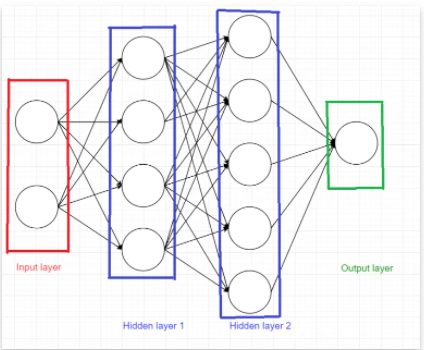
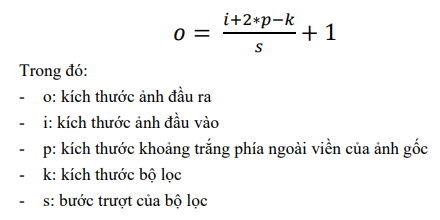
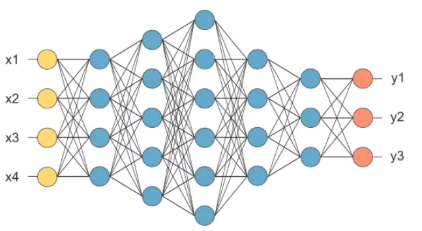
**Câu 1: Cấu trúc của một mạng CNN**

* Định nghĩa: Mạng cnn (convolutional neural network): hiểu đơn giản nó là tập hợp của các lớp Convolution chồng lên nhau và sử dụng các hàm kích hoạt phi tuyến tính (VD: Hàm ReLU, hàm tank) để kích hoạt trọng số cho các node, thông tin của node này là đầu vào của node kế tiếp (mô hình này được gọi là fully connected layer)
* Ý tưởng:
  1. Áp dụng các bộ lọc lên ảnh trước khi training mạng neural
  2. Sau khi đi qua các bộ lọc thì các đặc trưng của tấm ảnh sẽ trở lên nổi bật hơn
  3. Bộ lọc ở đây thực chất là tổ hợp của các phép nhân tích chập
  4. Kết hợp các bộ lọc lại với nhau bằng lớp tổng hợp pooling
  5. Lớp pooling sẽ gom một tập các điểm ảnh lại với nhau và cho ra một tập nhỏ hơn
* Cấu trúc của CNN:
  1. Lớp tích chập (Convolutional): Có nhiệm vụ trích xuất các đặc trưng của ảnh, output của convolutional layer đầu tiên là input của convolutional layer tiếp theo, convolutional layer bao gồm: input(dữ liệu đầu vào) và bộ lọc tích chập:
     + Bộ lọc: ví dụ một bộ lọc kích thước 3x3, di chuyển lần lượt qua từng vùng ảnh cho đến khi quét xong toàn bộ ảnh, tạo ra ảnh có kích thước nhỏ hơn hoặc bằng ảnh ban đầu. Kích thước này tùy theo khoảng trắng đc thêm ở viền
     + Tổng quan: Gỉa sử input của một convolutional layer là tensor kích thước H \* W \* D, Kernel kích thước F \* F \* D (kernel luôn có depth bằng depth của input và F là số lẻ), stride: S, padding: P.
     + => Output của layer là tensor 3 chiều có kích thước: () \*  () \* K
     + Output của convolutional layer sẽ đi qua các hàm kích hoạt trước khi trở thành input của convolutional layer kia
  2. Lớp kích hoạt phi tuyến tính (Rectified Linear Unit)
     + Lớp này được xây dựng với ý nghĩa đảm bảo tính phi tuyến của mô hình huấn luyện sau khi đã thực hiện một loạt các phép tính toán tuyến tính qua các lớp Tích chập. Lớp Kích hoạt phi tuyến nói chung sử dụng các hàm kích hoạt phi tuyến như ReLU hoặc sigmoid, tanh… để giới hạn phạm vi biên độ cho phép của giá trị đầu ra.
     + Phi tuyến tính có nghĩa là đầu ra không thể được sao chép từ tổ hợp tuyến tính của các đầu vào, nếu không có các hàm kích hoạt phi tuyến, thì mạng nơ-ron của chúng ta dù có nhiều lớp vẫn sẽ có hiệu quả như một lớp tuyến tính mà thôi
  3. Lớp tổng hợp (Pooling):
     + có chức năng tổng hợp các bộ lọc lại với nhau, pooling sẽ gom 1 tập các điểm ảnh và cho ra kích thước nhỏ hơn, làm giảm dần kích thước không gian để giảm số lượng tham số và tính toán của ảnh. Ví dụ maxpooling 2x2 thì sẽ chọn ra từ ảnh đầu vào giá trị điểm ảnh lớn nhất trong các nhóm điểm ảnh với kích thước 2x2.
     + Lớp lấy mẫu cũng sử dụng một cửa sổ trượt để quét toàn bộ các vùng trong ảnh tương tự như lớp Tích chập, và thực hiện phép lấy mẫu thay vì phép tích chập – tức là ta sẽ chọn lưu lại một giá trị duy nhất đại diện cho toàn bộ thông tin của vùng ảnh đó.
  4. Lớp kết nối đầy đủ (Fully-connected):
     + tất cả các điểm ảnh được kết nối đầy đủ với node trong lớp tiếp theo
     + output của layer cuối cùng sẽ được là phẳng thành vector và đưa vào một lớp được kết nối như một mạng nơ-ron.
     + Với FC layer được kết hợp với các tính năng lại với nhau để tạo ra một mô hình. Cuối cùng sử dụng softmax hoặc sigmoid để phân loại đầu ra.

