-hot vector from $p(\mathbf{z} \,|\, \mathbf{x})$. The resulting explanation is given by $\mathbf{t} = \mathbf{x} \odot \mathbf{z}$

Full markov-chain:

$$Y \longleftrightarrow X \longleftrightarrow (X, Z) \longleftrightarrow T$$

Information Bottleneck & Variational Bound

Information Bottleneck

The Information Bottleneck (IB) objective as stated by Tishby, Pereira, and Bialek, 2000:

$$\max_{p_{T|X},\, p_{Y|T},\, p_T} I(T;Y) - \beta I(X;T)$$

Variational Bound - $I(X;T)$ Upper Bound

Markov-chain:

$$Y \longleftrightarrow X \longleftrightarrow (X, Z) \longleftrightarrow T$$

$$I(X;T) \leq I(X;X,Z)$$

Variational Bound - $I(X;T)$ Upper Bound

Markov-chain:

$$Y \longleftrightarrow X \longleftrightarrow (X, Z) \longleftrightarrow T$$

$$\begin{aligned} I(X;T) &\leq I(X;X,Z) \\ &= I(X;Z) + I(X;X \,|\, Z) \end{aligned}$$

Variational Bound - $I(X;T)$ Upper Bound

Markov-chain:

$$Y \longleftrightarrow X \longleftrightarrow (X, Z) \longleftrightarrow T$$

$$\begin{aligned} I(X;T) &\leq I(X;X,Z) \\ &= I(X;Z) + I(X;X \mid Z) \\ &= I(X;Z) + H(X \mid Z) + H(X \mid Z) - H(X,X \mid Z) \end{aligned}$$

Variational Bound - $I(X;T)$ Upper Bound

Markov-chain:

$$Y \longleftrightarrow X \longleftrightarrow (X, Z) \longleftrightarrow T$$

$$\begin{aligned}
I(X;T) &\leq I(X;X,Z) \\
&= I(X;Z) + I(X;X \mid Z) \\
&= I(X;Z) + H(X \mid Z) + H(X \mid Z) - H(X, X \mid Z) \\
&= I(X;Z) + H(X \mid Z)
\end{aligned}$$

Variational Bound - $I(X;T)$ Upper Bound

Markov-chain:

$$Y \longleftrightarrow X \longleftrightarrow (X,Z) \longleftrightarrow T$$

$$\begin{aligned}
I(X;T) &\leq I(X;X,Z) \\
&= I(X;Z) + I(X;X \mid Z) \\
&= I(X;Z) + H(X \mid Z) + H(X \mid Z) - H(X,X \mid Z) \\
&= I(X;Z) + H(X \mid Z) \\
&\leq I(X;Z) + H(X)
\end{aligned}$$

Variational Bound - $I(X; Z)$ Upper Bound

$$I(X; Z) = \int p(\mathbf{x}, \mathbf{z}) \log \left(\frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x}) p(\mathbf{z})} \right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}$$

Variational Bound - $I(X;Z)$ Upper Bound

$$
\begin{aligned}
I(X;Z) &= \int p(\mathbf{x},\mathbf{z}) \log\left(\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})}\right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z}\mid\mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}
\end{aligned}
$$

Variational Bound - $I(X;Z)$ Upper Bound

$$
\begin{aligned}
I(X;Z) &= \int p(\mathbf{x},\mathbf{z}) \log \left(\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})} \right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \int p(\mathbf{z}) \log p(\mathbf{z}) \, \mathrm{d}\mathbf{z}
\end{aligned}
$$

Variational Bound - $I(X;Z)$ Upper Bound

$$
\begin{aligned}
I(X;Z) &= \int p(\mathbf{x},\mathbf{z}) \log \left(\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})} \right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \underbrace{\int p(\mathbf{z}) \log p(\mathbf{z}) \, \mathrm{d}\mathbf{z}}_{D_{\mathsf{KL}}(p||r) + H(p,r)}
\end{aligned}
$$

Variational Bound - $I(X;Z)$ Upper Bound

$$
\begin{aligned}
I(X;Z) &= \int p(\mathbf{x},\mathbf{z}) \log\left(\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})}\right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z}\,|\,\mathbf{x})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z}\,|\,\mathbf{x})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \underbrace{\int p(\mathbf{z}) \log p(\mathbf{z})\, \mathrm{d}\mathbf{z}}_{D_{\mathsf{KL}}(p||r)+H(p,r)} \\
&\leq \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z}\,|\,\mathbf{x})\, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \int p(\mathbf{z}) \log r(\mathbf{z})\, \mathrm{d}\mathbf{z}
\end{aligned}
$$

Variational Bound - $I(X;Z)$ Upper Bound

$$
\begin{aligned}
I(X;Z) &= \int p(\mathbf{x},\mathbf{z}) \log \left(\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})p(\mathbf{z})} \right) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \underbrace{\int p(\mathbf{z}) \log p(\mathbf{z}) \, \mathrm{d}\mathbf{z}}_{D_{\mathsf{KL}}(p||r) + H(p,r)} \\
&\leq \int p(\mathbf{x},\mathbf{z}) \log p(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z} - \int p(\mathbf{z}) \log r(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{x},\mathbf{z}) \log \left(\frac{p(\mathbf{z} \mid \mathbf{x})}{r(\mathbf{z})} \right) \mathrm{d}\mathbf{z}
\end{aligned}
$$

