Supplementary Material for "A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations"

A. Proof of Lemma 1

Proof: Saliency map includes only forward ReLUs without backward ReLUs, whereas DeconvNet includes only backward ReLUs without forward ReLUs. GBP has both types of ReLUs. Also, the norm of all the visualization results will be normalized to be in the range of $[0,1]$. Thus, by taking the (modified) derivative of $f_k(x)$ in Eq. (3) with respect to x and applying the proper normalization, these backpropagation-based visualizations for the k-th logit can be unified as

$$\begin{aligned} s_k(x) &\overset{(a)}{=} \frac{1}{Z_k}\sum_{i=1}^{N}\sum_{j=1}^{J} h(V_{q_{ij},k})\frac{\partial}{\partial x} g(w^{(i)T}y^{(j)}) \\ &\overset{(b)}{=} \frac{1}{Z_k}\sum_{j=1}^{J} {D_j}^T \sum_{i=1}^{N} h(V_{q_{ij},k})\frac{\partial}{\partial y^{(j)}} g(w^{(i)T}y^{(j)}) \qquad \text{(A.1)} \\ &\overset{(c)}{=} \frac{1}{Z_k}\sum_{j=1}^{J} {D_j}^T \sum_{i=1}^{N} h(V_{q_{ij},k})\tilde{w}^{(i,j)} \end{aligned}$$

where Z_k is the normalization coefficient to ensure $\|s_k(x)\| \in [0,1]$, (a) follows from the formal definitions of backpropagation-based visualization for a ReLU activation in Eq. (1) with $h(\cdot), g(\cdot)$ being given by Eq. (2), (b) is from applying $y^{(j)} = D_j x$ and swapping the two sums, and (c) is from taking the derivative of $g(\cdot)$ in the three cases with

$$\tilde{w}^{(i,j)} = \begin{cases} w^{(i)} & \text{for DeconvNet} \\ w^{(i)}\mathbb{I}\left(w^{(i)T}y^{(j)}\right) & \text{for saliency map and GBP} \end{cases}$$

as required. □

B. Proof of Theorem 1

Proof: In a random neural network where every entry of both V and W is assumed to be independently Gaussian distributed with a zero mean and variance c^2, we have $V_{q_{ij},k} \sim \mathcal{N}(0,c^2)$ and $w^{(i)} \sim \mathcal{N}(0,c^2 I)$ $\forall i \in \{1,\cdots,N\}, j \in \{1,\cdots,J\}$. For GBP, in order to ensure $\|s_k(x)\| \in [0,1]$ we first set $Z_k = \tilde{Z}_k N$. Assuming the number of filters N is sufficiently large (e.g. VGG-16 net usually has $N = 256$), then from Eq. (A.1) we have

$$\begin{aligned} s_k(x) &= \frac{1}{\tilde{Z}_k}\sum_{j=1}^{J} {D_j}^T \frac{1}{N}\sum_{i=1}^{N} h(V_{q_{ij},k})\tilde{w}^{(i,j)} \\ &\overset{(a)}{\approx} \frac{1}{\tilde{Z}_k}\sum_{j=1}^{J} {D_j}^T \mathbb{E}\left[h(V_{q_{ij},k})\tilde{w}^{(i,j)}\right] \\ &\overset{(b)}{=} \frac{1}{\tilde{Z}_k}\sum_{j=1}^{J} {D_j}^T \mathbb{E}\left[h(V_{q_{ij},k})\right]\mathbb{E}\left[\tilde{w}^{(i,j)}\right] \end{aligned}$$

where (a) follows from the asymptotic approximation of sample mean to the expectation and (b) follows from the fact that $V_{q_{ij},k}$ and $w^{(i)}$ are independent.

For GBP, we have $h(V_{q_{ij},k}) = \sigma(V_{q_{ij},k})$. Since we know $V_{q_{ij},k} \sim \mathcal{N}(0,c^2)$, then $h(V_{q_{ij},k})$ follows one-dimensional half normal distribution, and by its definition we can easily get $\mathbb{E}\left[h(V_{q_{ij},k})\right] = \sqrt{\frac{2}{\pi}}c$. Also, from the definition of $\tilde{w}^{(i,j)}$ for GBP, we know $\tilde{w}^{(i,j)}$ follows a p-dimensional half-normal distribution and its p.d.f. is

$$p(w) = \begin{cases} \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{w^T w}{2c^2}}, & \text{if } y^{(j)T}w > 0 \\ 0, & \text{otherwise} \end{cases} \qquad \text{(B.1)}$$

Then its expectation is given by

$$\begin{aligned} \mathbb{E}\left[\tilde{w}^{(i,j)}\right] &= \int_{y^{(j)T}w>0} w \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{w^T w}{2c^2}} dw \\ &\overset{(a)}{=} \int_{\phi_p>0} U\phi \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T \phi}{2c^2}} |U| d\phi \qquad \text{(B.2)} \\ &\overset{(b)}{=} U \int_{\phi_p>0} \phi \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T \phi}{2c^2}} d\phi \end{aligned}$$

where (a) follows from the change of variables $w = U\phi$ and U is an unitary matrix satisfying the condition that $U^T \cdot \frac{y^{(j)}}{\|y^{(j)}\|_2} = e^{(p)}$ and $e^{(p)}$ is an unit vector with only the p-th entry being 1. That is, $\frac{y^{(j)}}{\|y^{(j)}\|_2}$ is the p-th column of U. Thus, $y^{(j)T}w = y^{(j)T}U\phi = e^{(p)T}\phi\|y^{(j)}\|_2 = \phi_p\|y^{(j)}\|_2$ with ϕ_p being the p-th entry of ϕ, which means $y^{(j)T}w > 0$ is equivalent to $\phi_p > 0$. Also, by the change of variables in the integral, we have $dw = |U|d\phi$ where $|\cdot|$ denotes the determinant of a matrix. (b) follows from $|U| = 1$ by the definition of an unitary matrix, and the swap between matrix multiplication and the integral.

