SUPPLEMENTARY MATERIALS

A. DISCRIMINATIVE VS GENERATIVE LOG-LIKELIHOOD AND GRADIENT FOR BATCH TRAINING

During training, on a batch of training examples, $\{(x_i, y_i), i = 1, ..., n\}$, the generative log-likelihood is

$$l_{\mathrm{G}}(w) = \sum_i \log p(x_i|y_i, w) = \sum_i \log \frac{\exp\left(f_{y_i}(x_i; w)\right)}{Z_{y_i}(w)} \approx \sum_i \log \frac{\exp\left(f_{y_i}(x_i; w)\right)}{\sum_k \exp\left(f_{y_i}(x_k; w)\right)/n}.$$

The gradient with respect to w is

$$l'_{\mathrm{G}}(w) = \sum_i \left[\frac{\partial}{\partial w} f_{y_i}(x_i; w) - \sum_j \frac{\partial}{\partial w} f_{y_i}(x_j; w) \frac{\exp(f_{y_i}(x_j; w))}{\sum_k \exp(f_{y_i}(x_k; w))}\right].$$

The discriminative log-likelihood is

$$l_{\mathrm{D}}(w) = \sum_i \log p(y_i|x_i, w) = \sum_i \log \frac{\exp(f_{y_i}(x_i; w))}{\sum_y \exp(f_y(x_i; w))}.$$

The gradient with respect to w is

$$l'_{\mathrm{D}}(w) = \sum_i \left[\frac{\partial}{\partial w} f_{y_i}(x_i; w) - \sum_y \frac{\partial}{\partial w} f_y(x_i; w) \frac{\exp(f_y(x_i; w))}{\sum_y \exp(f_y(x_i; w))}\right].$$

l'_{D} and l'_{G} are similar in form and different in the summation operations. In l'_{D}, the summation is over category y while x_i is fixed, whereas in l'_{G}, the summation is over example x_j while y_i is fixed.

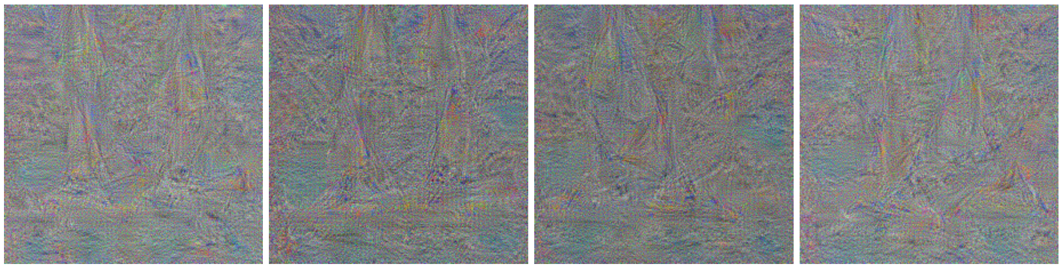
In the generative gradient, we want f_{y_i} to assign high score to x_i as well as those observations that belong to y_i, but assign low scores to those observations that do not belong to y_i. This constraint is for the *same* f_{y_i}, regardless of what other f_y do for $y \neq y_i$.

In the discriminative gradient, we want $f_y(x_i)$ to work together for all *different* y, so that f_{y_i} assigns high score to x_i than other f_y for $y \neq y_i$.

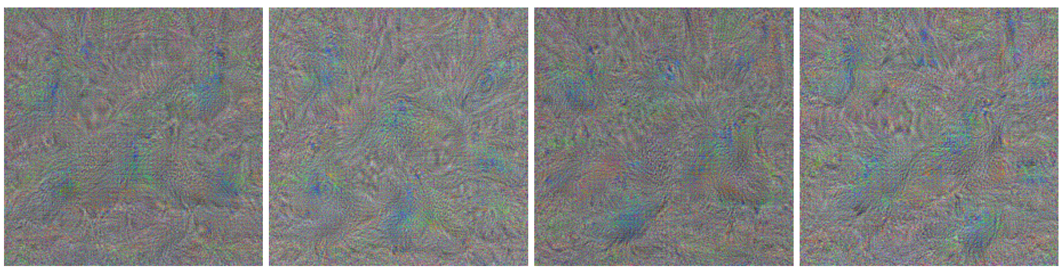
Apparently, the discriminative constraint is weaker because it involves all f_y, and the generative constraint is stronger because it involves single f_y. After generative learning, these f_y are well behaved and then we can continue to adjust them (probably the intercepts for different y) to satisfy the discriminative constraint.

