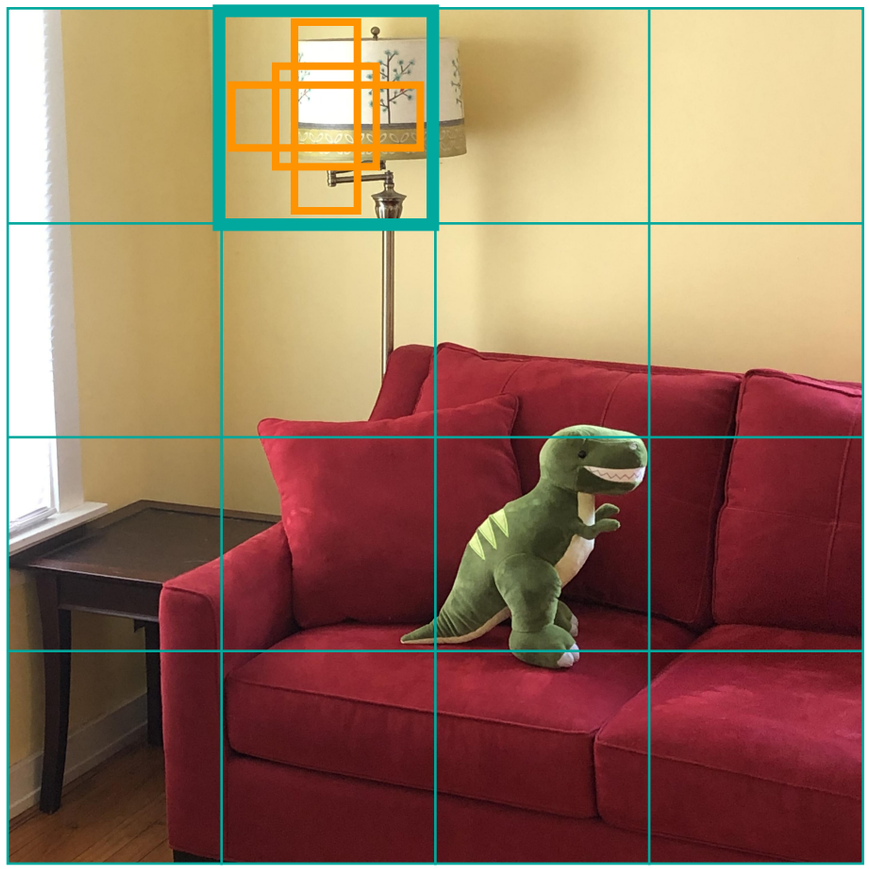
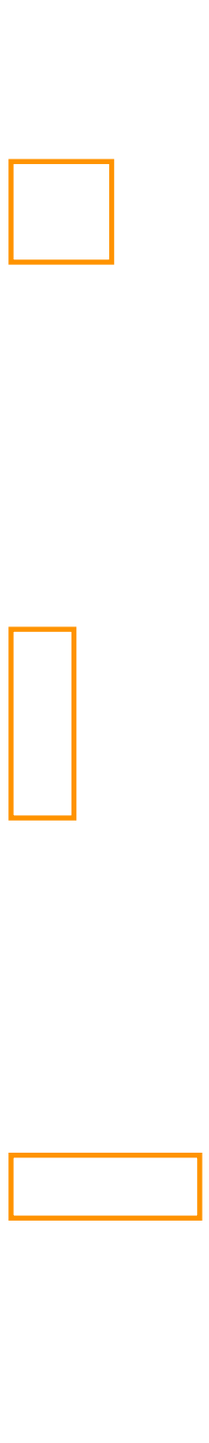
Set Up for Network Outputs



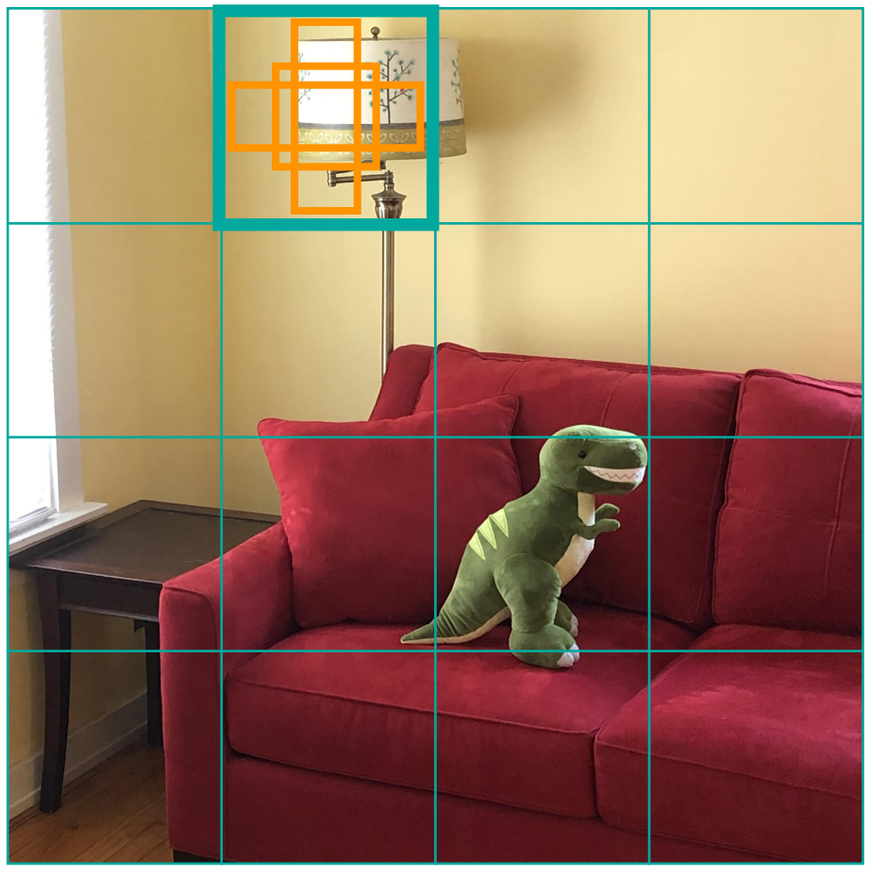
Our network must output predictions for all cells and anchor boxes!
Combined, for this image y is an array of shape **(4, 4, 3 * (1 + 4 + K))**

Outputs for all 3 anchor boxes in second grid cell of top row:

$$y[0, 1, :] = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0.8 \\ 0.9 \\ 0.8 \\ 1.7 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$



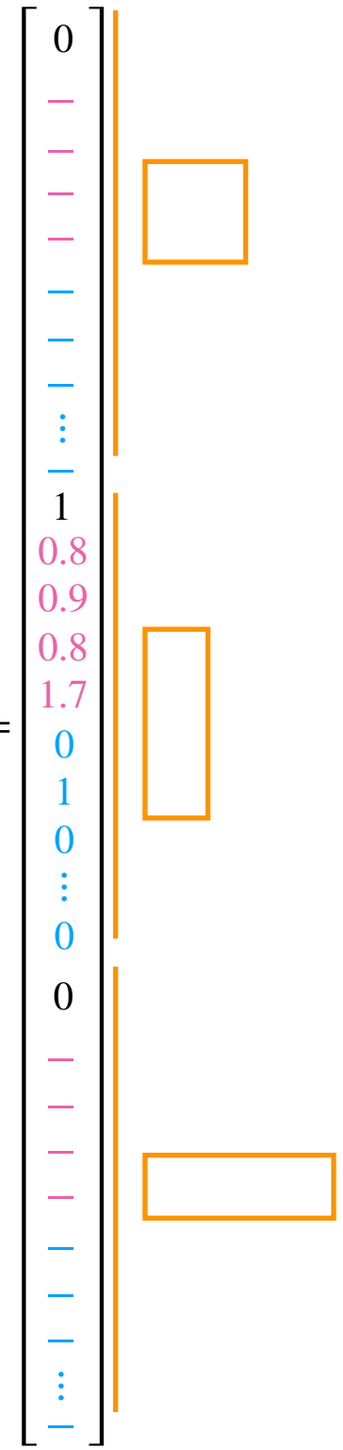
Set Up for Network Outputs



Our network must output predictions for all cells and anchor boxes!
Combined, for this image y is an array of shape **(4, 4, 3 * (1 + 4 + K))**

Outputs for all 3 anchor boxes in second grid cell of top row:

y[0, 1, :] =



Note:

- Although the image is divided into 16 cells, we run prediction for the model once ("you only look once").
- The output of a single pass through the network is an array of shape **(4, 4, 3 * (1 + 4 + K))** with predictions for all cells and anchor boxes.