Variational Bound - $I(X;T)$ Upper Bound

$$I(X;T) \leq \int p(\mathbf{x},\mathbf{z}) \log \left(\frac{p(\mathbf{z} \mid \mathbf{x})}{r(\mathbf{z})} \right) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} + C$$

Variational Bound - $I(X;T)$ Upper Bound

$$
\begin{aligned}
I(X;T) &\leq \int p(\mathbf{x},\mathbf{z}) \log\left(\frac{p(\mathbf{z}\,|\,\mathbf{x})}{r(\mathbf{z})}\right) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} + C \\
&= \int p(\mathbf{x}) p(\mathbf{z}\,|\,\mathbf{x}) \log\left(\frac{p(\mathbf{z}\,|\,\mathbf{x})}{r(\mathbf{z})}\right) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} + C
\end{aligned}
$$

Variational Bound - $I(X;T)$ Upper Bound

$$
\begin{aligned}
I(X;T) &\leq \int p(\mathbf{x},\mathbf{z}) \log\left(\frac{p(\mathbf{z}\mid\mathbf{x})}{r(\mathbf{z})}\right) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} + C \\
&= \int p(\mathbf{x}) p(\mathbf{z}\mid\mathbf{x}) \log\left(\frac{p(\mathbf{z}\mid\mathbf{x})}{r(\mathbf{z})}\right) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} + C \\
&\approx \frac{1}{N}\sum_{n=1}^{N} \int p(\mathbf{z}\mid\mathbf{x}^{(n)}) \log\left(\frac{p(\mathbf{z}\mid\mathbf{x}^{(n)})}{r(\mathbf{z})}\right) \mathrm{d}\mathbf{z} + C
\end{aligned}
$$

Variational Bound - $I(X;T)$ Upper Bound

$$
\begin{aligned}
I(X;T) &\leq \int p(\mathbf{x},\mathbf{z}) \log\left(\frac{p(\mathbf{z}\mid\mathbf{x})}{r(\mathbf{z})}\right) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} + C \\
&= \int p(\mathbf{x}) p(\mathbf{z}\mid\mathbf{x}) \log\left(\frac{p(\mathbf{z}\mid\mathbf{x})}{r(\mathbf{z})}\right) \mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x} + C \\
&\approx \frac{1}{N}\sum_{n=1}^{N} \int p(\mathbf{z}\mid\mathbf{x}^{(n)}) \log\left(\frac{p(\mathbf{z}\mid\mathbf{x}^{(n)})}{r(\mathbf{z})}\right) \mathrm{d}\mathbf{z} + C \\
&= \frac{1}{N}\sum_{n=1}^{N} D_{\mathsf{KL}}\left(f_{\mathsf{xpl}}(\mathbf{x}^{(n)}) \,||\, r\right) + C
\end{aligned}
$$

Variational Bound - $I(X;T)$ Upper Bound

$$
\begin{aligned}
I(X;T) &\leq \int p(\mathbf{x}, \mathbf{z}) \log \left(\frac{p(\mathbf{z} \mid \mathbf{x})}{r(\mathbf{z})} \right) \mathrm{d}\mathbf{z} \mathrm{d}\mathbf{x} + C \\
&= \int p(\mathbf{x}) p(\mathbf{z} \mid \mathbf{x}) \log \left(\frac{p(\mathbf{z} \mid \mathbf{x})}{r(\mathbf{z})} \right) \mathrm{d}\mathbf{z} \mathrm{d}\mathbf{x} + C \\
&\approx \frac{1}{N} \sum_{n=1}^{N} \int p(\mathbf{z} \mid \mathbf{x}^{(n)}) \log \left(\frac{p(\mathbf{z} \mid \mathbf{x}^{(n)})}{r(\mathbf{z})} \right) \mathrm{d}\mathbf{z} + C \\
&= \frac{1}{N} \sum_{n=1}^{N} D_{\mathsf{KL}} \left(f_{\mathsf{xpl}}(\mathbf{x}^{(n)}) \,||\, r \right) + C \quad =: U_{X,T}
\end{aligned}
$$

Variational Bound - $I(T;Y)$ Lower Bound

$$I(T;Y) = \int p(\mathbf{t},\mathbf{y}) \log \left(\frac{p(\mathbf{t},\mathbf{y})}{p(\mathbf{t})p(\mathbf{y})} \right) \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}$$