As ϕ is a p-dimensional vector, the integral above can be evaluated at each entry, denoted by ϕ_m, of ϕ separately. For $m \neq p$, we have

$$\begin{aligned}&\int_{\phi_p>0} \phi_m \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T\phi}{2c^2}} d\phi \\ &\overset{(a)}{=} 2 \underbrace{\int_{-\infty}^{\infty} \frac{\phi_m}{(2\pi c^2)^{\frac{1}{2}}} e^{-\frac{\phi_m^2}{2c^2}} d\phi_m}_{0} \cdot \int_{0}^{\infty} \frac{1}{(2\pi c^2)^{\frac{1}{2}}} e^{-\frac{\phi_p^2}{2c^2}} d\phi_p \\ &= 0\end{aligned}$$

where (a) follows from the expansion of the multiple integral, and all of the other $p-2$ integrals over ϕ_k for $k \notin \{p, m\}$ are 1. For $m = p$, we have

$$\begin{aligned}&\int_{\phi_p>0} \phi_p \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T\phi}{2c^2}} d\phi \\ &\overset{(a)}{=} 2 \int_{0}^{\infty} \phi_p \cdot \frac{1}{(2\pi c^2)^{\frac{1}{2}}} e^{-\frac{\phi_p^2}{2c^2}} d\phi_p \\ &\overset{(b)}{=} \sqrt{\frac{2}{\pi}} c\end{aligned}$$

where (a) also follows from the expansion of the multiple integral, and all other $p-1$ integrals over ϕ_k for $k \neq p$ are 1; (b) follows from evaluating the integral by the the change of variables $t = \frac{\phi_p^2}{2c^2}$. Putting them together, (B.2) becomes

$$\mathbb{E}\left[\tilde{w}^{(i,j)}\right] = \sqrt{\frac{2}{\pi}} c \cdot U e^{(p)} \overset{(a)}{=} \sqrt{\frac{2}{\pi}} c \cdot \frac{y^{(j)}}{\|y^{(j)}\|_2} \qquad \text{(B.3)}$$

where (a) follows from the the definition of the unitary matrix U satisfying $U^T \cdot \frac{y^{(j)}}{\|y^{(j)}\|_2} = e^{(p)}$.

Therefore, GBP at the k-th logit can be approximated as

$$\begin{aligned}s_k^{\text{GBP}}(x) &\approx \frac{2c^2}{\pi \tilde{Z}_k} \sum_{j=1}^{J} \frac{1}{\|y^{(j)}\|_2} {D_j}^T y^{(j)} \\ &\overset{(a)}{=} \frac{2c^2}{\pi \tilde{Z}_k} \left(\sum_{j=1}^{J} \frac{1}{\|y^{(j)}\|_2} \mathcal{I}_{p_j} \right) x\end{aligned}$$

where (a) follows from the definition

$$\mathcal{I}_{p_j} \triangleq {D_j}^T D_j = \begin{bmatrix} 0_{(j-1)b \times (j-1)b} & & \\ & I_{p \times p} & \\ & & 0 \end{bmatrix} \in \mathcal{R}^{d \times d}.$$

Ideally, if we assume $\|y^{(j)}\|_2 = C_0$, $\forall j$ (a constant) and ignore the boundary points (Note that using the "SAME" padding method instead of the "VALID" one is supposed to alleviate the boundary inconsistency to some extent), then $\sum_{j=1}^{J} \mathcal{I}_{p_j} \approx p I_{d \times d}$ and thus we can further approximate the GBP as

$$s_k^{\text{GBP}}(x) \approx \frac{2c^2 p}{\pi C_0 \tilde{Z}_k} x$$

Thus, by setting the normalization coefficient $\tilde{Z}_k = \frac{\pi C_0}{2c^2 p}$, we get the result. □

C. Proof of Theorem 2

In Eq. (4), we denote by $\Theta_j = \sum_{i=1}^{N} h(V_{q_{ij},k}) \tilde{w}^{(i,j)}$, which is a sum of N independent and identically distributed random variables. From the Central Limit Theorem, Θ_j is approximated as a Gaussian random variable if the number of filters N is sufficiently large. Since $s_k(x)$ is a linear function of Θ_j, i.e.

$$s_k(x) = \frac{1}{Z_k} \sum_{j=1}^{J} {D_j}^T \Theta_j \qquad \text{(C.1)}$$

we have $s_k(x)$ can also be approximated as a Gaussian random variable for both saliency map and DeconvNet.

In the first part of the proof, we evaluate the mean and variance of saliency map.

Since for saliency map we know $\mathbb{E}\left[h(V_{q_{ij},k})\right] = \mathbb{E}\left[V_{q_{ij},k}\right] = 0$, we can evaluate the mean of Θ_j as

$$\begin{aligned}\mathbb{E}\left[\Theta_j\right] &= \sum_{i=1}^{N} \mathbb{E}\left[h(V_{q_{ij},k}) \tilde{w}^{(i,j)}\right] \\ &\overset{(a)}{=} \sum_{i=1}^{N} \mathbb{E}\left[V_{q_{ij},k}\right] \mathbb{E}\left[\tilde{w}^{(i,j)}\right] \\ &= 0\end{aligned}$$

where (a) is from the fact that $V_{q_{ij},k}$ and $\tilde{w}^{(i,j)}$ are independent. Apparently, from Eq. (C.1) we have

$$\mathbb{E}\left[s_k^{\text{Sal}}(x)\right] = 0$$

Then to evaluate the variance of saliency map, we can also first evaluate the variance of Θ_j as

$$\begin{aligned}&\text{Var}\left[\Theta_j\right] = N \cdot \text{Var}\left[h(V_{q_{ij},k}) \tilde{w}^{(i,j)}\right] \\ &\overset{(a)}{=} N \cdot \Big\{ \text{Var}\left[V_{q_{ij},k}\right] \mathbb{E}\left[\tilde{w}^{(i,j)}\right]^2 + \\ &\qquad \text{Var}\left[V_{q_{ij},k}\right] \text{Var}\left[\tilde{w}^{(i,j)}\right] + \text{Var}\left[\tilde{w}^{(i,j)}\right] \mathbb{E}\left[V_{q_{ij},k}\right]^2 \Big\} \\ &\overset{(b)}{=} N \cdot c^2 \mathbb{E}\left[\tilde{w}^{(i,j)} \tilde{w}^{(i,j)T}\right]\end{aligned}$$

where (a) is also from the fact that $V_{q_{ij},k}$ and $\tilde{w}^{(i,j)}$ are independent and (b) follows from $\mathbb{E}\left[V_{q_{ij},k}\right] = 0$ and $\text{Var}\left[V_{q_{ij},k}\right] = c^2$. According to Eq. (B.1), we get

$$\begin{aligned}
\mathbb{E}\left[\tilde{w}^{(i,j)}\tilde{w}^{(i,j)T}\right] &= \int_{y^{(j)T}w>0} ww^T \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{w^Tw}{2c^2}} dw \\
&\overset{(a)}{=} \int_{\phi_p>0} U\phi\phi^T U^T \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T\phi}{2c^2}} |U| d\phi \\
&= U \underbrace{\left[\int_{\phi_p>0} \phi\phi^T \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T\phi}{2c^2}} d\phi\right]}_{A} U^T
\end{aligned}$$

where (a) also follows from the change of variables $w = U\phi$ and U is an unitary matrix satisfying the condition that $U^T \cdot \frac{y^{(j)}}{\|y^{(j)}\|_2} = e^{(p)}$.