**Câu 2: Lớp pooling để làm gì, maxpooling, average pooling là để làm gì**

* Ý nghĩa đầu tiên khi mà nói đến lớp pooling đó là: tổng hợp các điểm ảnh lại với nhau và cho ra tập nhỏ hơn tăng hiệu suất khi training
* Ý nghĩa quan trọng hơn là: làm giảm kích thước của ảnh những vẫn giữ lại được những đặc trưng nổi bật của ảnh, càng thu nhỏ thì những chi tiết quan trọng càng mất đi, càng hiển thị được ra cái chi tiết nổi bật mà có trong ảnh (VD: Khi hỏi trong ảnh cái gì là đặc trưng – thì chính là thằng nào là nổi bật nhất, khi giảm kích thước ảnh thì cái nào bé cái nào mờ nó sẽ bị xóa và giữ lại cái nổi bật từ đó nhận biết được cái nào nổi bật nhất)
* Max pooling là tính giá trị lớn nhất trong vùng đầu vào 2x2, như vậy qua lớp này sẽ giảm đi 1 nửa
  + Filter size: thường là ma trận 3x3 hoặc 5x5
  + Epochs: số lượng chu kì lặp đi lặp lại của CNN
  + Batch: để cải thiện việc tính toán trong CNN, ta sẽ dùng phương pháp chia nhỏ các hình ảnh ban đầu thành nhóm nhỏ hơn

**Câu 3: Tại sao lại phải trải ảnh ra thành mảng 1 chiều ví dụ ảnh 4x4 -> 1x16**

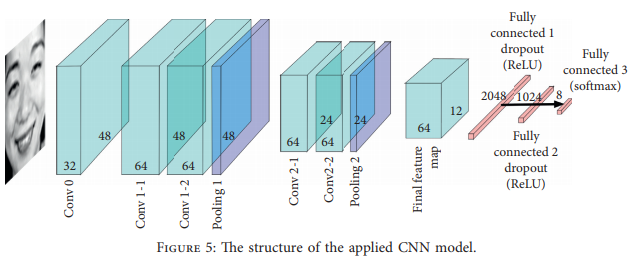
* Đây chỉ là một cách lưu trữ để cho dễ biểu diễn và xử lý hơn
* Phải trải ảnh từ 2 chiều sang 1 chiều vì: ví dụ nếu có 1 ảnh thì chỉ cần 4x4 thôi, nếu có 3 ảnh thì cần đến 3 cái ma trận 2 chiều thì n ảnh thì n mảng 2 chiều thì gom lại thành 3 chiều mỗi chiều 1 ảnh mỗi ảnh 2 chiều => khá phức tạp cho vc biểu diễn, còn 1 ảnh kéo thành 1 dòng thì chỉ cần 1 ma trận 2 chiều
* Nhờ đi khi training cho 1 mô hình chỉ nhận mảng 1 chiều thì ta hoàn toàn có thể chuyển ảnh 2 chiều về 1 chiều để nhét được vào trong model

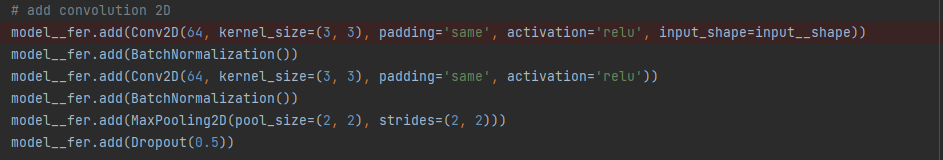
**Câu 4: Tại sao khi nhận ảnh vào thì và xử lý ảnh thì lại có 1 bước là chia 255**