YOLO Network Architecture

- Each YOLO paper used a different network architecture (later papers used bigger models).
- Here is the architecture from the first YOLO paper, which used 7x7 grid cells:

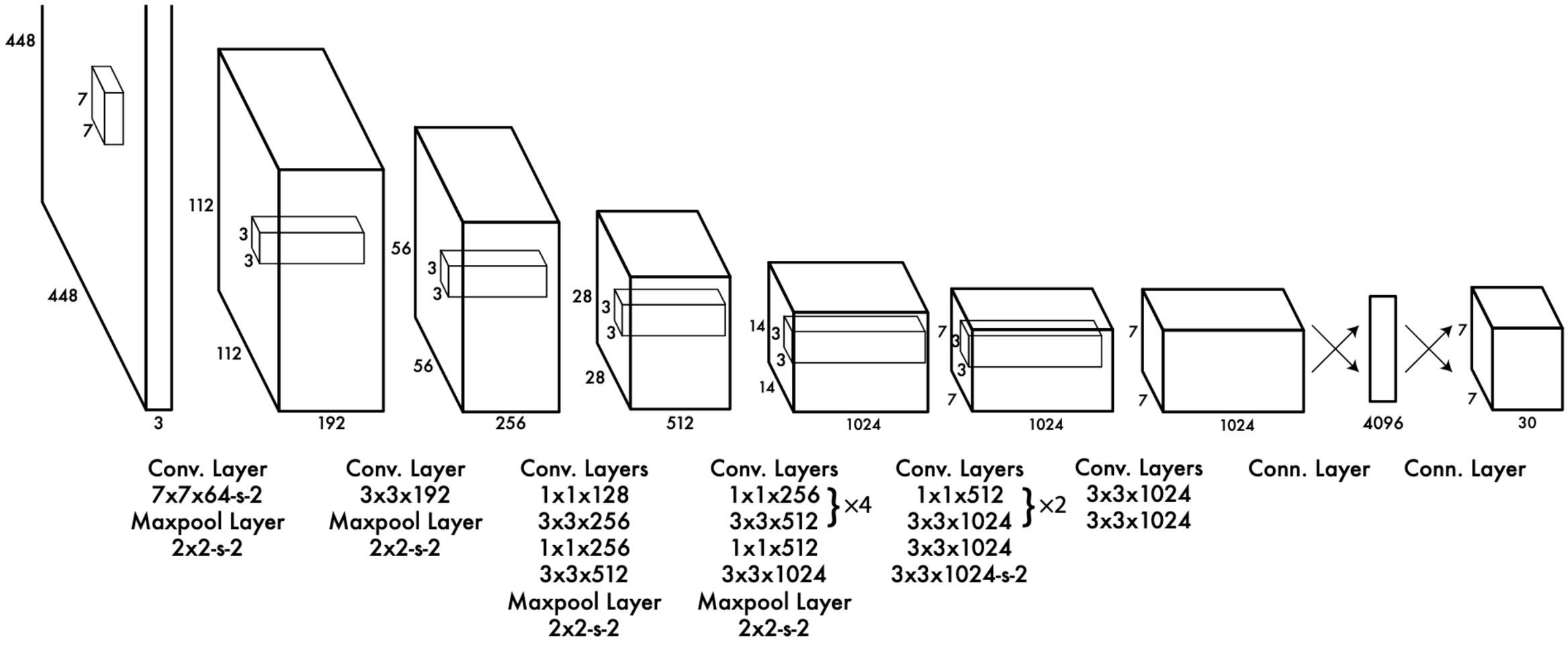
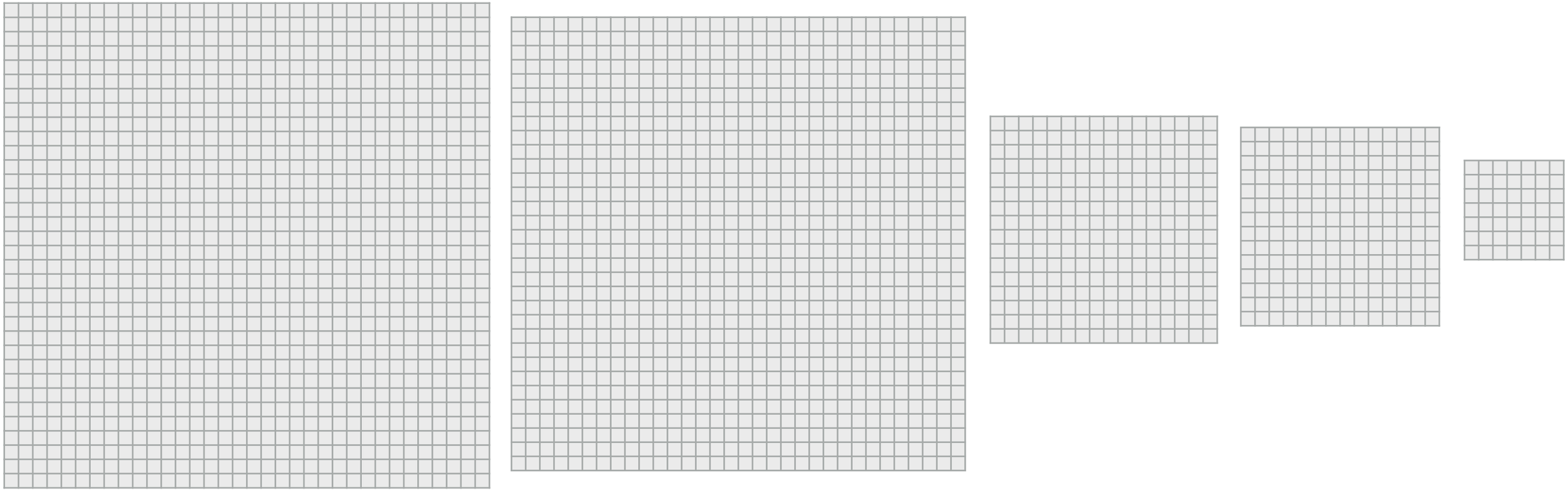


Figure from Redmon et al., “You Only Look Once: Unified, Real-Time Object Detection.” (2016)