Similarly, as $\phi\phi^T$ is a $p \times p$ matrix, the integral above can be evaluated at each entry, denoted by $\phi_m\phi_n$, of $\phi\phi^T$ separately where $m, n \in \{1, \cdots, p\}$.

First, for $m \neq n \neq p$, we have

$$\begin{aligned}
A_{mn} &= \int_{\phi_p>0} \phi_m\phi_n \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T\phi}{2c^2}} d\phi \\
&\overset{(a)}{=} \underbrace{\int_{-\infty}^{\infty} \frac{\phi_m}{(2\pi c^2)^{\frac{1}{2}}} e^{-\frac{\phi_m^2}{2c^2}} d\phi_m}_{0} \cdot \underbrace{\int_{-\infty}^{\infty} \frac{\phi_n}{(2\pi c^2)^{\frac{1}{2}}} e^{-\frac{\phi_n^2}{2c^2}} d\phi_n}_{0} \\
&= 0
\end{aligned}$$

where (a) follows from the expansion of the multiple integral, and all the other $p-2$ integrals over ϕ_k for $k \notin \{m, n\}$ are 1. Similarly, we can easily get that $A_{mn} = 0$ for $m \neq n$ with $i = p$ or $j = p$.

Also, for $m = n \neq p$, we have

$$\begin{aligned}
A_{mn} &= \int_{\phi_p>0} \phi_m^2 \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T\phi}{2c^2}} d\phi \\
&\overset{(a)}{=} 2 \underbrace{\int_{-\infty}^{\infty} \frac{\phi_m^2}{(2\pi c^2)^{\frac{1}{2}}} e^{-\frac{\phi_m^2}{2c^2}} d\phi_m}_{c^2} \cdot \underbrace{\int_{0}^{\infty} \frac{1}{(2\pi c^2)^{\frac{1}{2}}} e^{-\frac{\phi_p^2}{2c^2}} d\phi_p}_{\frac{1}{2}} \\
&= c^2
\end{aligned}$$

where (a) follows from the expansion of the multiple integral, and all other $p-2$ integrals are 1.

Finally, for $m = n = p$, we have

$$\begin{aligned}
A_{mn} &= \int_{\phi_p>0} \phi_p^2 \cdot \frac{2}{(2\pi c^2)^{\frac{p}{2}}} e^{-\frac{\phi^T\phi}{2c^2}} d\phi \\
&\overset{(a)}{=} 2 \underbrace{\int_{0}^{\infty} \frac{\phi_p^2}{(2\pi c^2)^{\frac{1}{2}}} e^{-\frac{\phi_p^2}{2c^2}} d\phi_p}_{\frac{1}{2}c^2} \\
&= c^2
\end{aligned}$$

where (a) follows from the expansion of the multiple integral, and all other $p-1$ integrals are 1.

Putting them together, we get $A = c^2 I$ and thus $\mathbb{E}\left[\tilde{w}^{(i,j)}\tilde{w}^{(i,j)T}\right] = c^2 UU^T = c^2 I$ which futher implies $\text{Var}\left[\Theta_j\right] = Nc^4 I$. Accordingly, from Eq. (C.1) we have

$$\begin{aligned}
\text{Var}\left[s_k^{\text{sal}}(x)\right] &= \frac{1}{Z_k^2} \sum_{j=1}^{J} D_j^T \text{Var}[\Theta_j] D_j \\
&= \frac{Nc^4}{Z_k^2} \sum_{j=1}^{J} D_j^T D_j \\
&\overset{(a)}{\approx} \frac{Nc^4 p}{Z_k^2} I \\
&\overset{(b)}{=} I
\end{aligned}$$

where (a) is from the approximation that the patching matrix D_j satisfies $\sum_{j=1}^{J} D_j^T D_j \approx pI$, (b) follows from setting the normalization coefficient to be $Z_k = c^2\sqrt{Np}$. Therefore, we have

$$s_k^{\text{Sal}} \sim \mathcal{N}(0, I)$$

In the second part of the proof, we evaluate the mean and variance of DeconvNet.

Similarly, for DeconvNet we have $\mathbb{E}\left[\tilde{w}^{(i,j)}\right] = \mathbb{E}\left[w^{(i)}\right] = 0$, then we can evaluate the mean of Θ_j as

$$\begin{aligned}
\mathbb{E}\left[\Theta_j\right] &= \sum_{i=1}^{N} \mathbb{E}\left[h(V_{q_{ij},k})\tilde{w}^{(i,j)}\right] \\
&\overset{(a)}{=} \sum_{i=1}^{N} \mathbb{E}\left[\sigma(V_{q_{ij},k})\right] \mathbb{E}\left[w^{(i)}\right] \\
&= 0
\end{aligned}$$

where (a) is from the fact that $V_{q_{ij},k}$ and $\tilde{w}^{(i,j)}$ are independent. Apparently, from Eq. (C.1) we have

$$\mathbb{E}\left[s_k^{\text{Deconv}}(x)\right] = 0$$

Then to evaluate the variance of DeconvNet, we can also first evaluate the variance of Θ_j as

$$\begin{aligned}
&\mathrm{Var}\left[\Theta_j\right] = N \cdot \mathrm{Var}\left[\sigma(V_{q_{ij},k}) w^{(i)}\right] \\
&\overset{(a)}{=} N \cdot \Big\{ \mathrm{Var}\left[\sigma(V_{q_{ij},k})\right] \mathrm{Var}\left[w^{(i)}\right] + \\
&\quad \mathrm{Var}\left[\sigma(V_{q_{ij},k})\right] \mathbb{E}\left[w^{(i)}\right]^2 + \mathrm{Var}\left[w^{(i)}\right] \mathbb{E}\left[\sigma(V_{q_{ij},k})\right]^2 \Big\} \\
&\overset{(b)}{=} Nc^2 \mathbb{E}\left[\sigma(V_{q_{ij},k})^2\right] \cdot I \\
&\overset{(c)}{=} Nc^4 I
\end{aligned}$$

where (a) is also from the fact that $\sigma(V_{q_{ij},k})$ and $w^{(i)}$ are independent, (b) follows from $\mathbb{E}\left[V_{q_{ij},k}\right] = 0$ and $\mathrm{Var}\left[w^{(i)}\right] = c^2 I$ and (c) follows from the fact that $\mathbb{E}\left[\sigma(V_{q_{ij},k})^2\right] = c^2$. Then, the rest of the proof follows the same derivations with saliency map, which yields

$$s_k^{\text{Deconv}} \sim \mathcal{N}(0, I)$$

Thus, we finish our proof by showing that both saliency map and DeconvNet are standard Gaussians which preserve no input information. □

