Understanding the Effective Receptive Field in Deep Convolutional Neural Networks

Wenjie Luo*, Yujia Li*, Raquel Urtasun and Richard Zemel
{wenjie, yujiali, urtasun, zemel}@cs.toronto.edu (* indicates equal contribution)

Introduction

- We introduce the notion of an **effective receptive field (ERF)**.
- We prove that ERF has a **Gaussian distribution** using Fourier analysis and central limit theorem.
- We show that ERF grows $O(\sqrt{n})$ over number of layers n in deep CNNs and occupies $O(\frac{1}{\sqrt{n}})$ of the full theoretical receptive field.
- We analyze the ERF in several architecture designs, and the effect of nonlinear activations, dropout, sub-sampling and skip connections on it.
- We show that ERF grows during training.

Be careful, receptive field is smaller than we thought.

Convolution by Fourier Transform

We are showing: the distribution of gradients in a receptive field for an output unit in a deep CNN correspond to coefficients of a (extended) binomial distribution.

Considering convolution with uniform weights. Given input $u(t) = \delta(t)$ and convolution kernel:

$$v(t) = \sum_{m=0}^{k-1} \delta(t-m), \quad \text{where } \delta(t) = \begin{cases} 1, & t = 0 \\ 0, & t \neq 0 \end{cases}$$

Using Fourier transform:

$$U(\omega) = \sum_{t=-\infty}^{\infty} u(t)e^{-j\omega t} = 1, \quad V(\omega) = \sum_{t=-\infty}^{\infty} v(t)e^{-j\omega t} = \sum_{m=0}^{k-1} e^{-j\omega m}$$

Applying the convolution theorem, we have the Fourier transform of o to be:

$$\mathcal{F}(o) = \mathcal{F}(u * v * \cdots * v)(\omega) = U(\omega) \cdot V(\omega)^n = \left(\sum_{m=0}^{k-1} e^{-j\omega m}\right)^n \qquad (1)$$

Using inverse Fourier transform:

$$o(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(\sum_{m=0}^{k-1} e^{-j\omega m}\right)^n d\omega, \qquad \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-j\omega s}e^{j\omega t} d\omega = \begin{cases} 1, & s = t \\ 0, & s \neq t \end{cases}$$

We can see that $\mathbf{o(t)}$ is simply the **coefficient of $\mathbf{e^{-j\omega t}}$** in the expansion of $\left(\sum_{m=0}^{k-1} e^{-j\omega m}\right)^n$.

Case $k = 2$: $\left(\sum_{m=0}^{k-1} e^{-j\omega m}\right)^n = (1 + e^{-j\omega})^n$. The coefficient for $e^{-j\omega t}$ is then the standard binomial coefficient $\binom{n}{t}$, i.e. $o(t) = \binom{n}{t}$.

Case $k > 2$: Coefficients are known as "extended binomial coefficients" or "polynomial coefficients".

Effective Receptive Field (ERF)

Receptive Field of an output unit is the region containing any input pixel with an impact on that unit.

Effective Receptive Field (ERF) of an output unit is the region containing any input pixel with a *non-negligible* impact on that unit.

non-negligible: region of impact within 2-standard deviation of center pixel's impact.

For CNNs, we measure the impact as the scale of the partial derivatives, which can be computed by back-propagation, i.e. convolving gradient with weight, similar as Eq. 1:

$$\mathcal{F}(o) = U(\omega) \cdot V(\omega) \cdots V(\omega) = \left(\sum_{m=0}^{k-1} w(m)e^{-j\omega m}\right)^n$$

$o(t)$, the impact at pixel location t, is the coefficient of $e^{-j\omega t}$ in the above expansion.

Uniform weights: Impact corresponds to binomial coefficient for $k = 2$ or "extended binomial coefficients" for $k > 2$, both distribute like Gaussian.

Non-Uniform weights: combinatorial literature shows:

$$o(t) = p(S_n = t), S_n = \sum_{i=1}^{n} X_i$$

where X_i's are i.i.d. multinomial variables distributed according to $w(m)$'s, i.e. $p(X_i = m) = w(m)$.

