Universal Approximator Theorem: Let $\phi(\cdot)$ be a non-constant, bounded and monotonically increasing fn. $\forall\epsilon > 0$ and any continuous fn $\in \mathbb{R}^m$, there exists an integer N, real constants $v_i b_i \in \mathbb{R}$ and real vectors $w_i \in \mathbb{R}$ where $i = 1, \dots N$ such that: $F(\vec{x}) = \sum_{i=1}^{N} v_i \phi(\vec{w}_i^T \vec{x} + b_i) \quad \text{with } |F(\vec{x}) - f(\vec{x})| < \epsilon$ where ϕ is a sensible activation function. <u>Problems</u>: ϵ can be very large in practice, making approximation less useful, and curse of dimensionality.

Shift Invariance: The unchanging response when the input is shifted. For classifier f and shift operator S_v, $f(\vec{x}) = f(S_v \vec{x})$ (no matter how the input is transformed, the output should remain constant) generalizes for unseen data.

Shift Equivariance: applying the shift operator after the function yields the same results as applying the function after the shift. i.e. $S_v \circ f(\vec{x}) = f(\vec{x}) \circ S_v$. It is about consistent transformation.

Translation Invariance: shift in input should have a predictable shift in hidden representation (location shouldn't matter)
Locality: we should not have to move far from location (i, j) to learn valuable information to asses what the area contains.

Fully connected net: every input feature n in an image influences ever neuron in the next layer: $n \times n$. **Sparsely connected net**: each neuron n is connected to a subset of neurons k: $k \times n$. **Weight sharing net**: weights are reused in a network

Convolution: $(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau \quad for f, g : [0, \infty] \to \mathbb{R}$
Correlation: $(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t + \tau) d\tau \quad for f, g : [0, \infty] \to \mathbb{R}$
commutative: $f * g = g * f$, associative $(f * g) * h = f * (g * h)$
distributive: $f_1 * (f_2 + f_3) = f_1 * f_2 + f_1 * f_3$

$$M = \left\lfloor \frac{M + 2 \times P - D \times (K-1) - 1}{S} + 1 \right\rfloor$$

standard CNNs are not naturally equivariant to rotations; Harmonic Networks/H-Nets; replacing kernel w/ 'circular harmonics'; rotation results in a proportionate rotation in the output.

Curse of Dimensionality:
<u>Sample Explosion</u>: As the number of features or dimensions grows, the amount of data we need to generalize accurately grows exponentially To approximate a (Lipschitz) continuous function $f : \mathbb{R}^d \to \mathbb{R}$ with ϵ accuracy one needs $O(\epsilon^{-d})$ samples.
<u>Sparseness</u>: The more features we use, the more sparse the data becomes such that accurate estimation of the classifier's parameters (i.e. its decision boundaries) becomes more difficult, this sparseness is not uniformly distributed over the search space; the higher dimensions you have the higher probability that a data-point will sit in its own distinct corner in the hypercube.
<u>Math</u>: $V_{rind}^n = (1 - \alpha^n) V_{original} \Rightarrow \frac{V_{rind}}{V_{original}} = 1 - \alpha^n \Rightarrow \frac{d(1-\alpha^n)}{d\alpha} = -n\alpha^{n-1}$ $\iff d(1 - \alpha^n) = -n\alpha^{n-1} d\alpha$. this shows that the volume of the rind initially grows much faster,n times faster than the rate at which the object is being shrunk (when $\alpha = 1$ and $d\alpha < 0$ then $d(1 - \alpha^n) = n|d\alpha|$); In higher dims, small changes in distance lead to vast changes in vol.

Factorized conv: 2 3x3 convolutions can act as an approx for 5x5 conv trading expressiveness for efficiency. Inserting a non-linearity between the 3x3s lets it capture more complex features.
Separable conv: approximate a 5x5 conv with a 5x1 and 1x5 reducing params (as above). Lossy.

Pooling: smaller res, hierarchal features (concentrates abstract features), shift/deform Invariance. Break shift-equivariance by blurring sample to avoid the shifting pooling issue.

Approximate Deformation Invar: $||f(\vec{x}) - f(D_\tau \vec{x})|| \approx ||\nabla \tau||$ deform img τ=deform factor