Output Layer Shapes Example

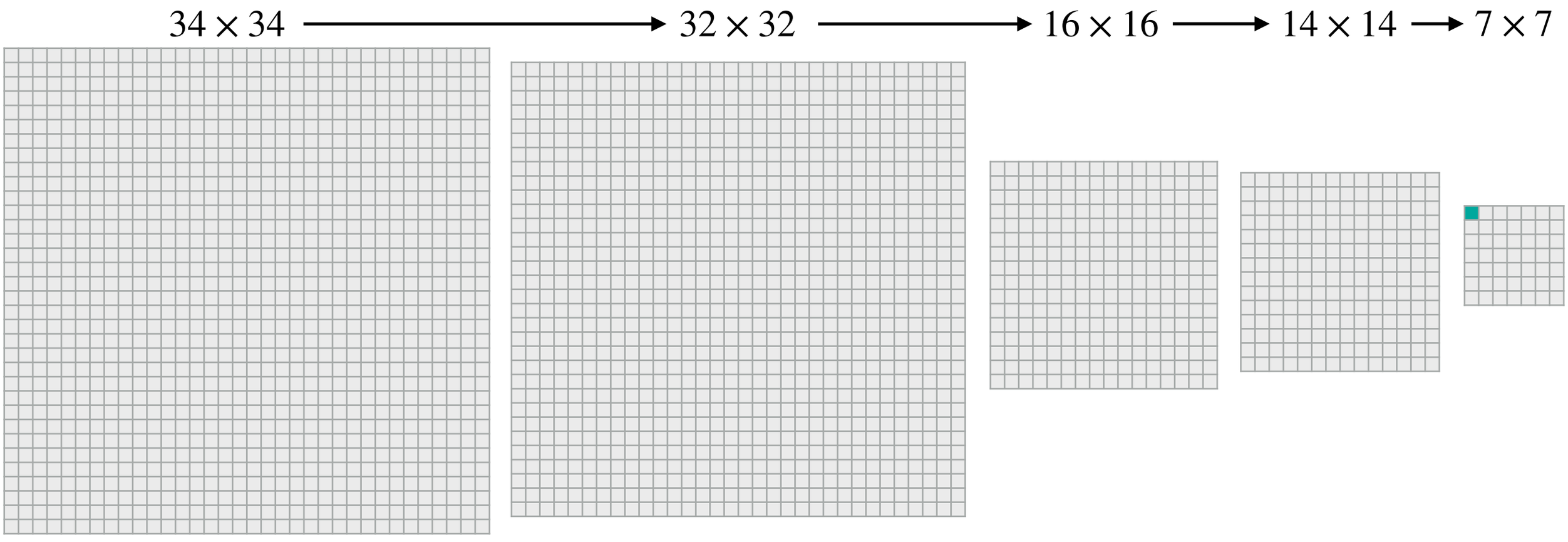
- Consider a network with: (1) a 3×3 convolutional filter; (2) 2×2 max pooling with stride 2; (3) a 3×3 convolutional filter; (4) 2×2 max pooling with stride 2
- Recall: if input is $n \times n$, output from 3×3 filter is $(n-2) \times (n-2)$
- If input is $n \times n$, output from 2×2 max pooling is $(n/2) \times (n/2)$
- Suppose input is 34×34

$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



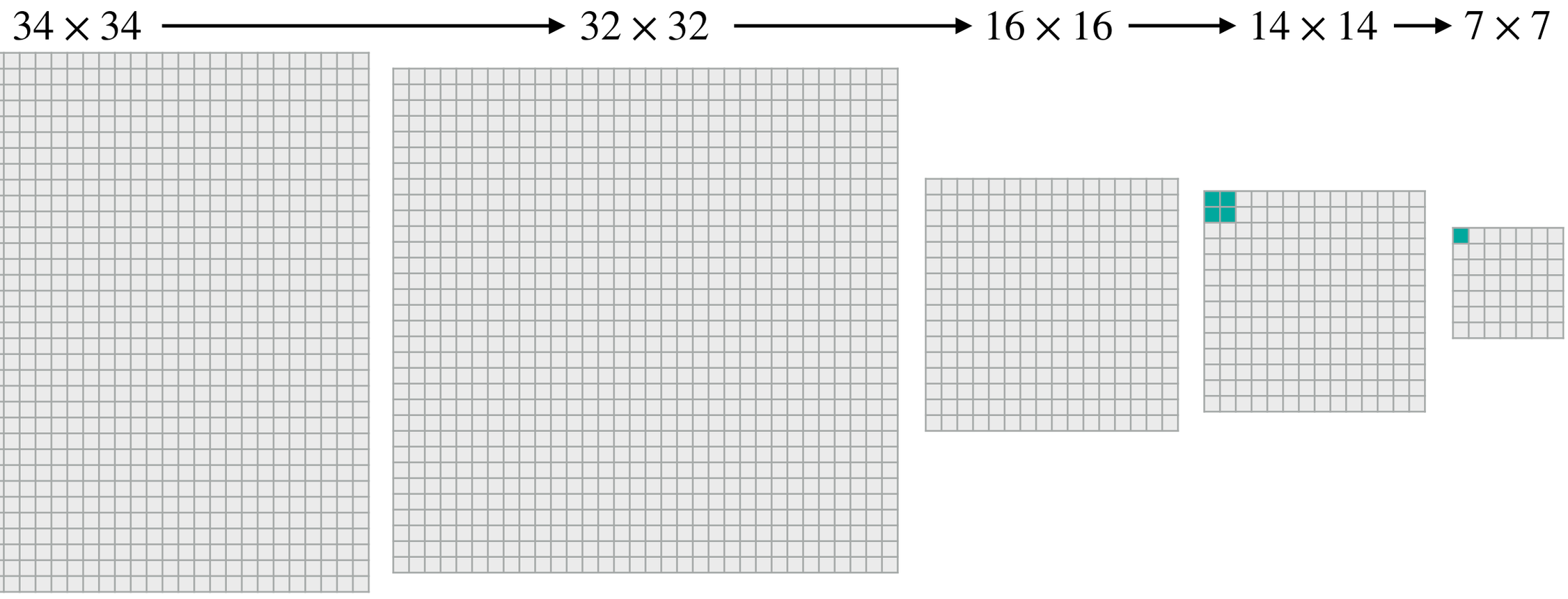
Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?



Effective Receptive Field

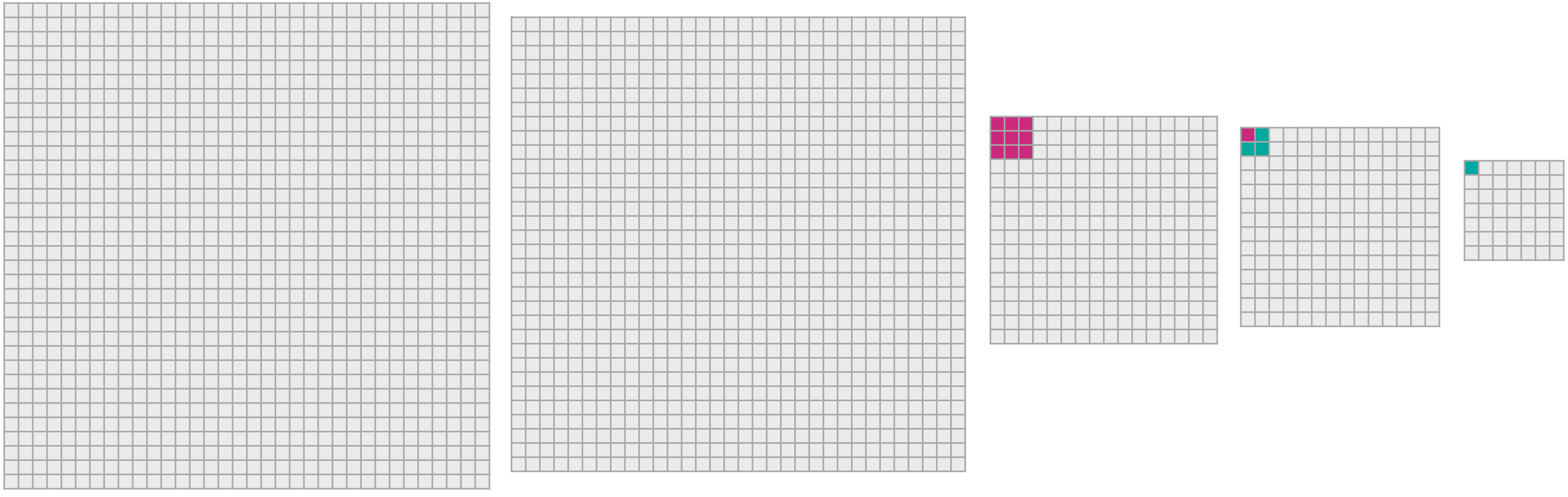
- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?



Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?

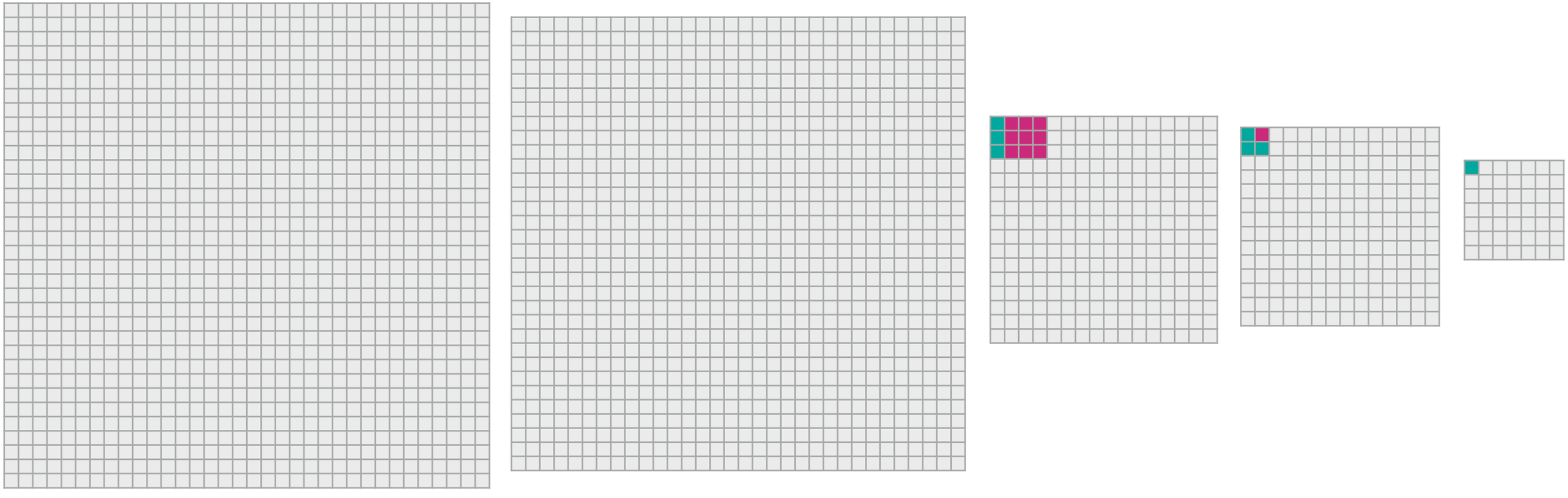
$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?

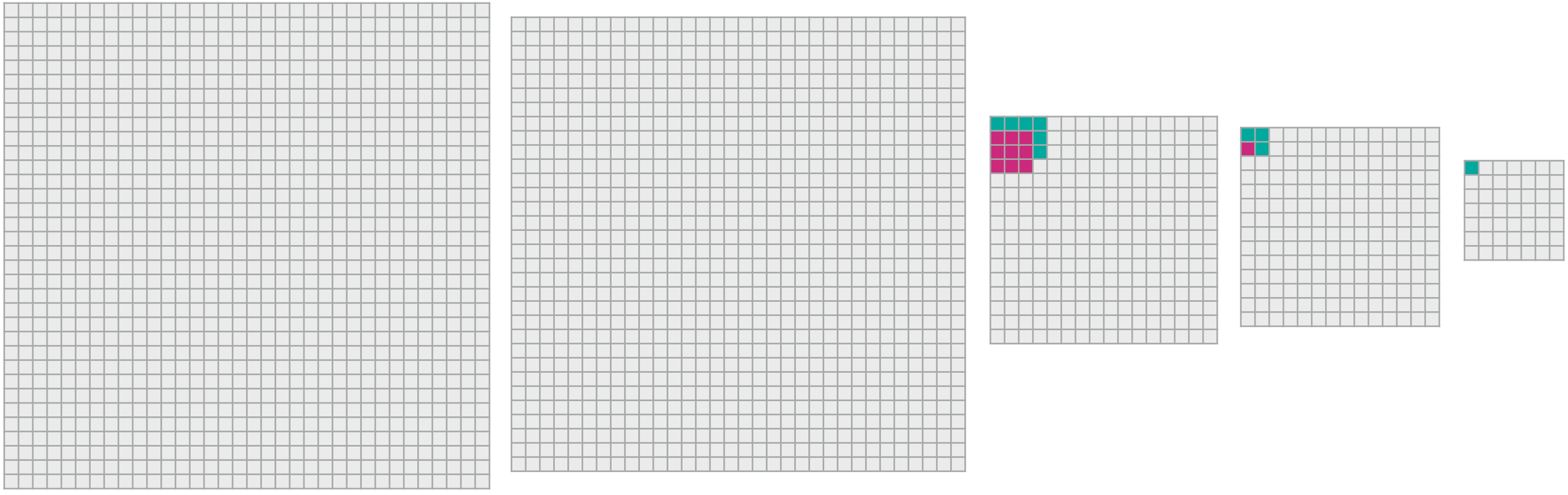
$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



Effective Receptive Field

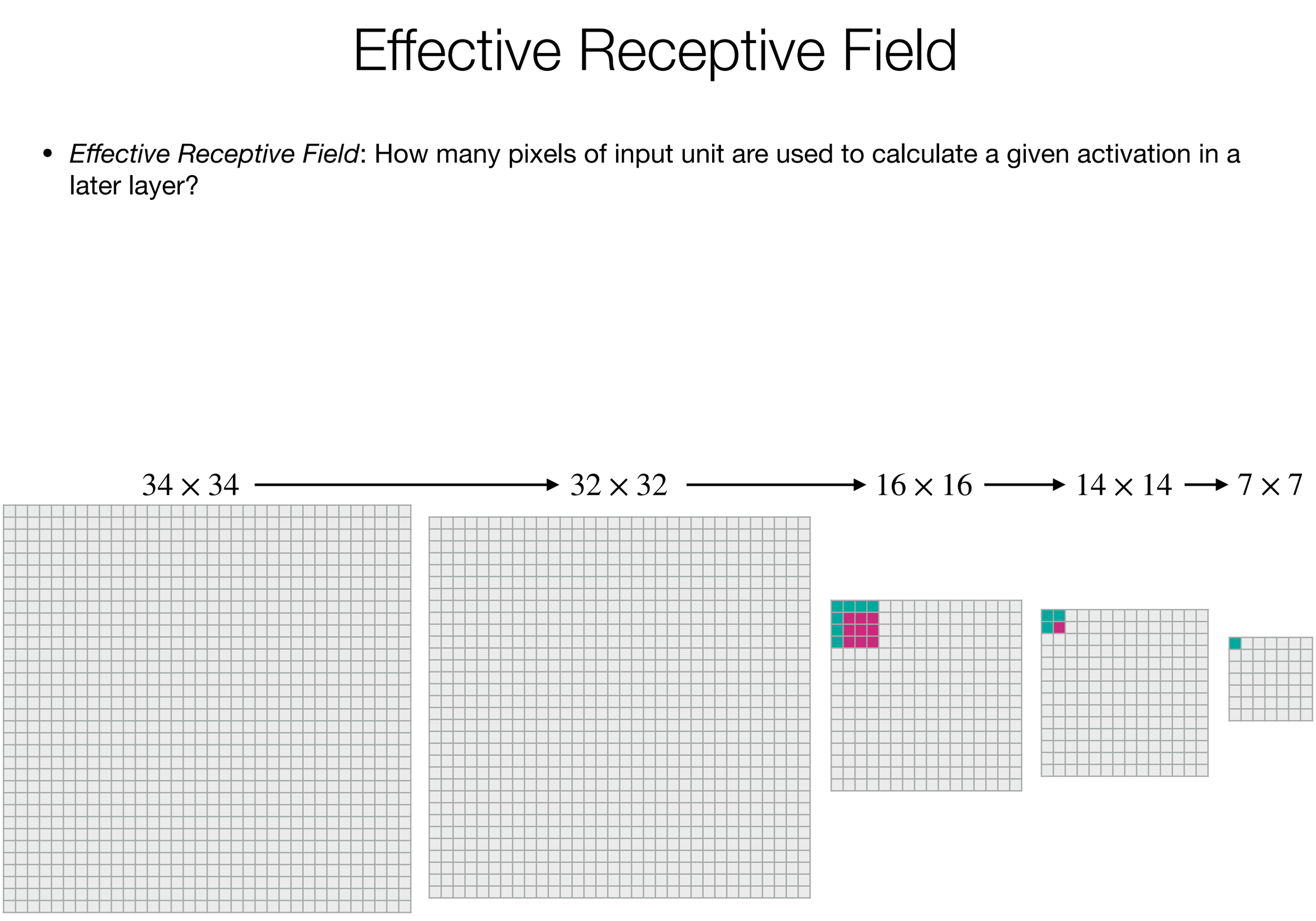
- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?