B. MORE GENERATIVE VISUALIZATION EXAMPLES

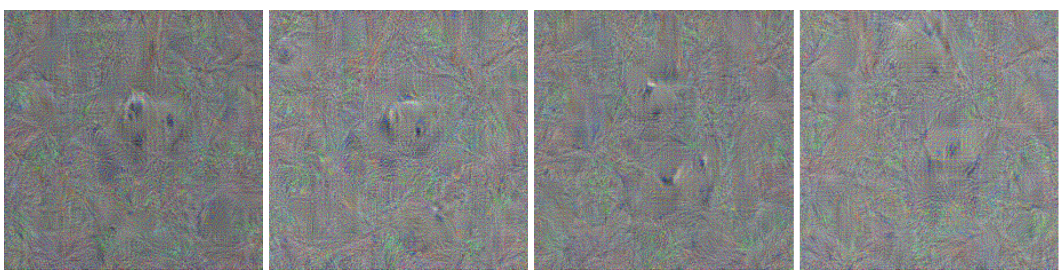
More generative visualization examples from the nodes at the final fully-connected layer in the fully trained AlexNet model are shown in Fig. B1, Fig. B2 and Fig. B3.



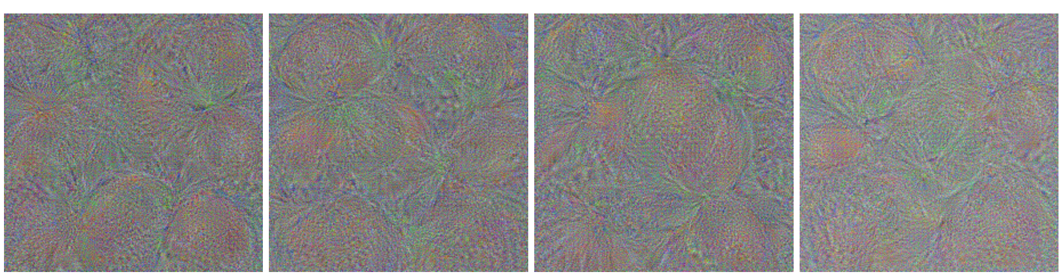
(a) catamaran



(b) Peacock

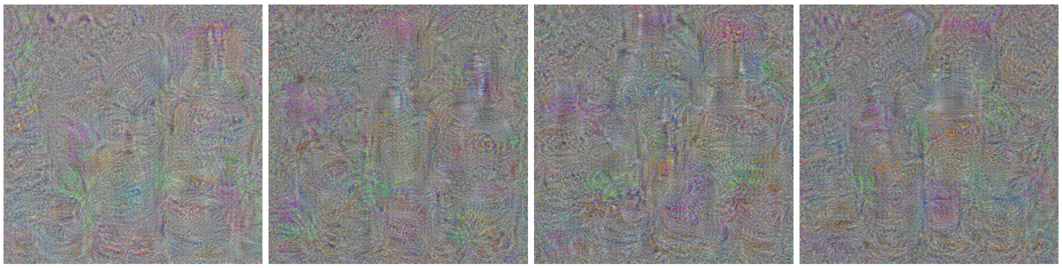


(c) Giant panda, panda, panda bear, coon bear, Ailuropoda melanoleuca

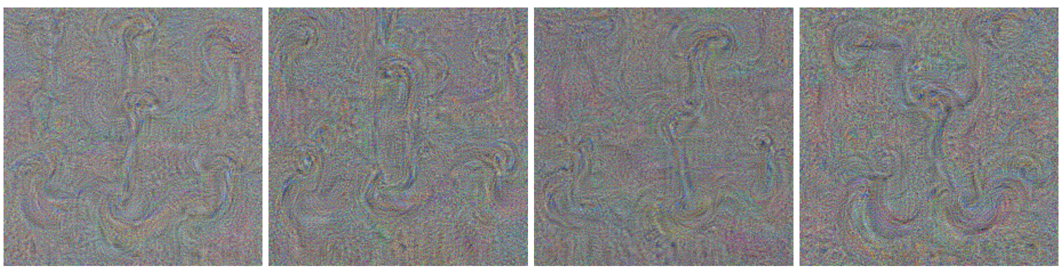


(d) Orange

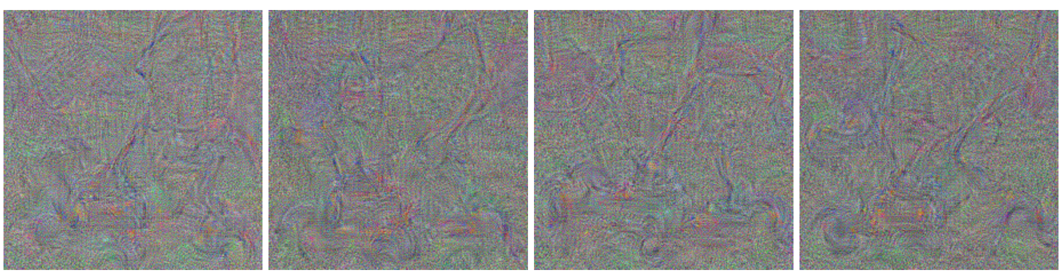
Figure B1: More samples from the nodes at the final fully-connected layer (fc8) in the fully trained AlexNet model, which correspond to different object categories (part 1).