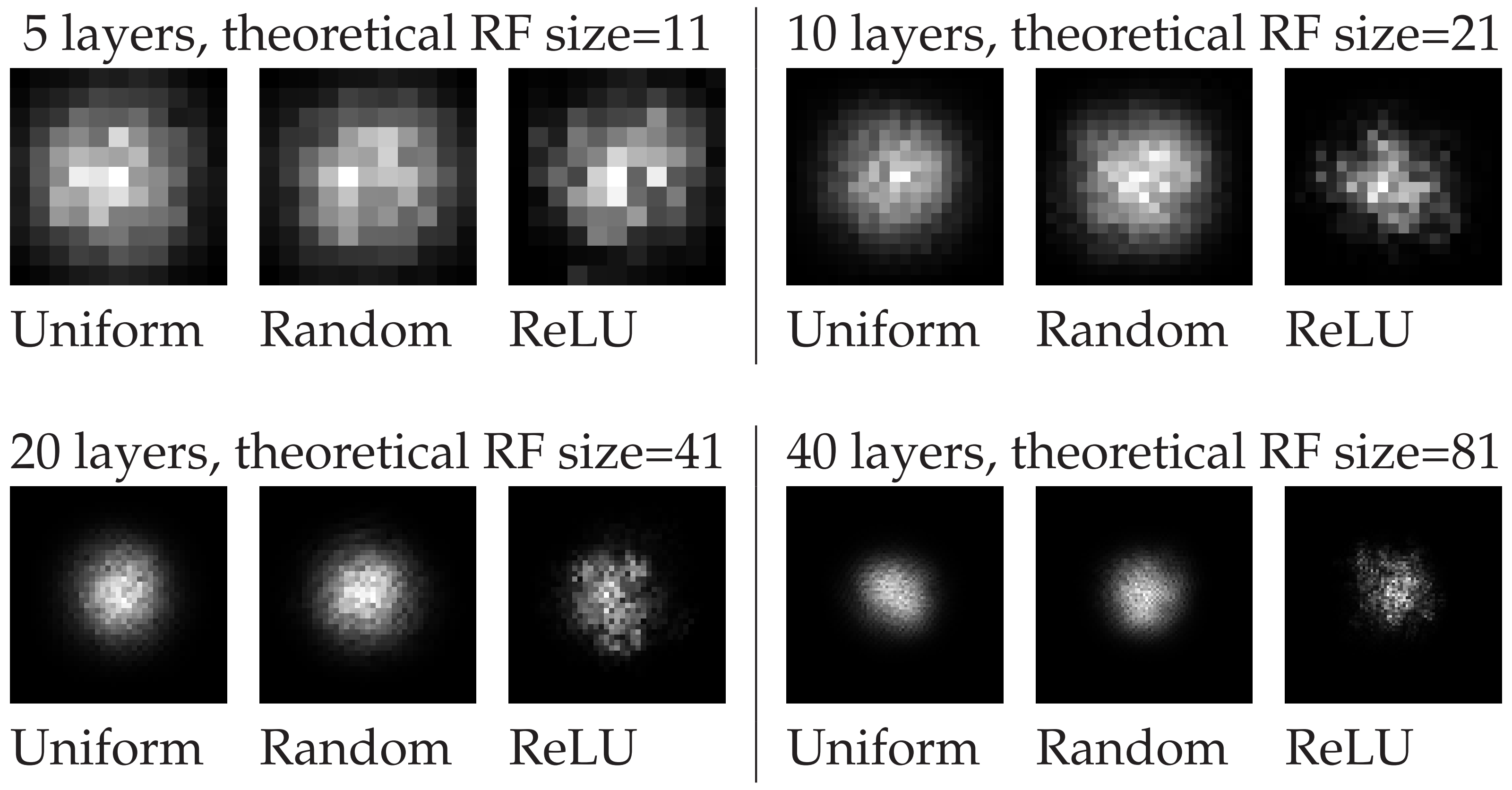
Central limit theorem says: as $n \to \infty$, the distribution of $\sqrt{n}(\frac{1}{n}S_n - \mathbb{E}[X])$ converges to Gaussian $\mathcal{N}(0, \mathrm{Var}[X])$ in distribution, i.e. $S_n \sim \mathcal{N}(n\mathbb{E}[X], n\mathrm{Var}[X])$ with

$$\mathbb{E}[S_n] = n\sum_{m=0}^{k-1} mw(m), \quad \mathrm{Var}[S_n] = n\left(\sum_{m=0}^{k-1} m^2 w(m) - \left(\sum_{m=0}^{k-1} mw(m)\right)^2\right)$$

Growth vs Shrinkage: ERF size is $\sqrt{\mathrm{Var}[S_n]} = \sqrt{n\mathrm{Var}[X_i]} = O(\sqrt{n})$; Correspondingly ERF ratio: $O(\frac{1}{\sqrt{n}})$.