$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



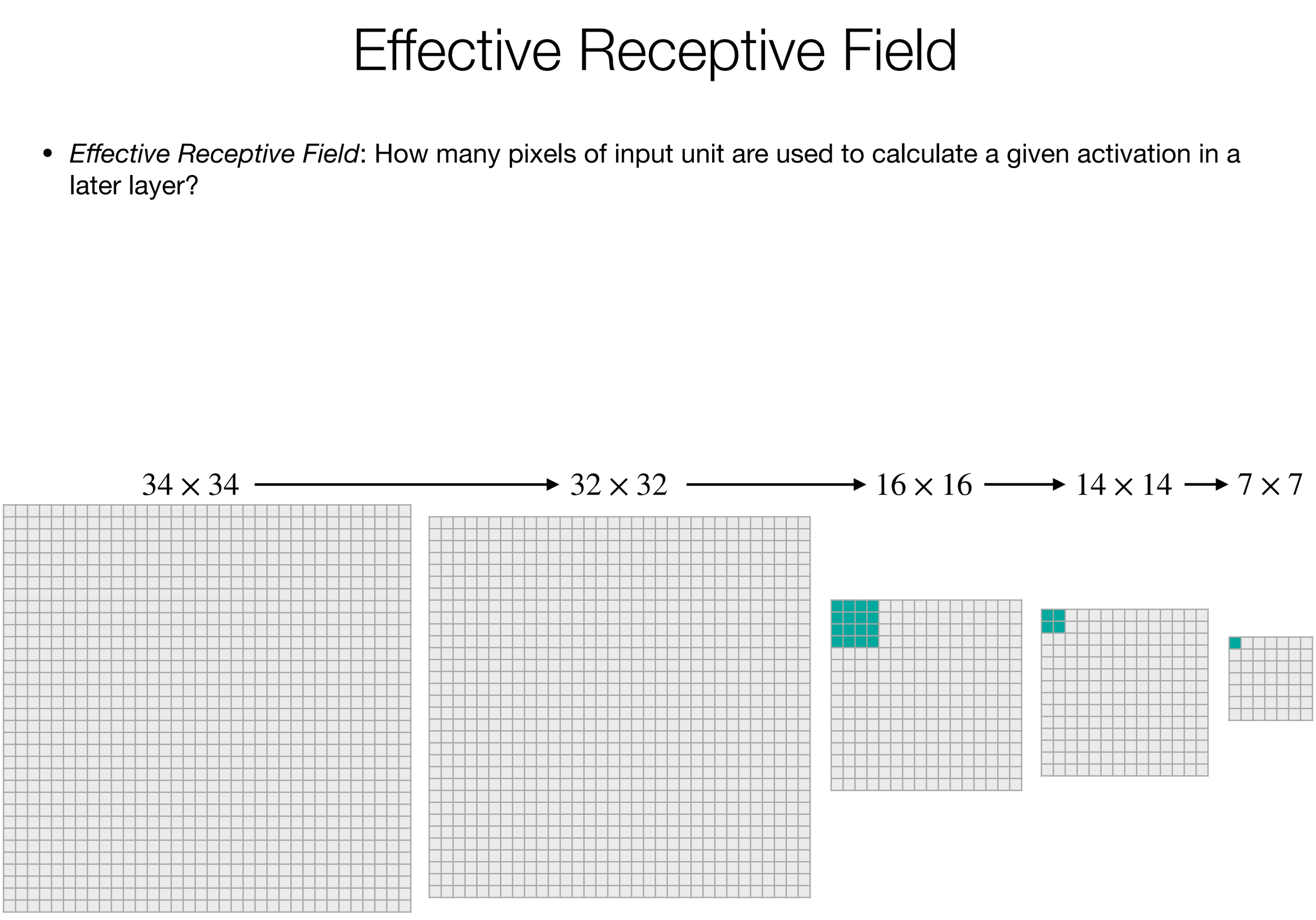
Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?



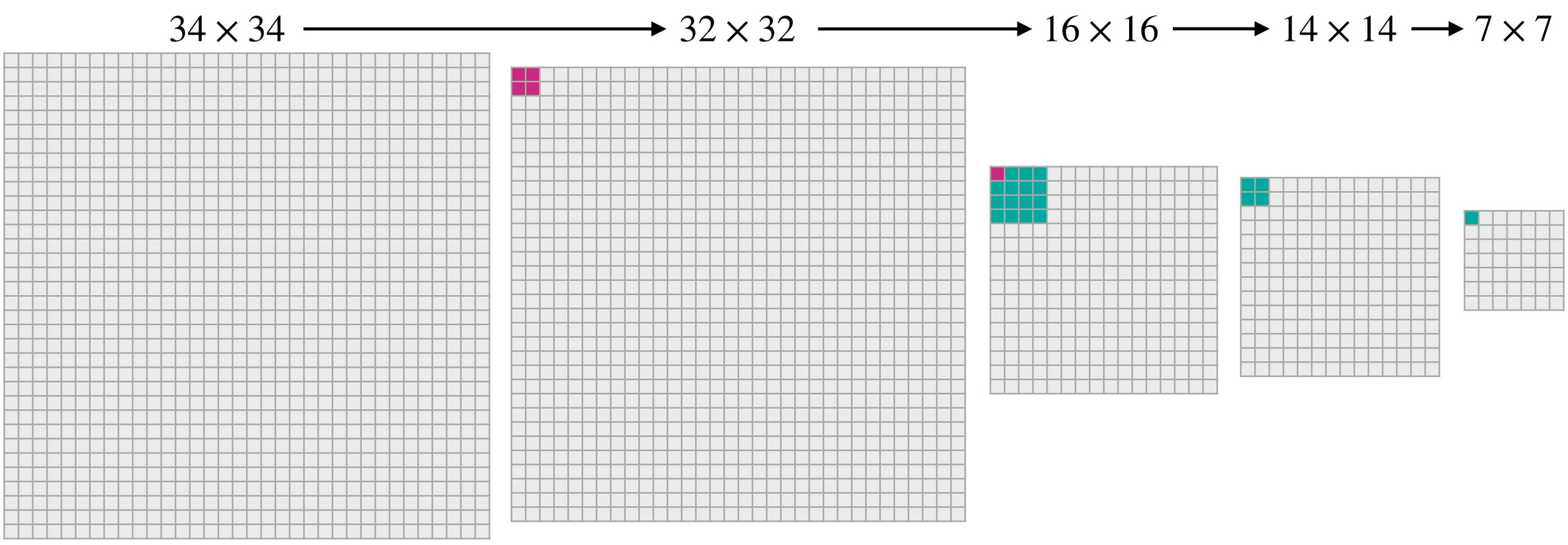
Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?