* Về nguyên tắc mỗi một điểm ảnh n là 1 gtri nguyên nằm trong khoảng từ 0-255 và khi xử lý thì n phải dùng các phép toán để xly chứ ko xl số nguyên
* Các kỹ thuật xử lý bắt buộc ảnh phải là số thực nằm trong đoạn 0-1
* Ví dụ phép xử lý bình phương điểm ảnh, nếu dùng số nguyên là 200 ^2 lên là 4k nằm ngoài khoảng 255
* Do đó chuyển 200 về đoạn 0-1 thì 200/255 -> 0.78 do đó nếu muốn chuyển lại cũng chỉ nhân ngược lại -> 200 ^2 = 155
* Khi số nằm trong khoảng 0 1 thì n vẫn ở trong khoảng đó (biểu diễn thực), nếu 255 nó sẽ vượt qua 255 (biểu diễn nguyên)
* Dẫn đến sự tương đồng giữa các data point nhiều hơn

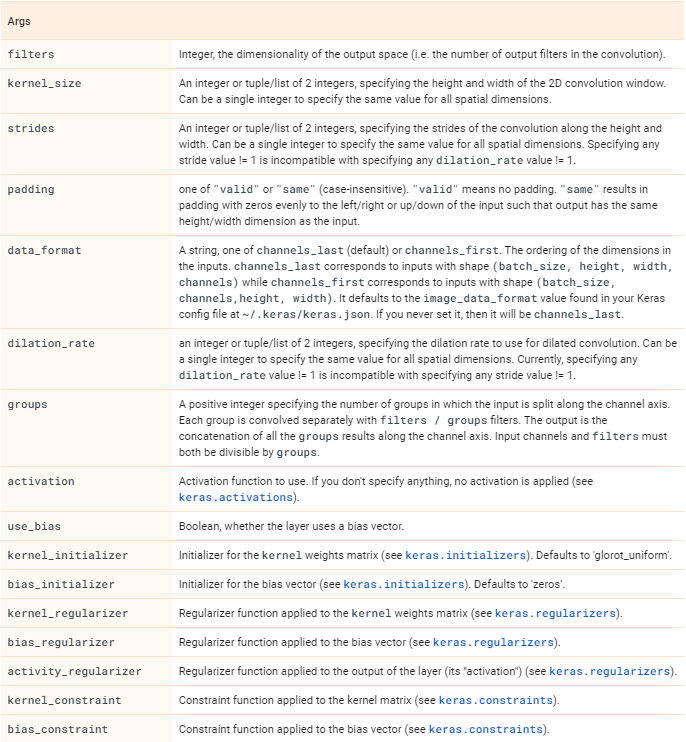
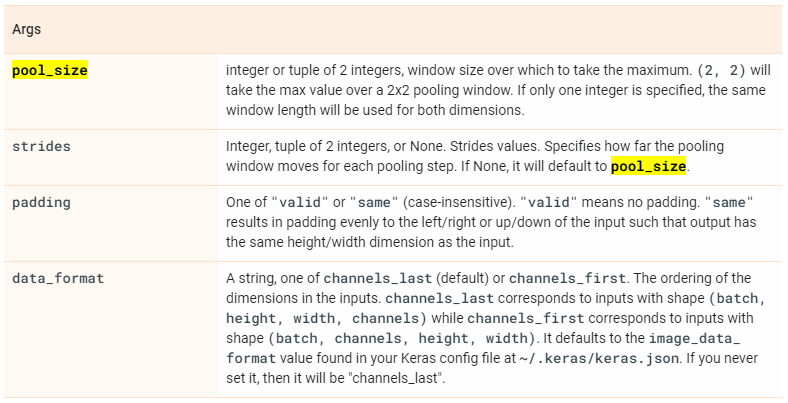
**Câu 5: Tại sao lại có 4 lớp?, sao không phải là 5 là 6?, ý nghĩa từng block?**

* Để trả lời cho câu hỏi bao nhiêu lớp thì nó còn phụ thuộc vào độ lớn vào số lượng của data, ví dụ có 1024 **params** thì nó sẽ phải tương đồng với **data point**, nếu ko tg đồng sẽ xảy ra trường hợp overfit (quá thông minh params >>> data point) hoặc underfit (params <<< data point ngu dốt mà lại học cái quá khó)
* Data point ở đây vi dụ như là các điểm ảnh