Gaussian Shape

Comparing the effect of number of layers, random weight initialization and nonlinear activation on the ERF.



5 layers, theoretical RF size=11 — Uniform, Random, ReLU

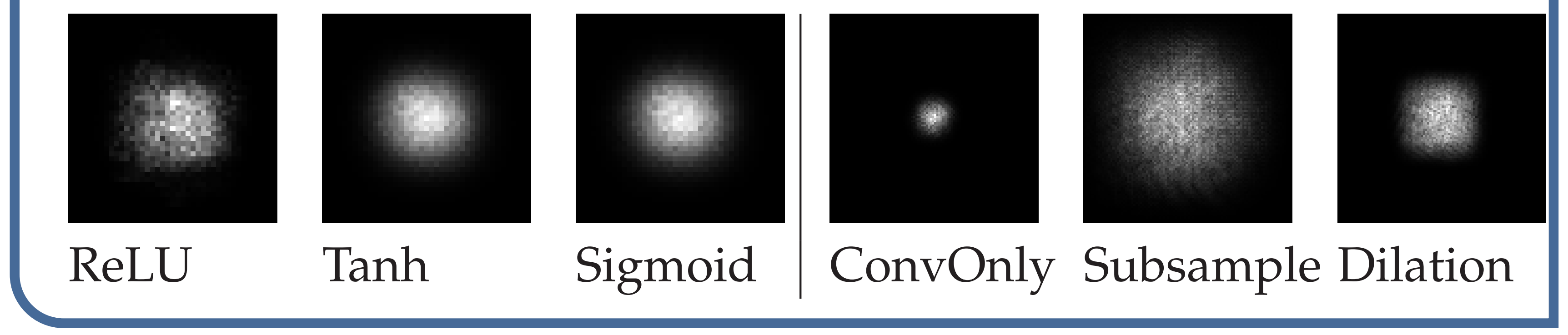
10 layers, theoretical RF size=21 — Uniform, Random, ReLU

20 layers, theoretical RF size=41 — Uniform, Random, ReLU

40 layers, theoretical RF size=81 — Uniform, Random, ReLU

Influence of Different Structures

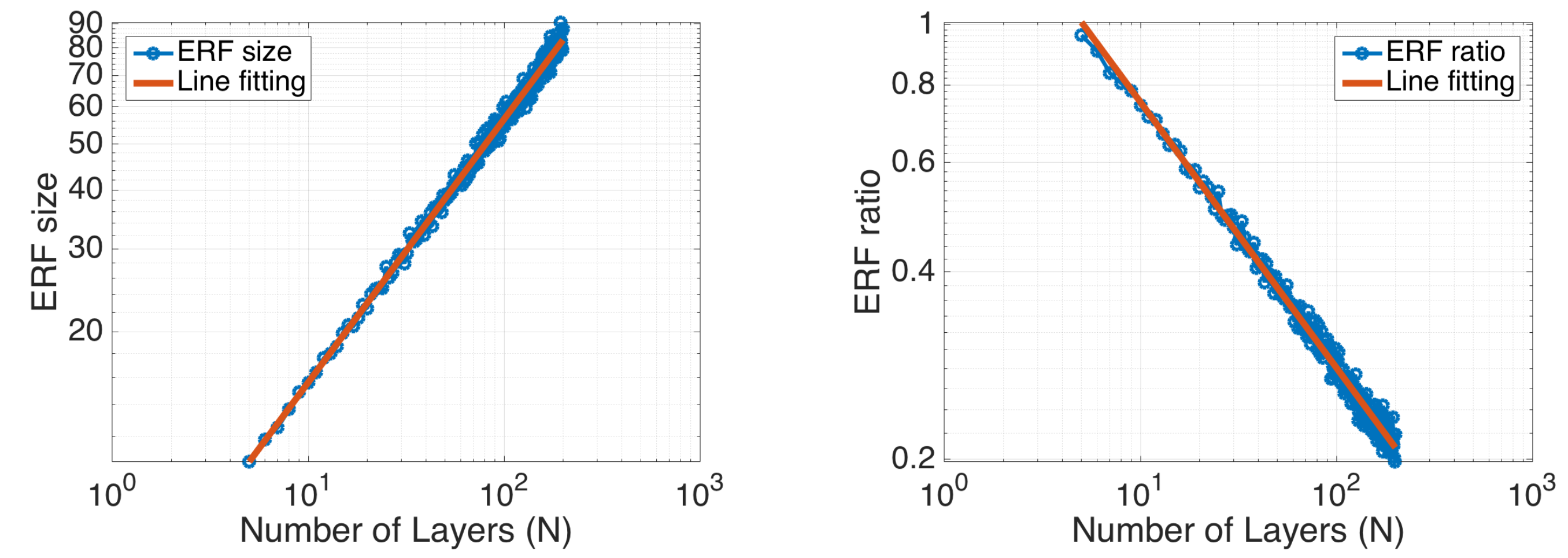
The left figure shows the effect of different non-linearity while the right figure shows the effect of subsampling and dilated convolution comparing to a pure convnet.