Effective Receptive Field

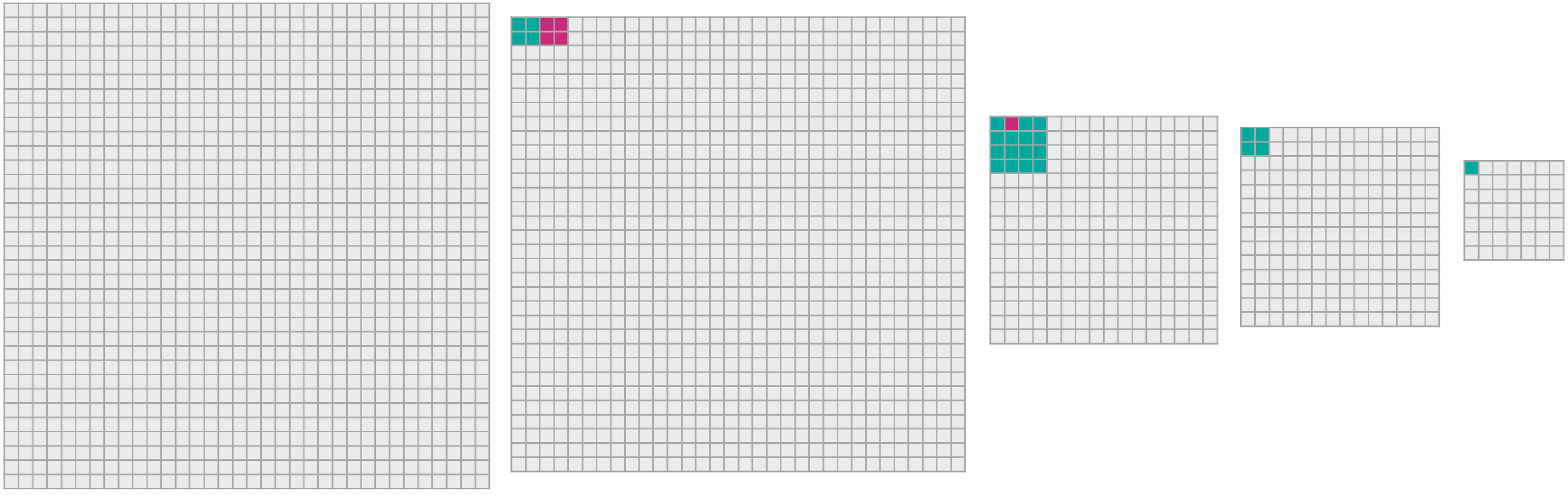
- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?



Effective Receptive Field

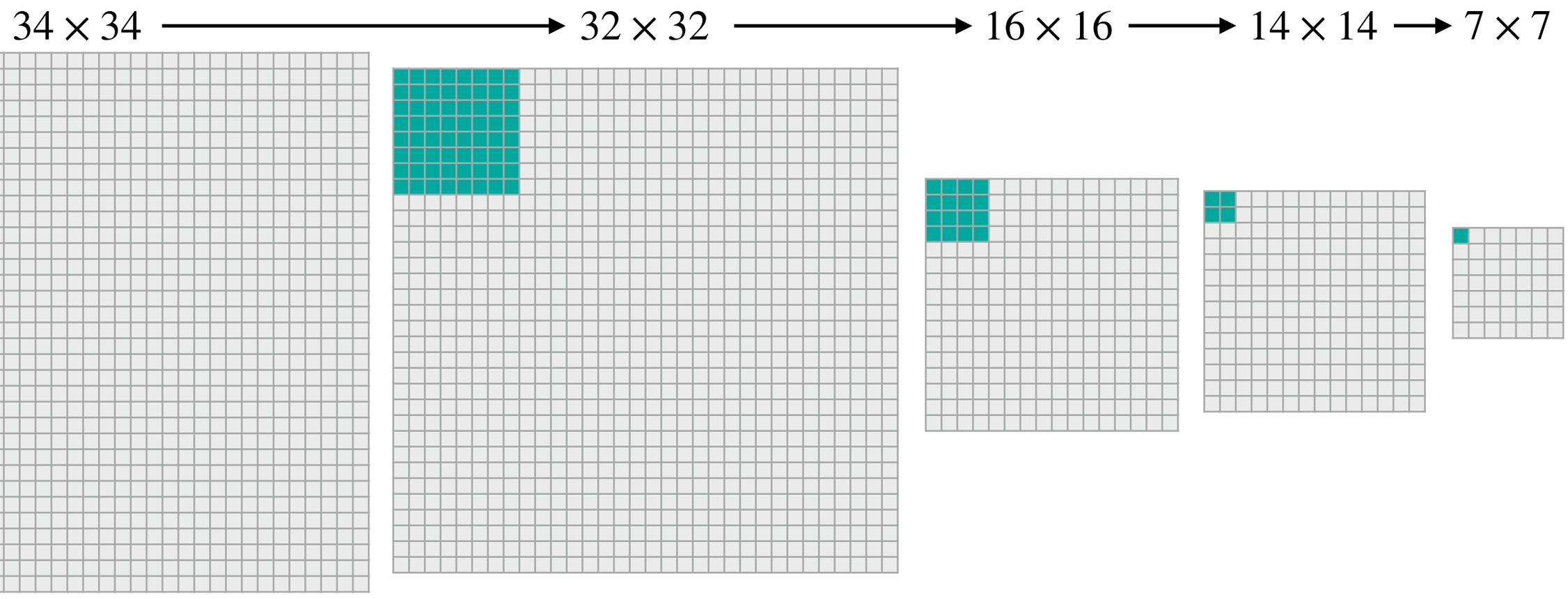
- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?