D. Proof of Proposition 1

First, let us focus on the GBP case. From Eq. (9), we know

$$\hat{V}^{(2)}_{\cdot,k} \triangleq \frac{\partial o^{(3)}}{\partial o^{(2)}} \cdots \sigma\left(\frac{\partial o^{(L-1)}}{\partial o^{(L-2)}} \sigma\left(\Gamma^{(L)}_k\right)\right) \qquad \text{(D.1)}$$

where $\hat{V}^{(2)}_{q_{ij},k} \in \mathbb{R}^{d_3 \times 1}$ as we know $\Gamma^{(l)} \in \mathbb{R}^{d_l \times d_{l+1}}$ in the l-th layer, and then $\hat{V}^{(1)}_{\cdot,k}$ for GBP becomes

$$\begin{aligned}
\hat{V}^{(1)}_{\cdot,k} &= \frac{\partial \sigma(\Gamma^{(2)T} o^{(1)})}{\partial o^{(1)}} \sigma(\hat{V}^{(2)}_{\cdot,k}) \\
&= \left[\Gamma^{(2)}_{\cdot,1} \mathbb{I}(\Gamma^{(2)T}_{\cdot,1} o^{(1)}) \quad \cdots \quad \Gamma^{(2)}_{\cdot,d_3} \mathbb{I}(\Gamma^{(2)T}_{\cdot,d_3} o^{(1)})\right] \sigma(\hat{V}^{(2)}_{\cdot,k}) \\
&= \sum_{t=1}^{d_3} \Gamma^{(2)}_{\cdot,t} \mathbb{I}(\Gamma^{(2)T}_{\cdot,t} o^{(1)}) \sigma(\hat{V}^{(2)}_{t,k})
\end{aligned}$$

which yields

$$\hat{V}^{(1)}_{q_{ij},k} = \sum_{t=1}^{d_3} \Gamma^{(2)}_{q_{ij},t} \mathbb{I}(\Gamma^{(2)T}_{\cdot,t} o^{(1)}) \sigma(\hat{V}^{(2)}_{t,k})$$

Since every entry of $\Gamma^{(2)}$ is *i.i.d.* Gaussian distributed with zero-mean, we have

$$\begin{aligned}
\mathbb{I}(\Gamma^{(2)T}_{\cdot,t} o^{(1)}) &= \mathbb{I}(\Gamma^{(2)}_{q_{ij},t} o^{(1)}_{q_{ij}} + \sum_{v \neq q_{ij}} \Gamma^{(2)}_{v,t} o^{(1)}_v) \\
&\overset{(a)}{\approx} \underbrace{\mathbb{I}(\sum_{v \neq q_{ij}} \Gamma^{(2)}_{v,t} o^{(1)}_v)}_{b_t}
\end{aligned}$$

where (a) follows from the assumption of the dimension d_2 is sufficiently high in the CNN, and thus the impact of $\Gamma^{(2)}_{q_{ij},t} o^{(1)}_{q_{ij}}$ can be ignored. Therefore, b_t is independent of $\Gamma^{(2)}_{q_{ij},t}$. Similarly, $\sigma(\hat{V}^{(2)}_{t,k})$ is also independent of $\Gamma^{(2)}_{q_{ij},t}$ under the same approximation.

As d_3 is also sufficiently large and $\hat{V}^{(1)}_{q_{ij},k}$ for GBP becomes

$$\hat{V}^{(1)}_{q_{ij},k} \approx \sum_{t=1}^{d_3} \Gamma^{(2)}_{q_{ij},t} b_t \sigma(\hat{V}^{(2)}_{t,k}) \qquad \text{(D.2)}$$

which approximately is a Gaussian random variable with zero mean due to the central limit theorem. Next, in order to show the independence of two Gaussian random variables, it is equivalent to show they are uncorrelated. Since for any $q' \neq q_{ij}$, we know

$$\begin{aligned}
&\mathbb{E}\left[\hat{V}^{(1)}_{q',k} \hat{V}^{(1)}_{q_{ij},k}\right] \\
&\approx \mathbb{E}\left[\sum_{t'=1}^{d_3} \Gamma^{(2)}_{q',t'} b_{t'} \sigma(\hat{V}^{(2)}_{t',k}) \sum_{t=1}^{d_3} \Gamma^{(2)}_{q_{ij},t} b_t \sigma(\hat{V}^{(2)}_{t,k})\right] \\
&\overset{(a)}{=} \sum_{t'=1}^{d_3} \sum_{t=1}^{d_3} \mathbb{E}\left[\Gamma^{(2)}_{q',t'} \Gamma^{(2)}_{q_{ij},t}\right] \mathbb{E}\left[b_{t'} b_t\right] \mathbb{E}\left[\sigma(\hat{V}^{(2)}_{t',k}) \sigma(\hat{V}^{(2)}_{t,k})\right] \\
&\overset{(b)}{=} 0
\end{aligned}$$

where (a) is from the mutual independence of b_t, $\Gamma^{(2)}_{q_{ij},t}$ and $\sigma(\hat{V}^{(2)}_{t,k})$, and (b) is from the independence of two *i.i.d.* zero-mean Gaussians $\Gamma^{(2)}_{q',t'}$ and $\Gamma^{(2)}_{q_{ij},t}$, by our assumption. Therefore, $\hat{V}^{(1)}_{q',k}$ and $\hat{V}^{(1)}_{q_{ij},k}$ are uncorrelated with each other for any $q' \neq q_{ij}$, as desired.

Second, we consider the saliency map and DeconvNet cases. As $\hat{V}^{(1)}_{q_{ij},k}$ for saliency map becomes

$$\hat{V}^{(1)}_{q_{ij},k} \approx \sum_{t=1}^{d_3} \Gamma^{(2)}_{q_{ij},t} b_t \hat{V}^{(2)}_{t,k} \qquad \text{(D.3)}$$

where $\hat{V}^{(2)}_{t,k}$ comes from the definition

$$\hat{V}^{(2)}_{\cdot,k} = \frac{\partial o^{(3)}}{\partial o^{(2)}} \cdots \frac{\partial o^{(L-1)}}{\partial o^{(L-2)}} \cdot \Gamma^{(L)}_k$$

which is also approximately independent of $\Gamma^{(2)}_{q_{ij},t}$ as before, and the other parameters are exactly the same with the GBP case, the independence approximation also holds for saliency map.