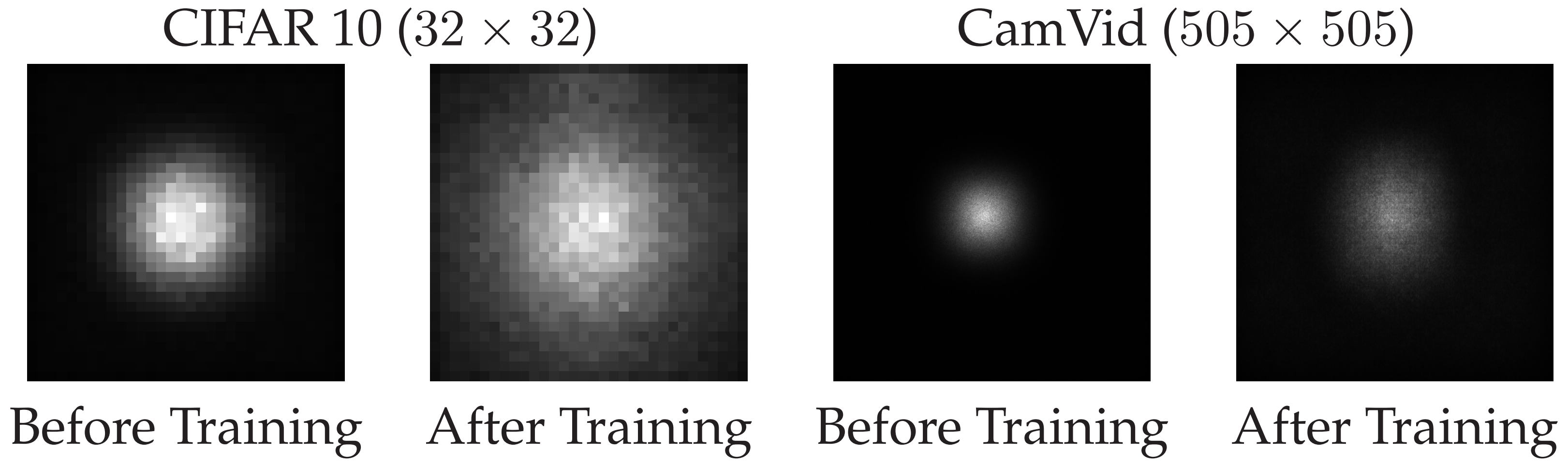
ReLU Tanh Sigmoid ConvOnly Subsample Dilation

Change of ERF

Absolute growth (left) and relative shrinkage (right) for ERF. The line for ERF growth has slope of 0.56 in log domain, while the line for ERF ratio has slope of -0.43. This indicates ERF size is growing linearly w.r.t $\sqrt{N}$ and ERF ratio is shrinking linearly w.r.t. $\frac{1}{\sqrt{N}}$.



Comparison of ERF before and after training for models trained on CIFAR-10 classification and CamVid semantic segmentation tasks. We can see ERF growth during training.



CIFAR 10 (32×32) — Before Training, After Training

CamVid (505×505) — Before Training, After Training

Connection to Other Work

Connection to biological neural networks: ERF in deep CNNs grows a lot slower than we used to think. It could preserve lots of local information; CNN may automatically create a form of foveal representation.

Connection to CNN applications: Variance analysis help better initialization [Xavier][He]; visualization of CNNs [Zeiler]; used as localization cue [Zhou] etc.