Trong đoạn mã của lớp Model\_0, khi tính toán kích thước đầu vào cho lớp **Fully Connected Layer** (lớp Linear), việc nhân với 16\*16 liên quan đến kích thước không gian (chiều cao và chiều rộng) của các feature map sau các lớp tích chập và pooling.

Dưới đây là giải thích chi tiết tại sao có sự nhân với 16\*16:

**1. Từ đầu vào ban đầu:**

* Giả sử bạn đang làm việc với đầu vào là ảnh có kích thước (C, H, W) (với C là số kênh, H là chiều cao, và W là chiều rộng).
* input\_shape là số lượng kênh của ảnh đầu vào (ví dụ: với ảnh RGB, input\_shape thường là 3).

**2. Tính toán kích thước sau conv\_block\_1:**

* Lớp tích chập đầu tiên (nn.Conv2d) có kernel\_size=3, stride=1, và padding=0.
  + Nếu bạn bắt đầu với đầu vào có kích thước (C, H, W), sau lớp tích chập đầu tiên, chiều rộng và chiều cao của ảnh sẽ giảm theo công thức:

H′=H−kernel\_size+2×paddingstride+1=H−31+1=H−2H' = \frac{H - \text{kernel\\_size} + 2 \times \text{padding}}{\text{stride}} + 1 = \frac{H - 3}{1} + 1 = H - 2H′=strideH−kernel\_size+2×padding​+1=1H−3​+1=H−2 W′=W−31+1=W−2W' = \frac{W - 3}{1} + 1 = W - 2W′=1W−3​+1=W−2

Do đó, nếu đầu vào có kích thước ban đầu là (C, H, W), sau lớp tích chập đầu tiên, kích thước sẽ là (hidden\_units, H-2, W-2).

* Lớp tích chập thứ hai trong conv\_block\_1 có kernel\_size=2, padding=1, và stride=1. Điều này sẽ làm thay đổi chiều cao và chiều rộng như sau:

H′=(H−2)−2+2×11+1=H−2H' = \frac{(H-2) - 2 + 2 \times 1}{1} + 1 = H-2H′=1(H−2)−2+2×1​+1=H−2 W′=W−2W' = W-2W′=W−2

Kích thước sau lớp này vẫn sẽ là (hidden\_units, H-2, W-2).

* Sau đó, lớp **MaxPool2D** với kernel\_size=2 và stride=2 sẽ giảm kích thước chiều cao và chiều rộng đi một nửa:

H′=H−22,W′=W−22H' = \frac{H-2}{2}, \quad W' = \frac{W-2}{2}H′=2H−2​,W′=2W−2​

Nếu bạn giả sử kích thước ảnh đầu vào ban đầu là 32x32, sau conv\_block\_1, kích thước feature map sẽ trở thành 16x16.

**3. Tính toán sau conv\_block\_2:**

* Quy trình tương tự diễn ra cho conv\_block\_2, tiếp tục sử dụng các lớp tích chập và MaxPool2D. Tuy nhiên, kích thước không gian sau khi qua conv\_block\_1 đã giảm xuống còn 16x16.

**4. Kích thước đầu vào của lớp Fully Connected:**

* Sau khi qua tất cả các lớp tích chập và pooling, dữ liệu trở thành một ma trận 4D có kích thước (batch\_size, hidden\_units, 16, 16).
* Trước khi đưa vào lớp **Linear (Fully Connected)**, ta cần "phẳng hóa" (flatten) nó thành một vector 2D. Khi đó, tổng số giá trị của ma trận sẽ là:

in\_features=hidden\_units×16×16\text{in\\_features} = hidden\\_units \times 16 \times 16in\_features=hidden\_units×16×16

* Do đó, trong lớp Linear, số lượng giá trị đầu vào (in\_features) phải là hidden\_units \* 16 \* 16.

**Tóm lại:**

Số nhân 16\*16 là kích thước không gian (chiều cao và chiều rộng) của feature map sau khi qua các lớp tích chập và pooling trong conv\_block\_1. Nếu kích thước ảnh ban đầu là 32x32, sau khi qua các lớp tích chập và pooling, kích thước giảm còn 16x16. Điều này giải thích tại sao đầu vào cho lớp Linear có hệ số hidden\_units \* 16 \* 16.

A screenshot of a black and white text

Description automatically generated