For DeconvNet, $\hat{V}^{(1)}_{q_{ij},k}$ becomes

$$\hat{V}^{(1)}_{q_{ij},k} \approx \sum_{t=1}^{d_3} \Gamma^{(2)}_{q_{ij},t} \hat{V}^{(2)}_{t,k} \qquad \text{(D.4)}$$

where $\hat{V}_{t,k}^{(2)}$ is defined identically as (D.1) for GBP. Since this is a special case of GBP, the analysis in the case trivially holds for DeconvNet as well. □

E. More Experiments on Random/Trained VGG-16 Net

We provide more results for backpropagation-based visualizations including saliency map, DeconvNet and GBP in both untrained (randomly initialized) and trained VGG-16 net. The input images – labeled as "dog", "panda", "forest" and "mastiff" – are randomly chosen from the ImageNet dataset. As we can see, all the results (Figures 1 - 8) are consistent with our previous empirical observations that GBP and DeconvNet are more visually compelling but less class-sensitive than saliency map.

F. Comparison Between GBP and Edge Detector

Here we compare the GBP visualization with a linear vertical edge detector, as shown in Figure 9. At the first glance, the GBP visualizations in a trained VGG-16 net are very similar to the results of an edge detector. In other words, GBP indeed pays much attention to the edge information like a Gabor filter. However, there exist subtle differences between GBP and linear edge detectors. As we can see, the linear vertical edge detector will highlight all the horizontal intensity changes, while GBP has the additional ability to filter out some background image patches.

G. More Experiments on Partly Trained VGG-16 Net

In this section, we provide more GBP visualizations by feeding more images to a partly trained VGG-16 net. Specifically, we consider two kinds of weights loading strategies for the VGG-16 net. The first one is to load trained weights **up to** a given layer as shown in Figure 10. The second one is to load trained weights for all the layers **except for** a given layer as shown in Figure 11. The results are consistent with our previous analysis: it is the trained weights in the convolutional layers rather than those in the dense layers that account for filtering out image patches. Also, earlier convolutional layers have a greater impact on the GBP visualization than later convolutional layers.

H. More Experiments on ResNet

Our theoretical analysis shows that for GBP it is the local connections in CNNs, together with the backward ReLU, that contribute to the clean-looking visualizations. Here we further investigate backpropagation-based visualizations on both randomly initialized (Figure 11) and trained (Figure 12) ResNet-50. In general, the results are very similar to those in the VGG-16 net. However, we do observe some additional grid-like textures here and we conjecture that this deterioration of visual quality is due to the skip connections, as we have shown earlier that network structure has a significant impact on the visualizations. We leave the rigorous analysis of this phenomenon for future work.

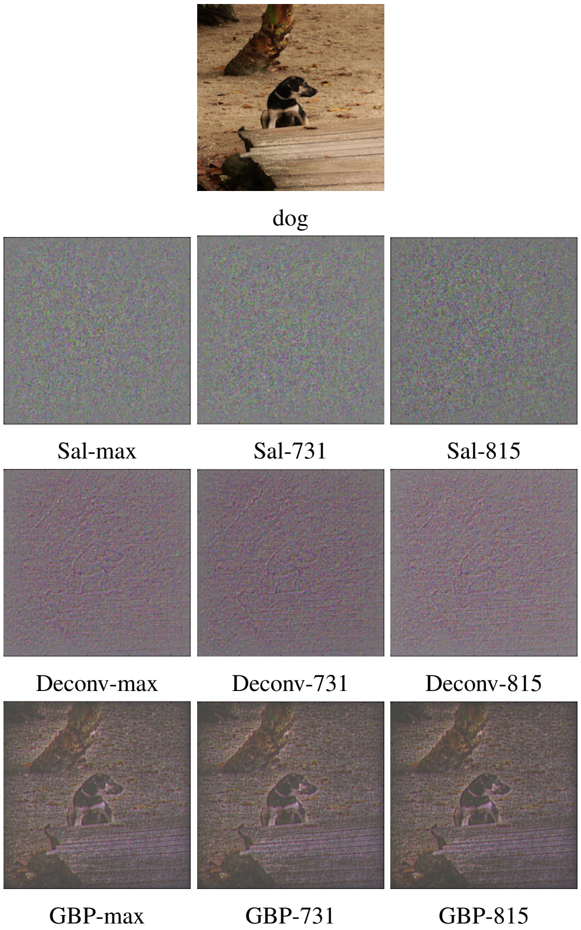


Figure 1. Saliency map, DeconvNet and GBP visualizations for the random VGG-16 net with the input image “dog”.

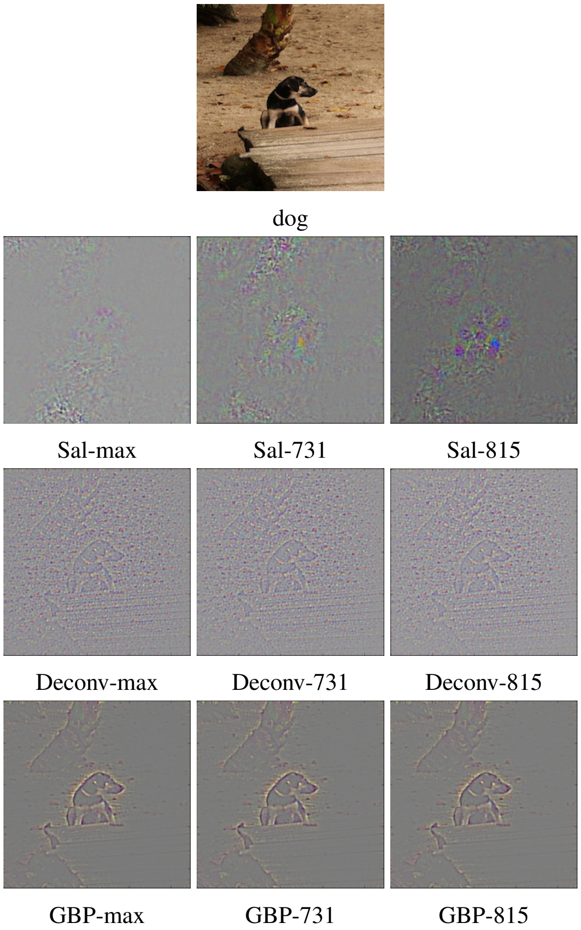


Figure 2. Saliency map, DeconvNet and GBP visualizations for the trained VGG-16 net with the input image “dog”.