$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



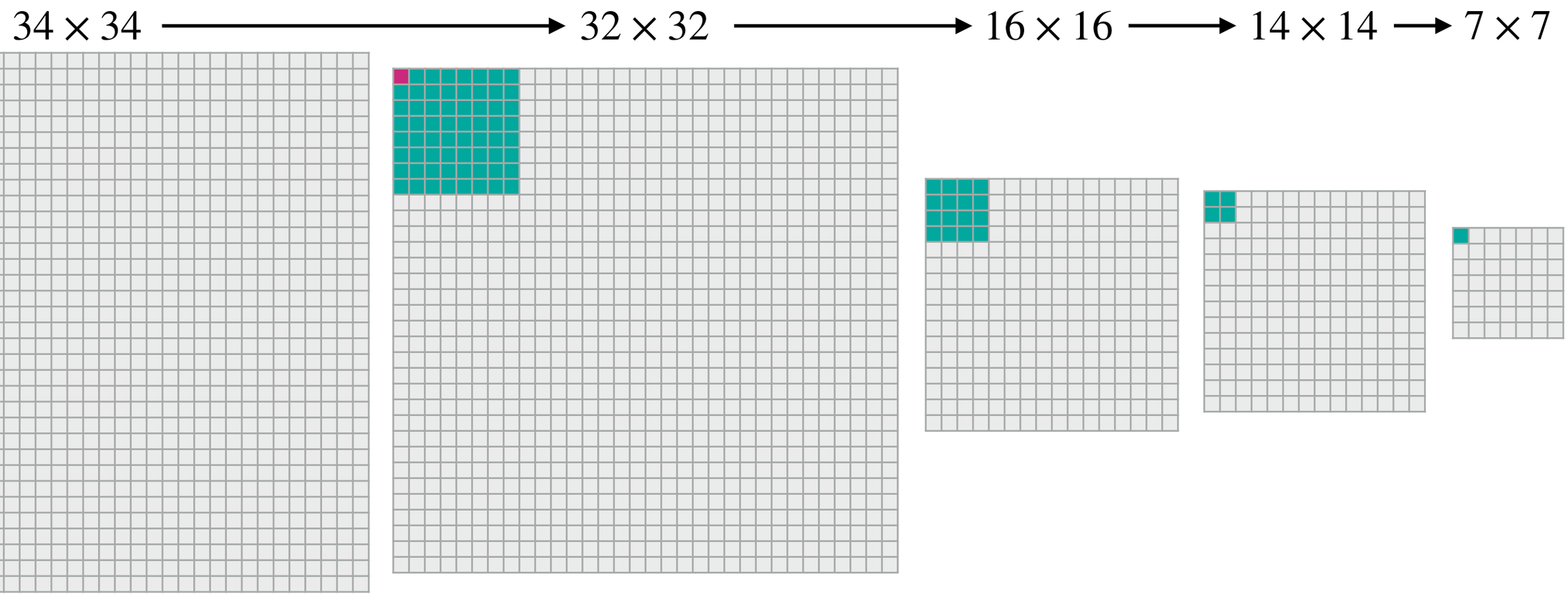
Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?



Effective Receptive Field

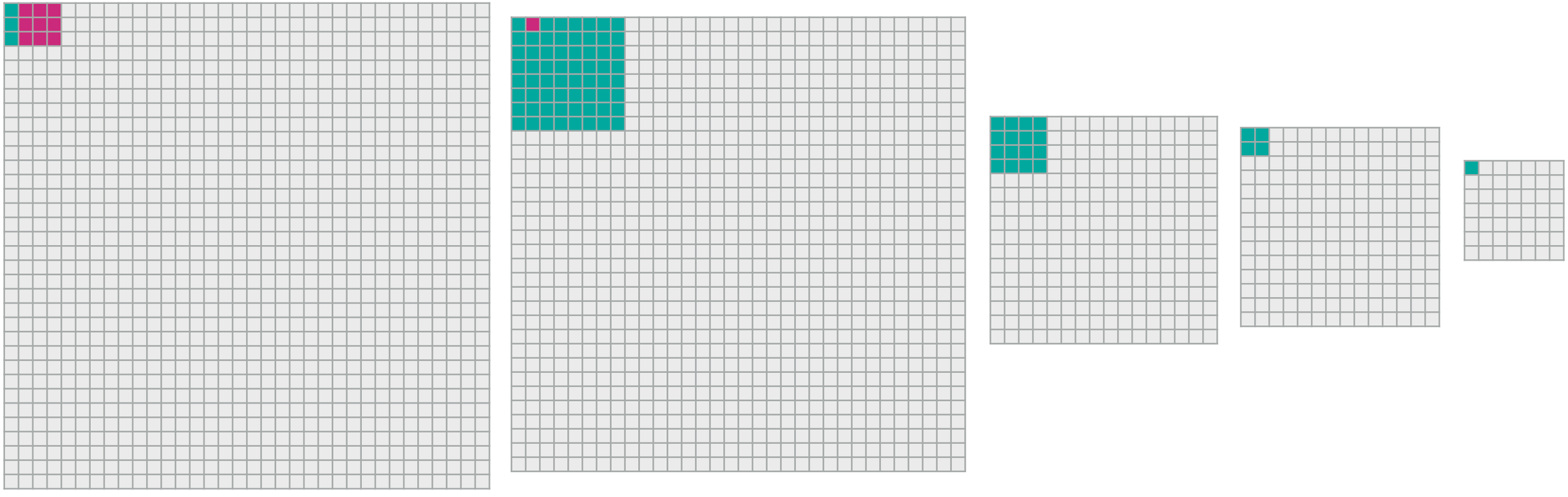
- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?



Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?

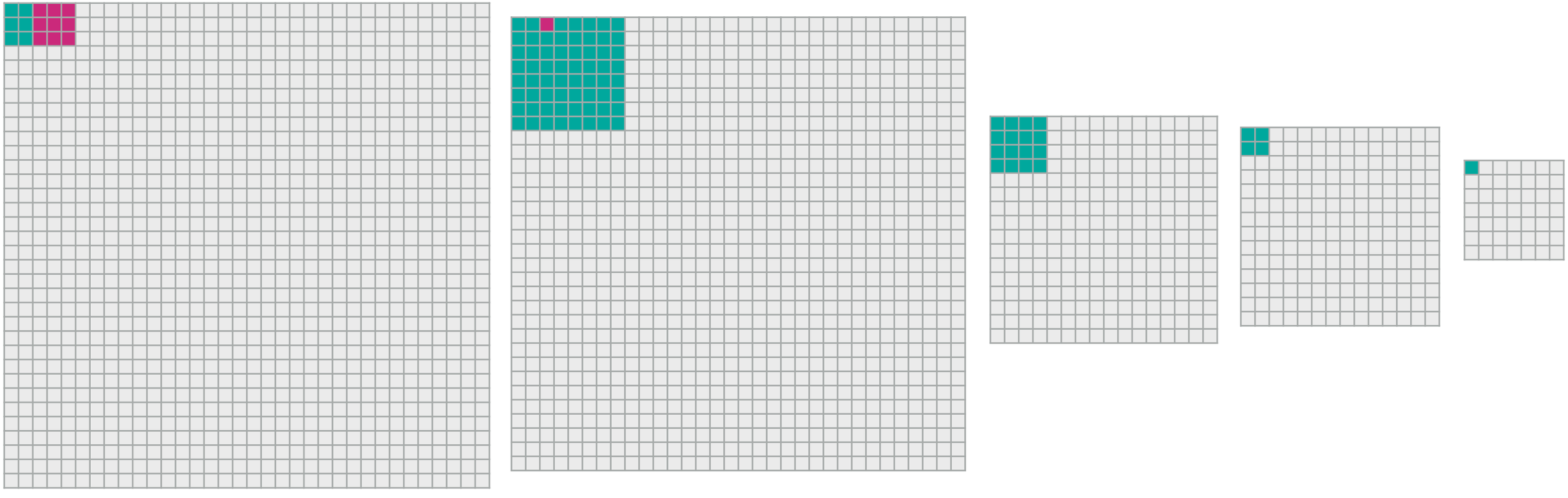
$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?