Variational Bound - $I(T;Y)$ Lower Bound

$$
\begin{aligned}
I(T;Y) &= \int p(\mathbf{t},\mathbf{y}) \log \left(\frac{p(\mathbf{t},\mathbf{y})}{p(\mathbf{t})p(\mathbf{y})} \right) \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y}) \log p(\mathbf{y} \mid \mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} - \int p(\mathbf{t},\mathbf{y}) \log p(\mathbf{y}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}
\end{aligned}
$$

$$
\begin{aligned}
I(T;Y) &= \int p(\mathbf{t},\mathbf{y}) \log\left(\frac{p(\mathbf{t},\mathbf{y})}{p(\mathbf{t})p(\mathbf{y})}\right) \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y}) \log p(\mathbf{y} \mid \mathbf{t})\, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} - \int p(\mathbf{t},\mathbf{y}) \log p(\mathbf{y})\, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{y}) \int p(\mathbf{y} \mid \mathbf{t}) \log p(\mathbf{y} \mid \mathbf{t})\, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} - \int p(\mathbf{y}) \log p(\mathbf{y})\, \mathrm{d}\mathbf{y}
\end{aligned}
$$

Variational Bound - $I(T;Y)$ Lower Bound

$$
\begin{aligned}
I(T;Y) &= \int p(\mathbf{t},\mathbf{y}) \log\left(\frac{p(\mathbf{t},\mathbf{y})}{p(\mathbf{t})p(\mathbf{y})}\right) \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y}) \log p(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} - \int p(\mathbf{t},\mathbf{y}) \log p(\mathbf{y})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{y}) \int p(\mathbf{y}\,|\,\mathbf{t}) \log p(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} - \int p(\mathbf{y}) \log p(\mathbf{y})\,\mathrm{d}\mathbf{y} \\
&\overset{\text{(var. approx)}}{\geq} \int p(\mathbf{y}) \int p(\mathbf{y}\,|\,\mathbf{t}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} + H(Y)
\end{aligned}
$$

Variational Bound - $I(T;Y)$ Lower Bound

$$
\begin{aligned}
I(T;Y) &= \int p(\mathbf{t},\mathbf{y}) \log\left(\frac{p(\mathbf{t},\mathbf{y})}{p(\mathbf{t})p(\mathbf{y})}\right) \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y}) \log p(\mathbf{y}\,|\,\mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} - \int p(\mathbf{t},\mathbf{y}) \log p(\mathbf{y}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{y}) \int p(\mathbf{y}\,|\,\mathbf{t}) \log p(\mathbf{y}\,|\,\mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} - \int p(\mathbf{y}) \log p(\mathbf{y}) \, \mathrm{d}\mathbf{y} \\
&\overset{\text{(var. approx)}}{\geq} \int p(\mathbf{y}) \int p(\mathbf{y}\,|\,\mathbf{t}) \log q(\mathbf{y}\,|\,\mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} + H(Y) \\
&\geq \int p(\mathbf{t},\mathbf{y}) \log q(\mathbf{y}\,|\,\mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}
\end{aligned}
$$

Variational Bound - $I(T;Y)$ Lower Bound

$$I(T;Y) \geq \int p(\mathbf{t},\mathbf{y}) \log q(\mathbf{y} \mid \mathbf{t}) \,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}$$

Variational Bound - $I(T;Y)$ Lower Bound

$$\begin{aligned} I(T;Y) &\geq \int p(\mathbf{t},\mathbf{y}) \log q(\mathbf{y} \mid \mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\ &= \int p(\mathbf{t},\mathbf{y},\mathbf{x}) \log q(\mathbf{y} \mid \mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \end{aligned}$$

Variational Bound - $I(T;Y)$ Lower Bound

$$
\begin{aligned}
I(T;Y) &\geq \int p(\mathbf{t},\mathbf{y}) \log q(\mathbf{y} \mid \mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y},\mathbf{x}) \log q(\mathbf{y} \mid \mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&= \int p(\mathbf{x}) p(\mathbf{t} \mid \mathbf{x}) p(\mathbf{y} \mid \mathbf{x}) \log q(\mathbf{y} \mid \mathbf{t}) \, \mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x}
\end{aligned}
$$

Variational Bound - $I(T; Y)$ Lower Bound

$$
\begin{aligned}
I(T;Y) &\geq \int p(\mathbf{t},\mathbf{y}) \log q(\mathbf{y} \,|\, \mathbf{t}) \,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y},\mathbf{x}) \log q(\mathbf{y} \,|\, \mathbf{t}) \,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&= \int p(\mathbf{x}) p(\mathbf{t} \,|\, \mathbf{x}) p(\mathbf{y} \,|\, \mathbf{x}) \log q(\mathbf{y} \,|\, \mathbf{t}) \,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&\approx \frac{1}{N} \sum_{n=1}^{N} \int p(\mathbf{t} \,|\, \mathbf{x}^{(n)}) \int p(\mathbf{y} \,|\, \mathbf{x}^{(n)}) \log q(\mathbf{y} \,|\, \mathbf{t}) \,\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{t}
\end{aligned}
$$

