#3230

Sampling the “Inverse Set” of a Neuron

Suryabhan Singh Hada and **Miguel Á. Carreira-Perpiñán**,
Dept. CSE, UC Merced

1 Motivation and summary

- With the recent success of deep neural networks in computer vision, it is important to understand the internal working of these networks. What does a given neuron represent? The concepts captured by a neuron may be hard to understand or express in simple terms.

- We solve this by characterizing the region of input space that excites a given neuron to a certain level; we call this the inverse set.

- This inverse set is a complicated high dimensional object that we explore by an optimization-based sampling approach. Inspection of samples of this set by a human can reveal regularities that help to understand the neuron.

The inverse set of a neuron: definition

- We say an input **x** is in the inverse set of a given neuron having a real-valued activation function f if it satisfies the following two properties:

$$z_1 \le f(\mathbf{x}) \le z_2 \qquad \mathbf{x} \text{ is a valid input} \qquad (1)$$

  where z_1, $z_2 \in \mathbb{R}$ are activation values of the neuron.

- In general for deep neural networks we approximate the inverse set with a sample that covers it in a representative way.

2 Sampling the inverse set of a neuron

- To create a sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ that covers the inverse set, we transform eq. (1) into a constrained optimization problem:

$$\arg\max_{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n} \sum_{i,j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \quad \text{s.t.} \quad z_1 \le f(\mathbf{x}_1), \dots, f(\mathbf{x}_n) \le z_2.$$

- The objective function ensures that the samples are different from each other and satisfy eq. (1).

- It has two issues. The generated images are noisy and are very sensitive to small changes in their pixels.

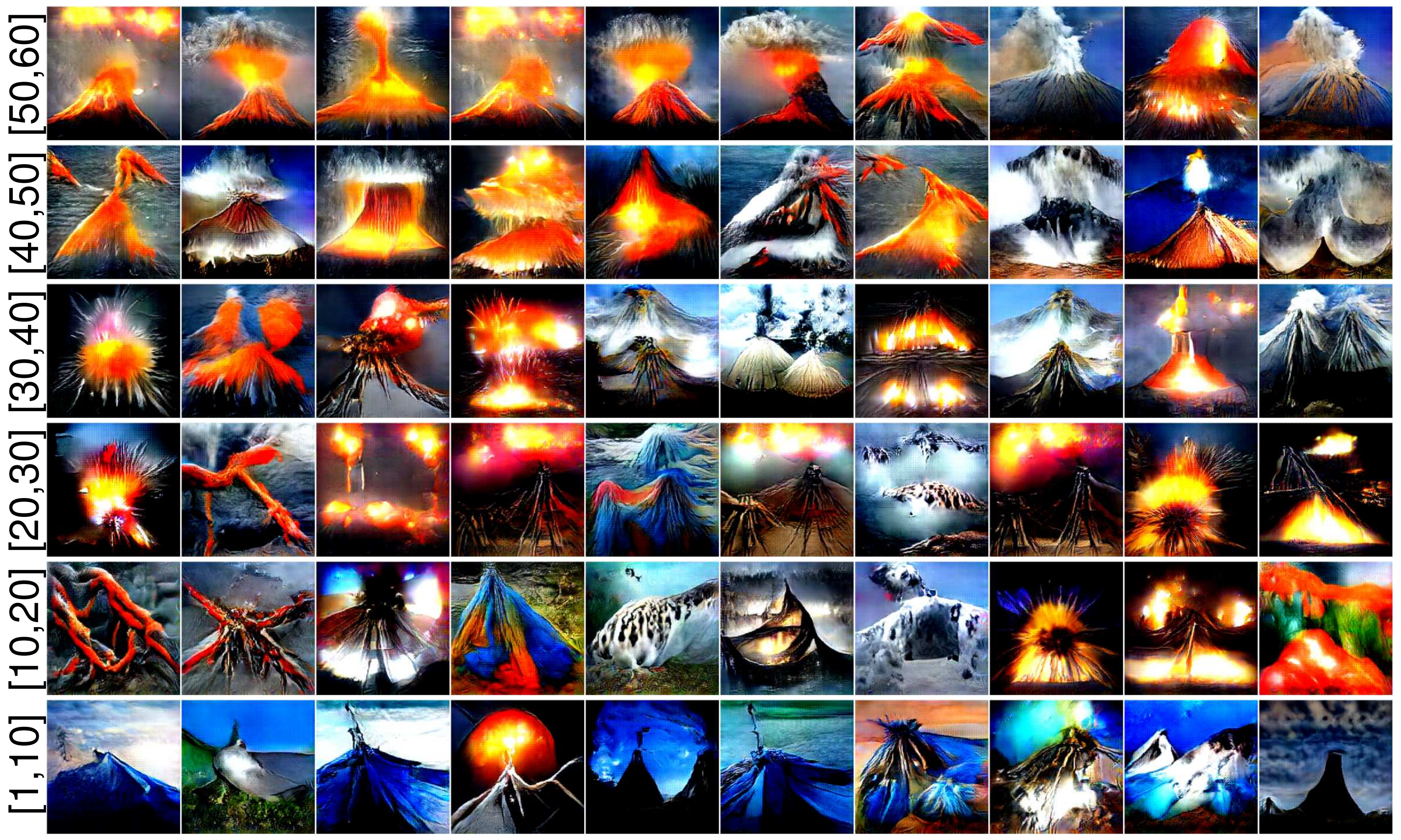
- To counter the first issue, we use generator network **G** to generate images from a code vector **c**. Next, for the second issue we compute distances on a low-dimensional encoding $\mathbf{E}(\mathbf{G}(\mathbf{c}))$ of the generated images constructed by an encoder **E**. This gives us our final formulation for generating n samples.

$$\arg\max_{\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_n} \sum_{i,j=1}^{n} \|\mathbf{E}(\mathbf{G}(\mathbf{c}_i)) - \mathbf{E}(\mathbf{G}(\mathbf{c}_j))\|_2^2$$
$$\text{s.t.} \quad z_1 \le f(\mathbf{G}(\mathbf{c}_1)), \dots, f(\mathbf{G}(\mathbf{c}_n)) \le z_2.$$

- It is computationally expensive to generate many samples due to the quadratic complexity of the objective function over the number of samples n. We use the following two approximations to make it faster.
  - We stop the optimization algorithm once the samples enter the feasible set, as, by that time, the samples are already separated.
  - We create the samples incrementally, K samples at a time (with $K \ll n$). We optimize the objective function for the first K samples, initializing the code vectors **c** with random values. These samples are then fixed. The next K samples use the objective plus their distances to the previous K samples. We initialize them with the previous K samples and take a single gradient step in the feasible region. The resultant samples are the new K samples.

3 Experiments

Six inverse sets with different activation range for the neuron # 981 in fc8 layer of the CaffeNet, which represents volcano class. Both the first and last row have volcanoes, but lava and smoke create a huge difference in the activation value of the neuron. The activation value of the neuron is proportional to the amount of lava and smoke.



Sampling the intersection of two inverse sets. The sample images from left to right are from the inverse set of neuron # 664 (class monastery), of neuron # 862 (class toilet seat), and of their intersection, all in the activation range [50,60]. Both neurons are from layer fc8 of CaffeNet.

neuron #664, [50,60]

neuron #862, [50,60]

Inverse set Intersection