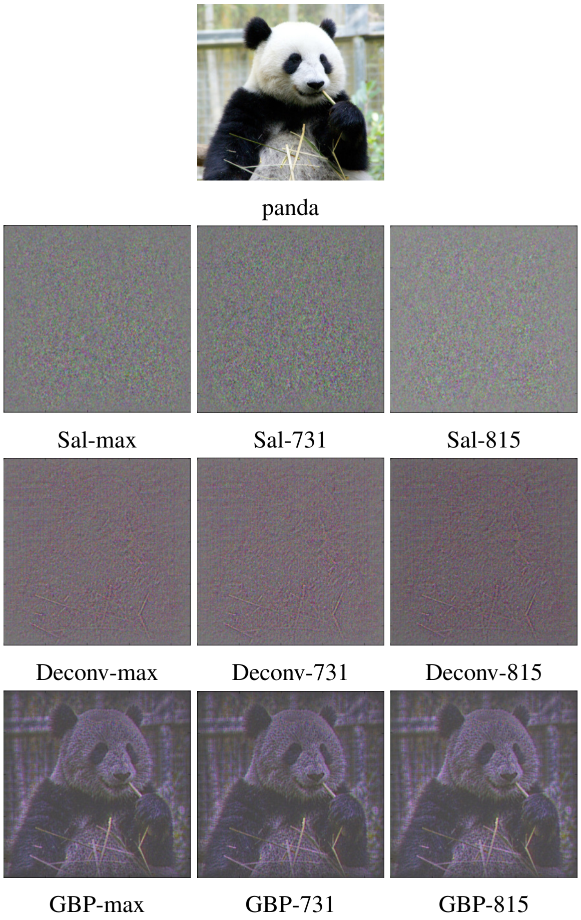


Figure 3. Saliency map, DeconvNet and GBP visualizations for the random VGG-16 net with the input image "panda".

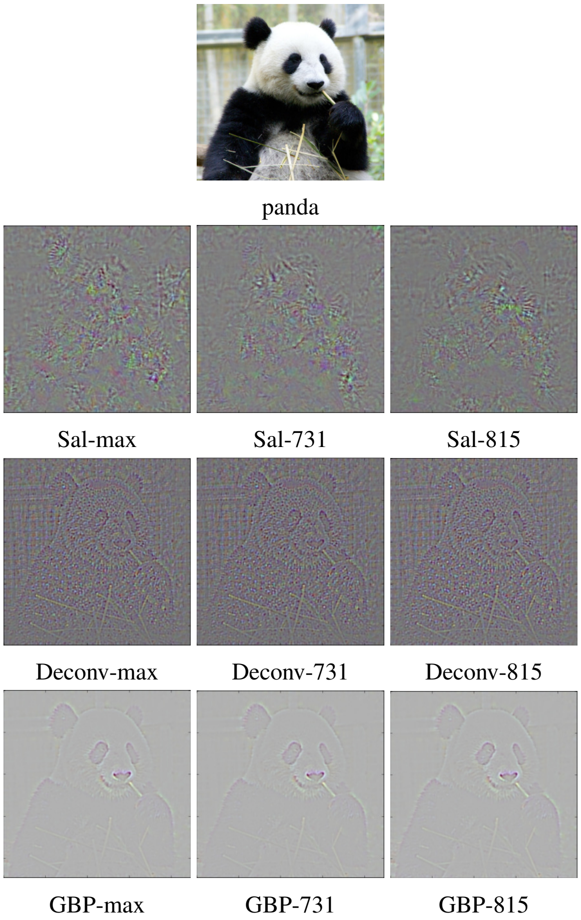
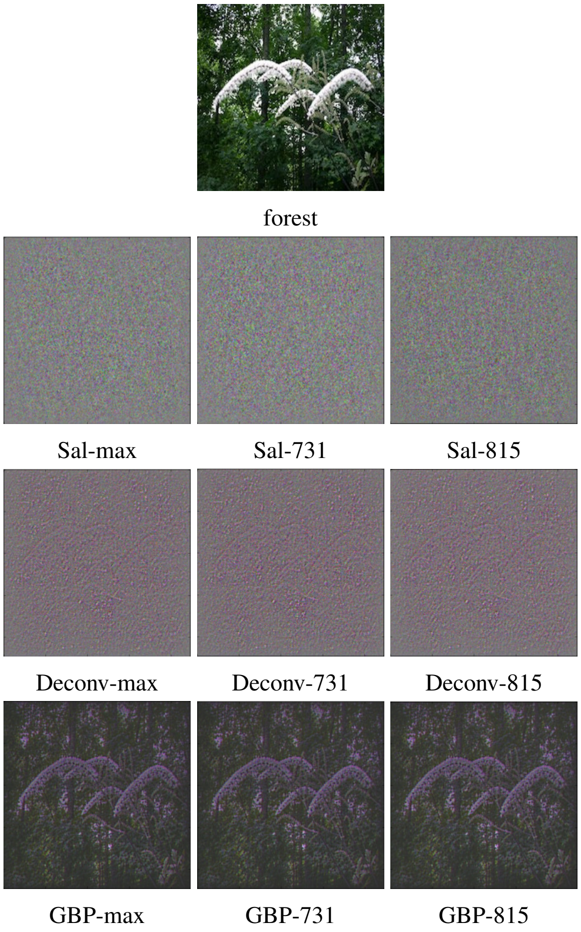


Figure 4. Saliency map, DeconvNet and GBP visualizations for the trained VGG-16 net with the input image "panda".



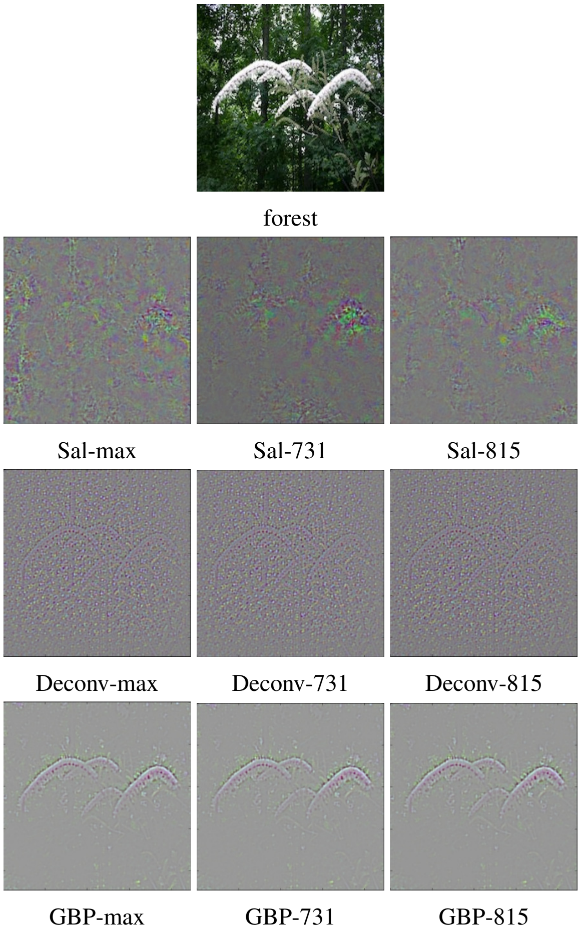
forest

Sal-max Sal-731 Sal-815

Deconv-max Deconv-731 Deconv-815

GBP-max GBP-731 GBP-815

Figure 5. Saliency map, DeconvNet and GBP visualizations for the random VGG-16 net with the input image "forest".