Variational Bound - $I(T;Y)$ Lower Bound

$$
\begin{aligned}
I(T;Y) &\geq \int p(\mathbf{t},\mathbf{y}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y},\mathbf{x}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&= \int p(\mathbf{x})p(\mathbf{t}\,|\,\mathbf{x})p(\mathbf{y}\,|\,\mathbf{x}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&\approx \frac{1}{N}\sum_{n=1}^{N} \int p(\mathbf{t}\,|\,\mathbf{x}^{(n)}) \int p(\mathbf{y}\,|\,\mathbf{x}^{(n)}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{t} \\
&\approx \frac{1}{NM} \sum_{\substack{n=1\ldots N \\ m=1\ldots M}} \int p(\mathbf{y}\,|\,\mathbf{x}^{(n)}) \log q(\mathbf{y}\,|\,\mathbf{t}^{(m;n)})\,\mathrm{d}\mathbf{y}
\end{aligned}
$$

Variational Bound - $I(T;Y)$ Lower Bound

$$
\begin{aligned}
I(T;Y) &\geq \int p(\mathbf{t},\mathbf{y}) \log q(\mathbf{y} \mid \mathbf{t}) \,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y},\mathbf{x}) \log q(\mathbf{y} \mid \mathbf{t}) \,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&= \int p(\mathbf{x}) p(\mathbf{t} \mid \mathbf{x}) p(\mathbf{y} \mid \mathbf{x}) \log q(\mathbf{y} \mid \mathbf{t}) \,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&\approx \frac{1}{N} \sum_{n=1}^{N} \int p(\mathbf{t} \mid \mathbf{x}^{(n)}) \int p(\mathbf{y} \mid \mathbf{x}^{(n)}) \log q(\mathbf{y} \mid \mathbf{t}) \,\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{t} \\
&\approx \frac{1}{NM} \sum_{\substack{n=1\ldots N \\ m=1\ldots M}} \int p(\mathbf{y} \mid \mathbf{x}^{(n)}) \log q(\mathbf{y} \mid \mathbf{t}^{(m;n)}) \,\mathrm{d}\mathbf{y} \\
&\approx -\frac{1}{NM} \sum_{\substack{n=1\ldots N \\ m=1\ldots M}} H\left(b(\mathbf{x}^{(n)}), f_{\mathsf{apx}}(\mathbf{t}^{(m;n)})\right)
\end{aligned}
$$

Variational Bound - $I(T;Y)$ Lower Bound

$$
\begin{aligned}
I(T;Y) &\geq \int p(\mathbf{t},\mathbf{y}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y} \\
&= \int p(\mathbf{t},\mathbf{y},\mathbf{x}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&= \int p(\mathbf{x})p(\mathbf{t}\,|\,\mathbf{x})p(\mathbf{y}\,|\,\mathbf{x}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{t}\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \\
&\approx \frac{1}{N}\sum_{n=1}^{N} \int p(\mathbf{t}\,|\,\mathbf{x}^{(n)}) \int p(\mathbf{y}\,|\,\mathbf{x}^{(n)}) \log q(\mathbf{y}\,|\,\mathbf{t})\,\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{t} \\
&\approx \frac{1}{NM} \sum_{\substack{n=1\ldots N\\ m=1\ldots M}} \int p(\mathbf{y}\,|\,\mathbf{x}^{(n)}) \log q(\mathbf{y}\,|\,\mathbf{t}^{(m;n)})\,\mathrm{d}\mathbf{y} \\
&\approx -\frac{1}{NM} \sum_{\substack{n=1\ldots N\\ m=1\ldots M}} H\left(b(\mathbf{x}^{(n)}), f_{\mathsf{apx}}(\mathbf{t}^{(m;n)})\right) \qquad =: L_{T,Y}
\end{aligned}
$$

Variational Bound - Combined

$$I(T;Y) - \beta I(X;T) \gtrapprox L_{T,Y} - \beta U_{X,T}$$

$$\max_{\boldsymbol{\theta}} \quad L_{T,Y} - \beta U_{X,T}$$

Variational Bound - Combined

$$I(T;Y) - \beta I(X;T) \gtrapprox L_{T,Y} - \beta U_{X,T}$$

$$\begin{aligned} & \max_{\boldsymbol{\theta}} \quad L_{T,Y} - \beta U_{X,T} \\ \iff & \min_{\boldsymbol{\theta}} -L_{T,Y} + \beta U_{X,T} \end{aligned}$$

Variational Bound - Combined

$$I(T;Y) - \beta I(X;T) \gtrapprox L_{T,Y} - \beta U_{X,T}$$

$$\begin{aligned} & \max_{\boldsymbol{\theta}} \quad L_{T,Y} - \beta U_{X,T} \\ \iff & \min_{\boldsymbol{\theta}} -L_{T,Y} + \beta U_{X,T} \quad =: J_{\text{IB}} \end{aligned}$$