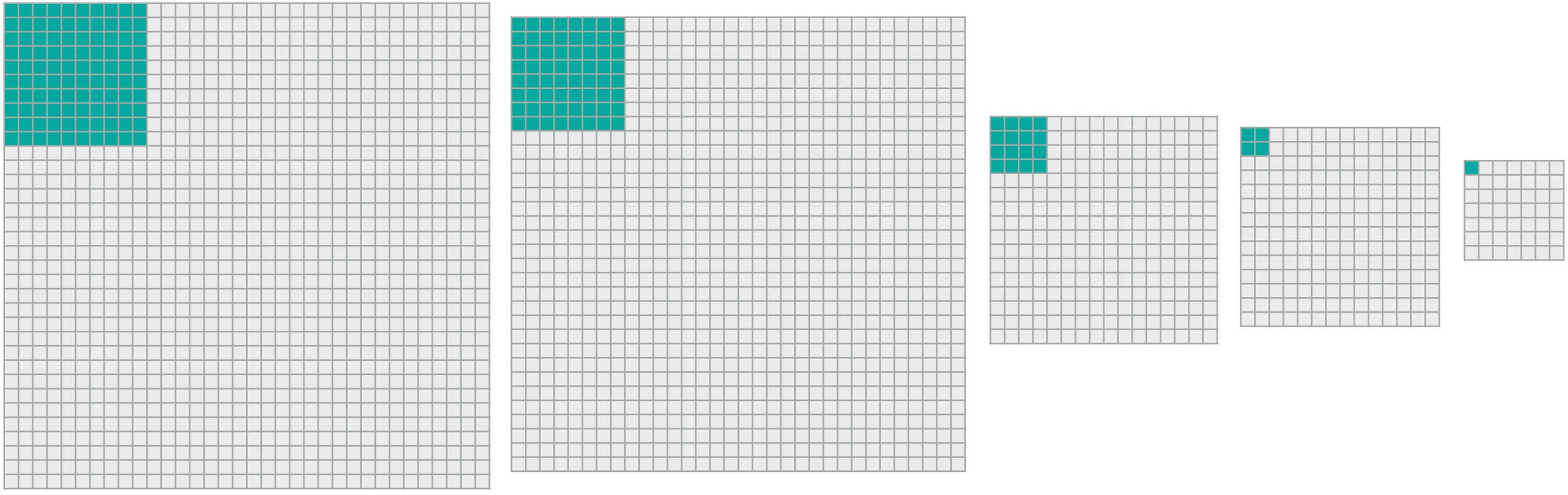
$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?

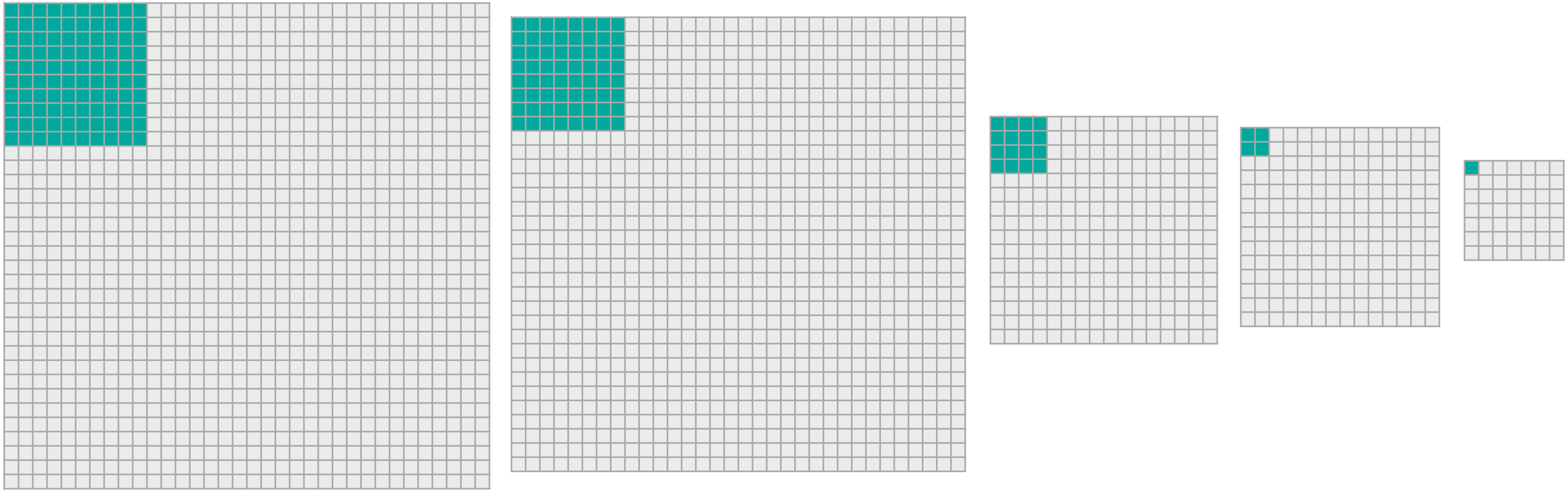
$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?
  - For this example, effective receptive field is 10×10
  - Note: If we had 7 cells, each would be about 5×5 in the input image
  - We use information from outside of a given grid cell to inform predictions for that grid cell.
  - For a deeper network, effective receptive field is even larger

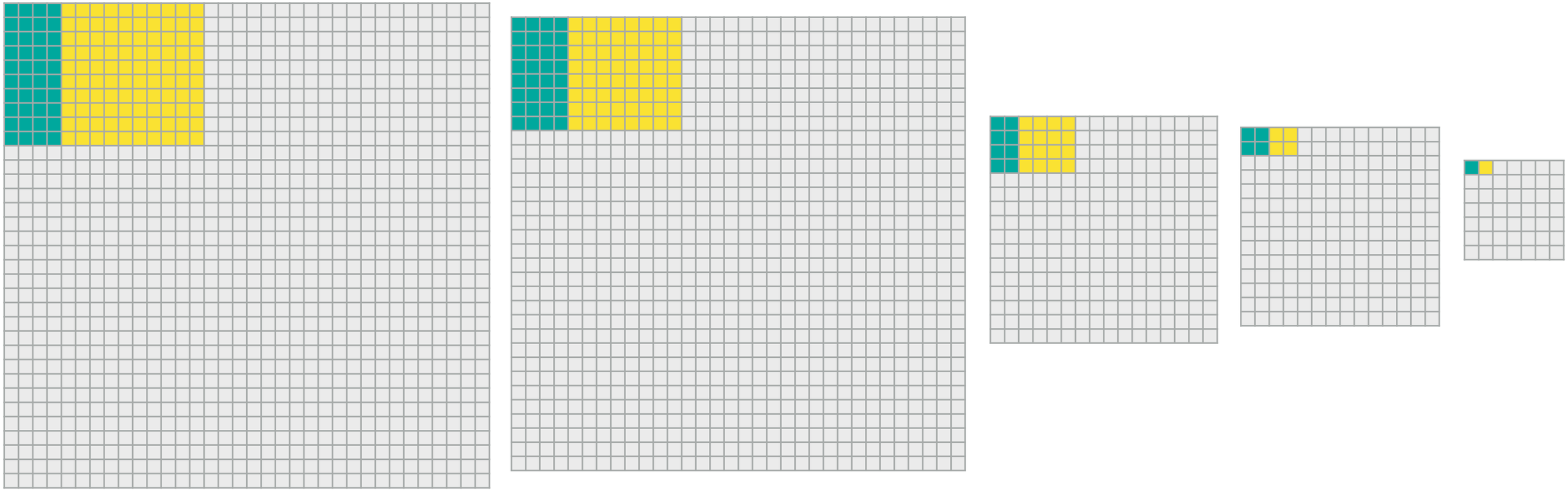
$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?
  - For this example, effective receptive field is 10×10
  - Note: If we had 7 cells, each would be about 5×5 in the input image
  - We use information from outside of a given grid cell to inform predictions for that grid cell.
  - For a deeper network, effective receptive field is even larger
- There is a lot of overlap in effective receptive fields for neighboring cells

$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$



Effective Receptive Field

- *Effective Receptive Field*: How many pixels of input unit are used to calculate a given activation in a later layer?
  - For this example, effective receptive field is 10×10
  - Note: If we had 7 cells, each would be about 5×5 in the input image
  - We use information from outside of a given grid cell to inform predictions for that grid cell.
  - For a deeper network, effective receptive field is even larger
- There is a lot of overlap in effective receptive fields for neighboring cells
  - An advantage to using convolutions is that the same computations are re-used!

$34 \times 34 \longrightarrow 32 \times 32 \longrightarrow 16 \times 16 \longrightarrow 14 \times 14 \longrightarrow 7 \times 7$