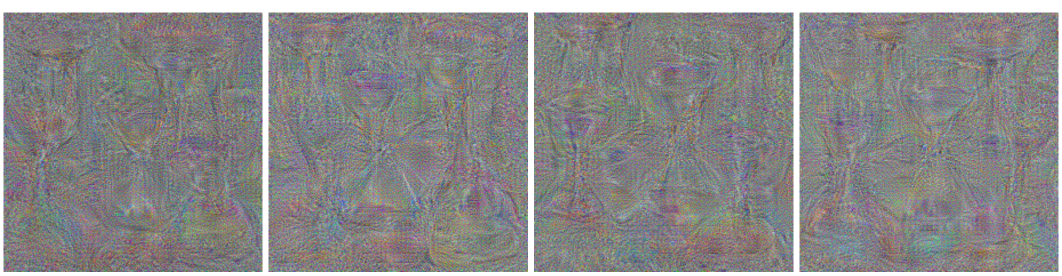
(a) Lotion



(b) Hook, claw

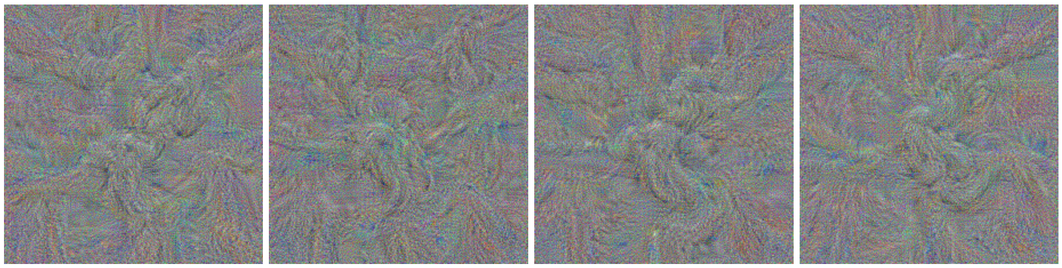


(c) Lawn mower, mower

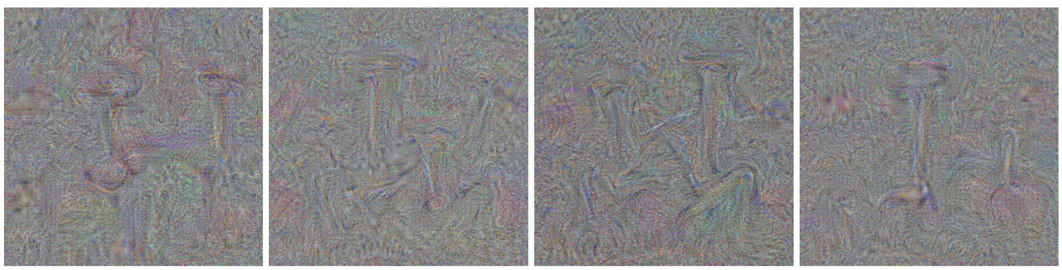


(d) Hourglass

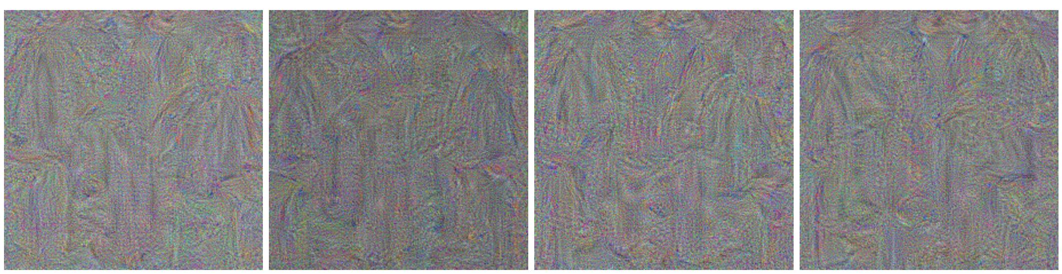
Figure B2: More samples from the nodes at the final fully-connected layer (fc8) in the fully trained AlexNet model, which correspond to different object categories (part 2).



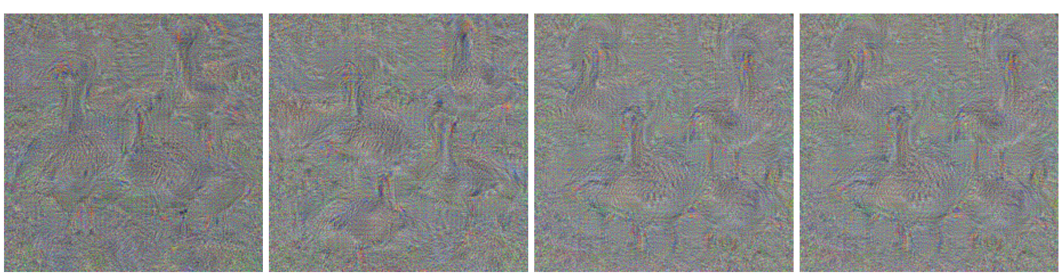
(a) Knot



(b) Nail



(c) academic gown, academic robe, judge's robe



(d) goose

Figure B3: More samples from the nodes at the final fully-connected layer (fc8) in the fully trained AlexNet model, which correspond to different object categories (part 3).