Variational Bound - Combined

$$I(T;Y) - \beta I(X;T) \gtrapprox L_{T,Y} - \beta U_{X,T}$$

$$\begin{aligned}
& \max_{\boldsymbol{\theta}} \quad L_{T,Y} - \beta U_{X,T} \\
\iff & \min_{\boldsymbol{\theta}} -L_{T,Y} + \beta U_{X,T} \qquad =: J_{\mathsf{IB}} \\
\iff & \min_{\boldsymbol{\theta}} \frac{1}{NM} \sum_{\substack{n=1\ldots N \\ m=1\ldots M}} H\Big(b(\mathbf{x}^{(n)}), f_{\mathsf{apx}}(\mathbf{t}^{(m;n)})\Big) \\
& \qquad + \beta \frac{1}{N} \sum_{n=1\ldots N} D_{\mathsf{KL}}\Big(f_{\mathsf{xpl}}(\mathbf{x}^{(n)}) \,||\, r\Big)
\end{aligned}$$

Variational Bound - Combined

$$I(T;Y) - \beta I(X;T) \gtrapprox L_{T,Y} - \beta U_{X,T}$$

$$\begin{aligned}
&\max_{\boldsymbol{\theta}} \quad L_{T,Y} - \beta U_{X,T} \\
\iff &\min_{\boldsymbol{\theta}} -L_{T,Y} + \beta U_{X,T} \quad =: J_{\mathsf{IB}} \\
\iff &\min_{\boldsymbol{\theta}} \frac{1}{NM} \sum_{\substack{n=1\ldots N \\ m=1\ldots M}} H\Big(\underbrace{b(\mathbf{x}^{(n)})}_{p(\mathbf{y}\,|\,\mathbf{x})}, \underbrace{f_{\mathsf{apx}}(\mathbf{t}^{(m;n)})}_{q(\mathbf{y}\,|\,\mathbf{t})} \Big) \\
&\qquad + \beta \frac{1}{N} \sum_{n=1\ldots N} D_{\mathsf{KL}}\Big(\underbrace{f_{\mathsf{xpl}}(\mathbf{x}^{(n)})}_{p(\mathbf{z}\,|\,\mathbf{x})} \;||\; r \Big)
\end{aligned}$$

Gumbel Softmax

Sampling T

$$\mathbf{t} = \mathbf{x} \odot \mathbf{z}$$

$$p(\mathbf{t} \mid \mathbf{x}) = p(\mathbf{z} \mid \mathbf{x}) \stackrel{?}{=} f_{\mathsf{xpl}}(\mathbf{x})$$

Sampling T

$$\mathbf{t} = \mathbf{x} \odot \mathbf{z}$$

$$p(\mathbf{t} \mid \mathbf{x}) = p(\mathbf{z} \mid \mathbf{x}) \stackrel{?}{=} f_{\mathsf{xpl}}(\mathbf{x})$$

outputs of f_{xpl} network are unnormalized logits...

Sampling T

$$\mathbf{t} = \mathbf{x} \odot \mathbf{z}$$

$$p(\mathbf{t} \mid \mathbf{x}) = p(\mathbf{z} \mid \mathbf{x}) \stackrel{?}{=} f_{\text{xpl}}(\mathbf{x})$$

$$p(\mathbf{z}^* \mid \mathbf{x}) = \text{relaxed\_k\_hot}_\tau(f_{\text{xpl}}(\mathbf{x}))$$

outputs of f_{xpl} network are unnormalized logits...

Gumbel Softmax

$$\mathbf{g} = -\log(-\log(\boldsymbol{\varepsilon}))\,,\ \boldsymbol{\varepsilon} \sim \mathcal{U}[0,1]$$
$$\mathbf{c} = \mathsf{softmax}\left(\frac{\log(\mathbf{p}) + \mathbf{g}}{\tau}\right)$$

Gumbel Softmax

$$\mathbf{g} = -\log(-\log(\varepsilon)) \,, \; \varepsilon \sim \mathcal{U}[0,1]$$
$$\mathbf{c} = \text{softmax}\left(\frac{\log(\mathbf{p}) + \mathbf{g}}{\tau}\right)$$

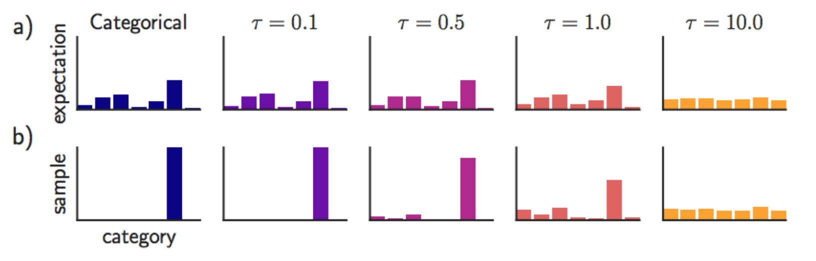


Figure 1: Relaxed categorical gumbel softmax distribution with varying temperature τ (Jang, Gu, and Poole, 2017).