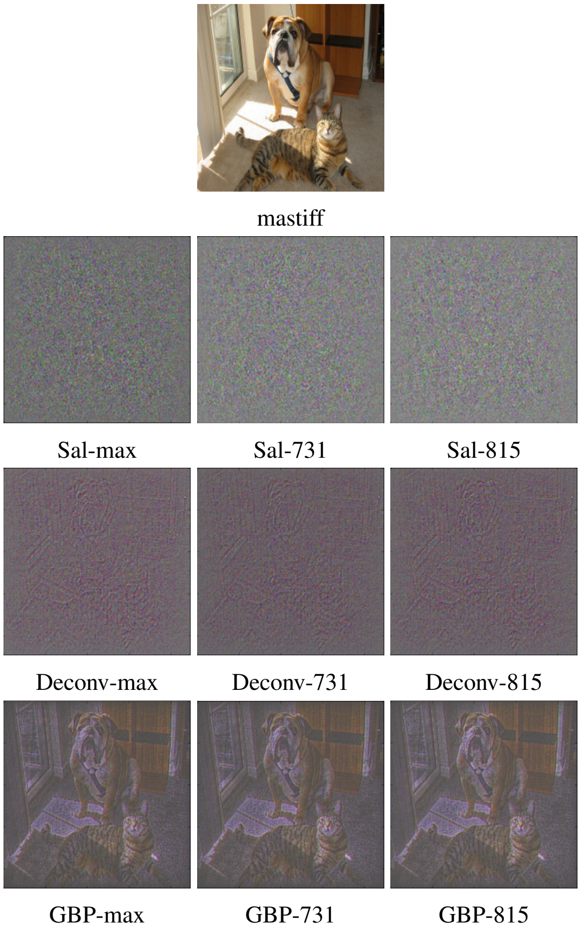
forest

Sal-max Sal-731 Sal-815

Deconv-max Deconv-731 Deconv-815

GBP-max GBP-731 GBP-815

Figure 6. Saliency map, DeconvNet and GBP visualizations for the trained VGG-16 net with the input image "forest".

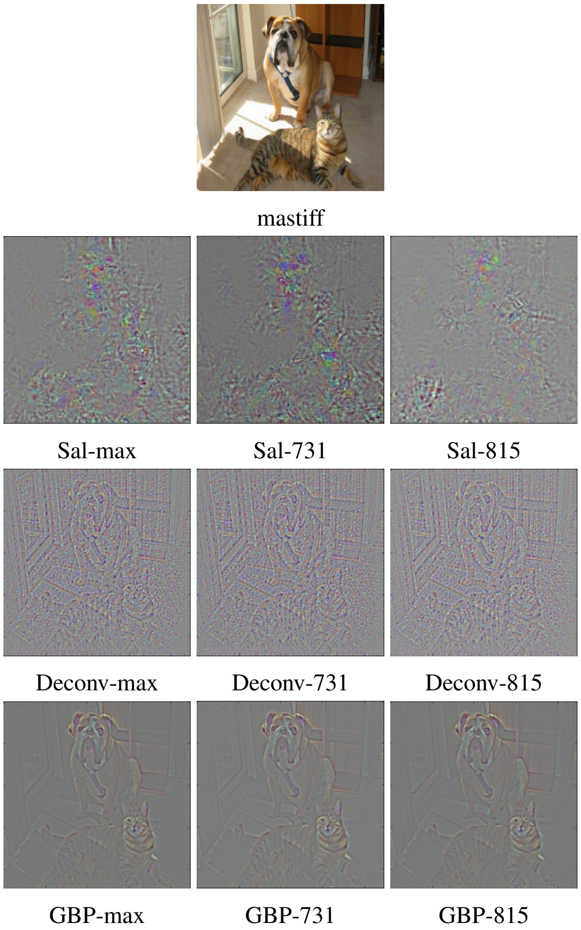

mastiff

Sal-max Sal-731 Sal-815

Deconv-max Deconv-731 Deconv-815

GBP-max GBP-731 GBP-815

Figure 7. Saliency map, DeconvNet and GBP visualizations for the random VGG-16 net with the input image “mastiff”.


mastiff

Sal-max Sal-731 Sal-815

Deconv-max Deconv-731 Deconv-815

GBP-max GBP-731 GBP-815

Figure 8. Saliency map, DeconvNet and GBP visualizations for the trained VGG-16 net with the input image “mastiff”.

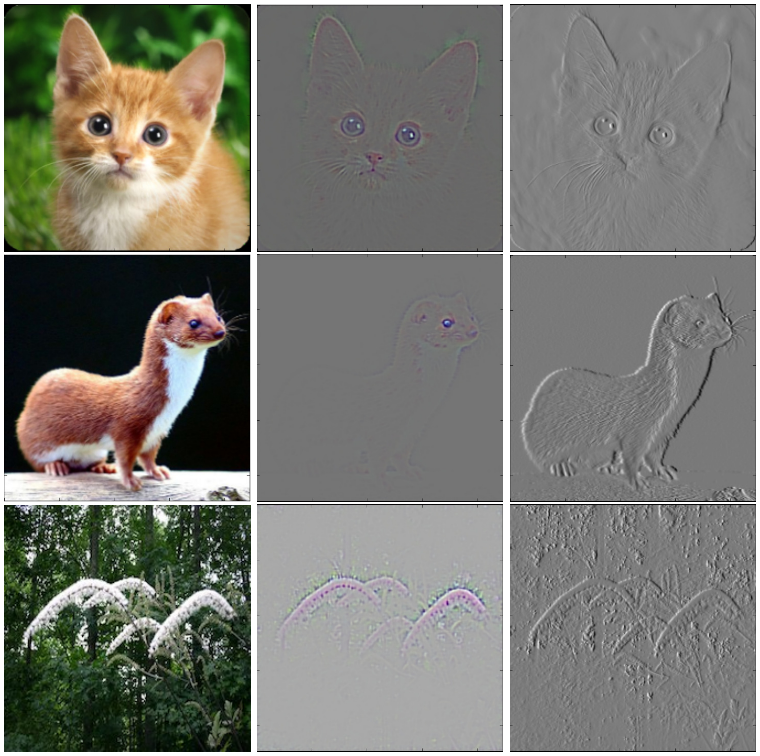


Figure 9. Comparison between the GBP visualization and the linear edge detector. The left column contains three sample inputs. The middle column contains the GBP visualization for each input. The right column is a linear vertical edge detector applied to each input. Specifically, the edge detector is designed by taking each pixel in the image and subtracting the neighboring pixel on the left.