Gumbel Softmax

$$\mathbf{g} = -\log(-\log(\boldsymbol{\varepsilon}))\,,\ \boldsymbol{\varepsilon} \sim \mathcal{U}[0,1]$$
$$\mathbf{c} = \mathsf{softmax}\left(\frac{\log(\mathbf{p}) + \mathbf{g}}{\tau}\right)$$

Gumbel Softmax

$$\mathbf{g} = -\log(-\log(\boldsymbol{\varepsilon}))\,,\ \boldsymbol{\varepsilon} \sim \mathcal{U}[0,1]$$
$$\mathbf{c} = \text{softmax}\left(\frac{\log(\mathbf{p}) + \mathbf{g}}{\tau}\right)$$

Sample k relaxed one-hot vectors $\{\mathbf{c}^{(n)}\}_{n=1}^{k}$.

$$\mathbf{z}_i^* = \max_n \mathbf{c}_i^{(n)}$$

Resulting vector z^* will be at most k-hot.

Results

Authors' - Results

Figure 2: Interpretable results from MNIST black box classifier (Bang et al., 2019).

Authors' - Results

	chunk size	k	L2X	VIBI (Ours)					
			0	0.001	0.01	0.1	1	10	100
Accuracy	1×1	64	0.694	0.690	0.726	0.689	0.742	0.729	**0.766**
	1×1	96	0.814	0.831	0.780	0.806	**0.859**	0.765	0.826
	1×1	160	0.903	0.907	0.905	0.917	0.917	**0.928**	0.902
	2×2	16	0.735	**0.795**	0.750	0.771	0.732	0.753	0.769
	2×2	24	0.776	0.855	0.834	0.856	**0.868**	0.854	0.847
	2×2	40	0.811	0.914	0.914	0.915	0.903	0.918	**0.935**
	2×2	80	0.905	0.949	0.940	0.939	**0.962**	0.941	0.923
	4×4	4	0.650	0.655	0.650	**0.775**	0.717	0.682	0.681
	4×4	6	0.511	**0.858**	0.706	0.701	0.708	0.690	0.730
	4×4	10	0.835	0.835	0.824	**0.933**	0.875	0.854	0.782
	4×4	20	0.954	**0.962**	0.815	0.934	0.929	0.946	0.943

Figure 3: Ablations for β parameter on MNIST (Bang et al., 2019).

Authors' - Results

	chunk size	k	Approximator Fidelity					Rationale Fidelity	
			Saliency	LIME	SHAP	L2X	VIBI (Ours)	L2X	VIBI (Ours)
IMDB	sentence	1	38.7 ± 0.9	72.7 ± 0.8	49.5 ± 1.0	87.6 ± 0.6	**87.7 ± 0.6**	72.7 ± 0.8	**73.1 ± 0.8**
	word	5	41.9 ± 0.9	**75.6 ± 0.8**	50.1 ± 1.0	73.8 ± 0.8	74.4 ± 0.8	63.8 ± 0.8	**65.7 ± 0.8**
	5 words	1	42.4 ± 0.9	29.0 ± 0.8	49.7 ± 1.0	75.9 ± 0.7	**76.4 ± 0.7**	60.1 ± 0.9	**63.2 ± 0.8**
	5 words	3	41.4 ± 0.9	67.9 ± 0.8	49.1 ± 1.0	83.3 ± 0.7	**83.5 ± 0.7**	**69.4 ± 0.8**	66.0 ± 0.8
MNIST	2×2	16	91.2 ± 0.6	77.0 ± 0.8	94.2 ± 0.5	93.4 ± 0.5	**94.8 ± 0.4**	73.5 ± 0.9	**77.1 ± 0.8**
	2×2	24	93.8 ± 0.5	80.7 ± 0.8	95.4 ± 0.4	95.1 ± 0.4	**95.3 ± 0.4**	77.6 ± 0.8	**85.6 ± 0.7**
	2×2	40	95.7 ± 0.4	85.9 ± 0.7	95.4 ± 0.4	**96.7 ± 0.4**	96.2 ± 0.4	81.1 ± 0.8	**91.5 ± 0.5**
	4×4	4	86.3 ± 0.7	60.9 ± 1.0	94.8 ± 0.4	**95.3 ± 0.4**	94.8 ± 0.4	65.0 ± 0.9	**77.5 ± 0.8**
	4×4	6	90.6 ± 0.6	63.7 ± 0.9	93.6 ± 0.5	**95.7 ± 0.4**	95.6 ± 0.4	51.1 ± 1.0	**70.1 ± 0.9**
	4×4	10	94.9 ± 0.4	70.5 ± 0.9	95.1 ± 0.4	96.5 ± 0.4	**96.7 ± 0.4**	83.5 ± 0.7	**93.3 ± 0.5**

Figure 4: Comparison with other interpretability frameworks (Bang et al., 2019).

CIFAR10 Results

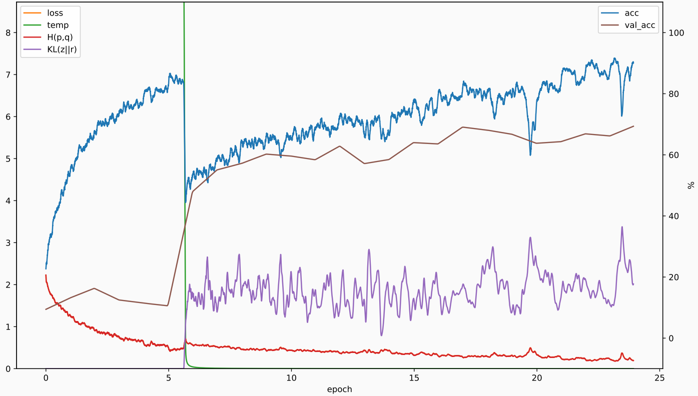
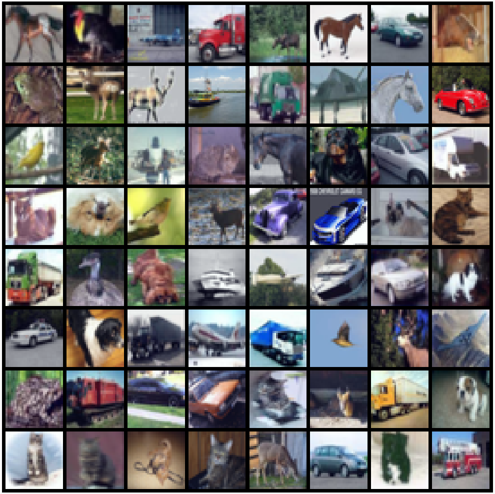


Figure 5: CIFAR10 VIBI training metrics. Explainer is a *Unet*, $k = 64$, β = 0.001

CIFAR10 Results

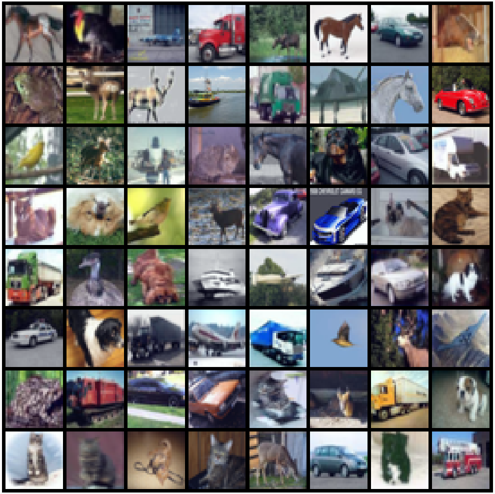


CIFAR10 test set batch.

CIFAR10 Results

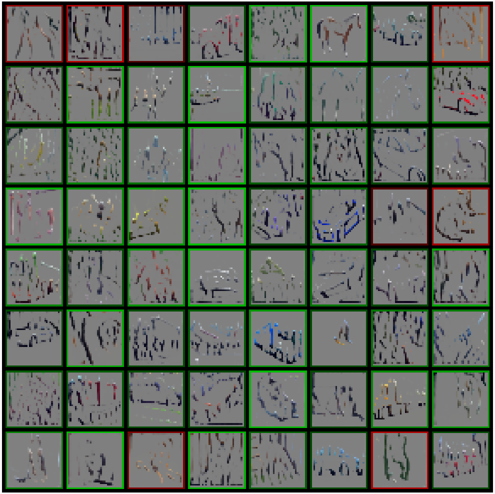
CIFAR10 VIBI **distribution** over explanation. Explainer is a *Unet*, $k = 64$, β = 0.001

CIFAR10 Results



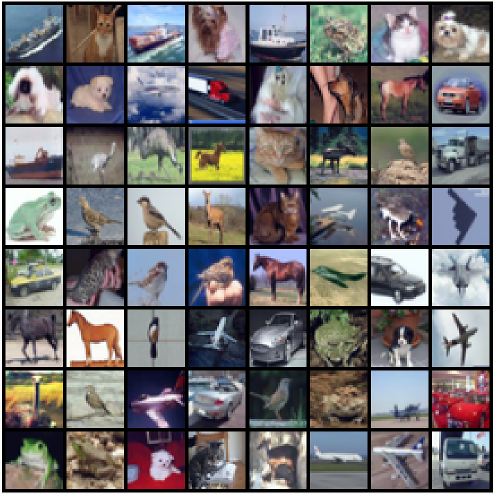
CIFAR10 test set batch.

CIFAR10 Results



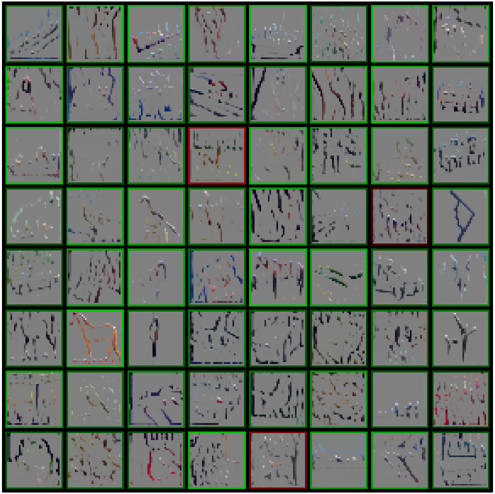
CIFAR10 VIBI explanations with black box prediction context. Explainer is a *Unet*, $k = 64$, β = 0.001