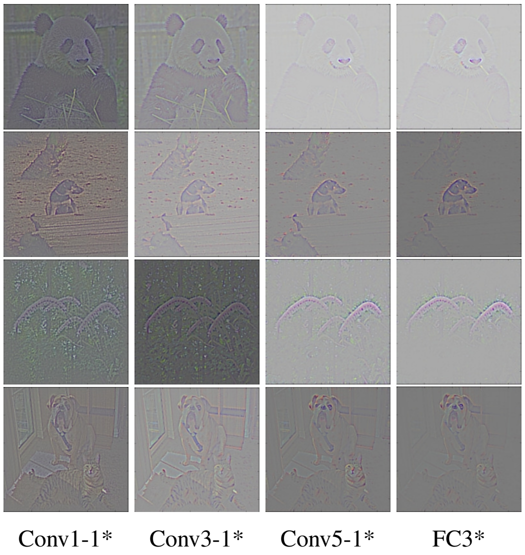


Figure 10. Load the trained weights of the VGG-16 net **up to** the indexed layer and leave the rest layers to be randomly initialized (denoted by the star sign) with different input images.

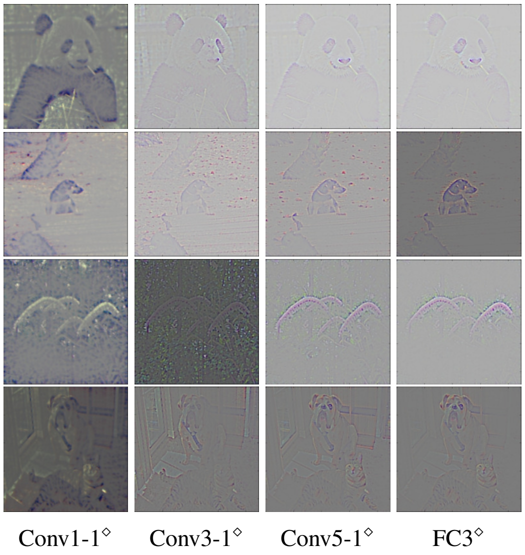


Figure 11. Load the trained weights of the VGG-16 net **except for** the indexed layer which is randomly initialized instead (denoted by the diamond sign) with different input images.

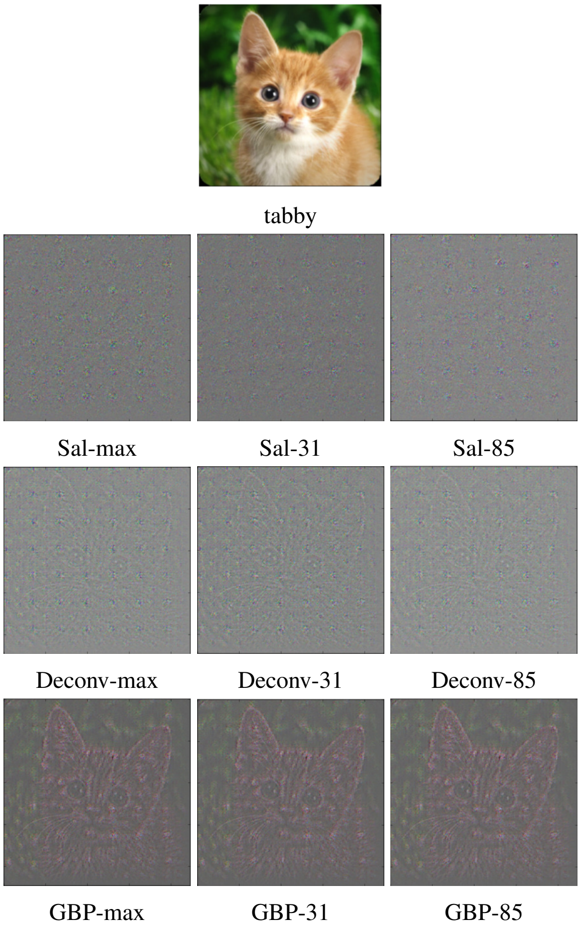


Figure 12. Saliency map, DeconvNet and GBP visualizations for the random ResNet-50 with the input image “tabby”.

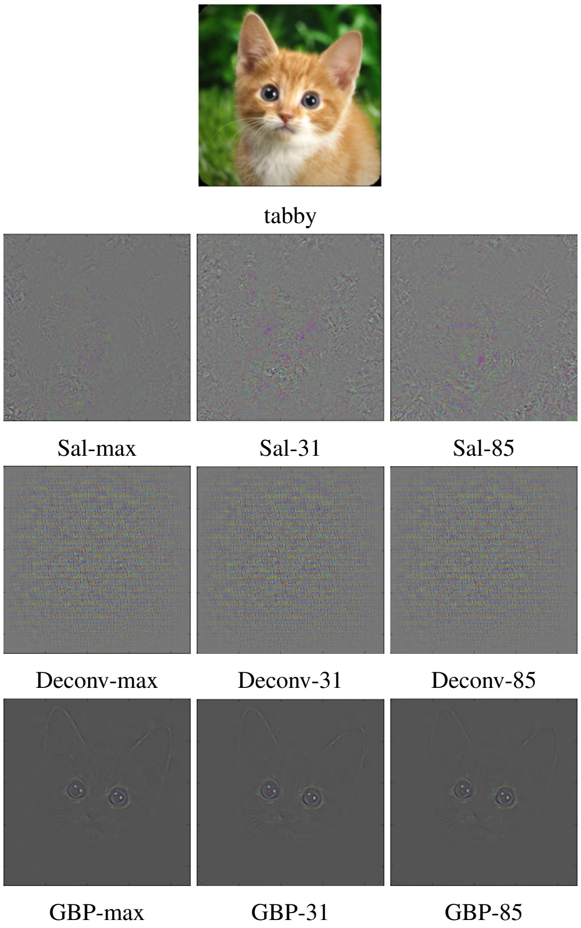


Figure 13. Saliency map, DeconvNet and GBP visualizations for the trained ResNet-50 net with the input image “tabby”.