CIFAR10 More Results



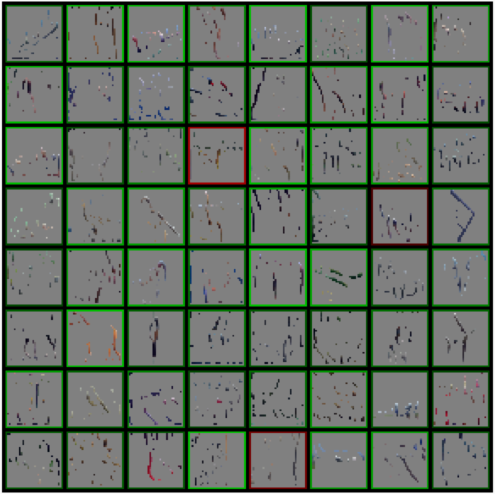
CIFAR10 test set batch.

CIFAR10 More Results



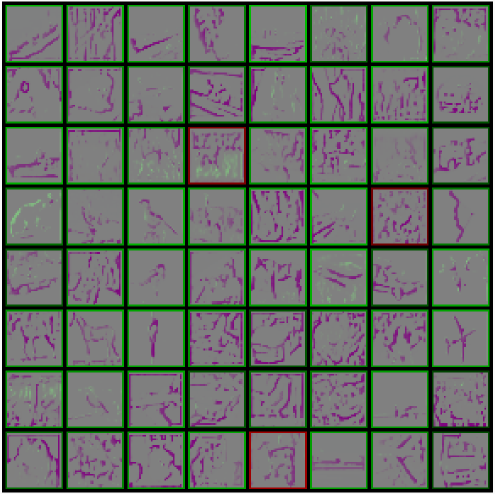
CIFAR10 VIBI **distribution** over explanation with black box prediction context.
Explainer is a *Unet*, $k = 64$, $\beta = 0.001$

CIFAR10 More Results



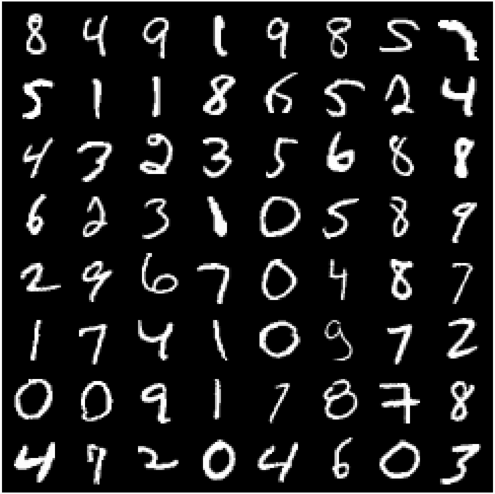
CIFAR10 VIBI **top-k** explanations with black box prediction context. Explainer is a ***Unet***, $k = 64$, $\beta = 0.001$

CIFAR10 More Results



CIFAR10 VIBI top-k explanations with black box prediction context. Explainer is a ***Unet***, $k = 64$, $\beta = 0.001$, ***out_channels*** $= 3$

MNIST Results

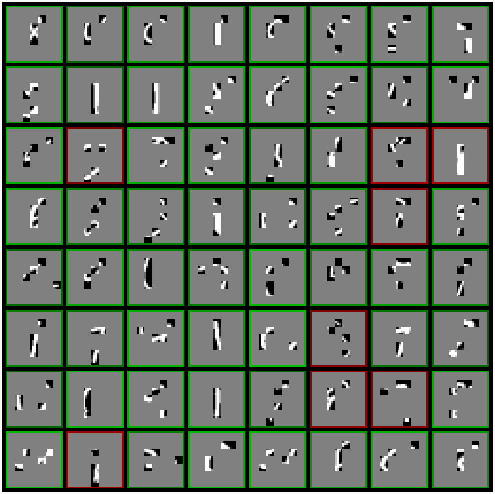


MNIST test set batch.

MNIST Results

MNIST VIBI **distribution** over explanations with black box prediction context.
Explainer is a ***ResNet***, $k = 4$, $\beta = 0.01$, ***chunk_size=4x4***

MNIST Results



MNIST VIBI **top-k** explanations with black box prediction context. Explainer is a *ResNet*, $k = 4$, $\beta = 0.01$, *chunk_size=4x4*

References

- Seojin Bang et al. *Explaining a black-box using Deep Variational Information Bottleneck Approach*. 2019. arXiv: *1902.06918 [cs.LG]*.
- Eric Jang, Shixiang Gu, and Ben Poole. *Categorical Reparameterization with Gumbel-Softmax*. 2017. arXiv: *1611.01144 [stat.ML]*.
- Naftali Tishby, Fernando C Pereira, and William Bialek. "The information bottleneck method". In: *arXiv preprint physics/0004057* (2000).

Code available at github.com/willisk/VIBI