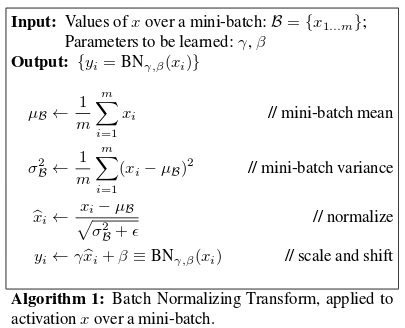


* Sau mỗi một lần đi qua các lớp conv thì kích thước của ảnh giảm nhưng số lượng channel sẽ tăng lên theo ý tưởng của VGG16
* Thường thì lớp đầu tiên sẽ là giữ lại hầu hết thông tin của ảnh, lớp đầu thường sẽ trích xuất viền và cạnh của ảnh
* Càng đi sâu các lớp về sau thì khó nhận ra đc là ảnh gì bởi vì: càng về sâu, các feature map biểu diễn những hình trừu tượng hơn
* Các thông tin về sau sẽ càng thưa thớt hơn. Điều này là hợp lý vì ban đầu phải giữ lại các cạnh của ảnh, ảnh nào mà chả phải có cạnh, càng về sâu thì phải biểu diễn những cái nó phức tạp hơn thứ mà không phải hình ảnh nào cũng c
* 1. Xây dựng bài toán 2. Chuẩn bị dữ liệu (dataset) 3. Xây dựng model 4. Định nghĩa loss function 5. Thực hiện backpropagation và áp dụng gradient descent để tìm các parameter gồm weight và bias để tối ưu loss function. 6. Dự đoán dữ liệu mới bằng model với các hệ số tìm được ở trên

**Câu 6: Lớp conv2d để làm gì?, các tham số trong đó là gì?, activation trong conv2d là gì?**

* Các tham số trong conv2d: ví dụ
* model\_\_fer.add(Conv2D(64, kernel\_size=(3, 3), padding='same', activation='relu', input\_shape=input\_\_shape))
* Ở đây có 64 kernel và mỗi kernel có kích thước là 3x3, giữ lại cạnh của anh, hàm kích hoạt phi tuyến relu
* Các thông số trong maxpooling2d:

**Câu 7: Lớp batchnormalization là để làm gì**

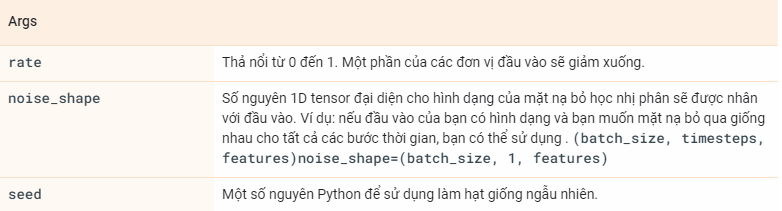
* ****Chuẩn hóa lại data để nó đồng nhất với input ban đầu
* **Batch Normalization** là một phương pháp hiệu quả khi training một mô hình mạng nơ ron. Mục tiêu của phương pháp này chính là việc muốn chuẩn hóa các feature (đầu ra của mỗi layer sau khi đi qua các activation) về trạng thái **zero-mean** với độ lệch chuẩn 1. Vậy hiện tượng ngược lại đó là **non-zero mean** có ảnh hưởng như thế nào đến việc training mô hình:
  + Thứ nhất: **Non zero mean** là hiện tượng dữ liệu không phân bố quanh giá trị 0, mà dữ liệu có phần nhiều giá trị lớn hơn không, hoặc nhỏ hơn không. Kết hợp với vấn đề high variance khiến dữ liệu trở nên có nhiều thành phần rất lớn hoặc rất nhỏ. Vấn đề này rất phổ biến khi training các mạng nơ ron với số layer sâu. Việc feature không phân phối trong những khoảng ổn định (giá trị to nhỏ thất thường) sẽ có ảnh hưởng đến quá trình tối ưu của mạng. Vì như chúng ta đã biết việc tối ưu một mạng nơ ron sẽ cần phải sử dụng đến tính toán đạo hàm. Giả sử như một công thức tính layer đơn giản là y = (Wx + b)*y*=(*Wx*+*b*) thì đạo hàm của y*y* theo w*w* có dạng: dy = dWx*dy*=*dWx*. Như vậy giá trị x*x* ảnh hưởng trực tiếp đến giá trị của đạo hàm. Do đó nếu x*x* mang các giá trị thay đổi không ổn định dẫn đến đạo hàm sẽ có thể bị quá lớn, hoặc quá nhỏ dẫn đến việc learning model không được ổn định. Và điều đó cũng đồng nghĩa với việc chúng ta có thể sử dụng các learning rate cao hơn trong quá trình training khi sử dụng Batch Normalization.
  + Thứ hai: **Batch normalization** có thể giúp chúng ta tránh được hiện tượng giá trị của x*x* rơi vào **khoảng bão hòa** sau khi đi qua các hàm kích hoạt phi tuyển. Vậy nên nó đảm bảo rằng không có sự kích hoạt nào bị vượt quá cao hoặc quá thấp. Điều này giúp cho các weights mà khi không dùng **BN** có thể sẽ không bao giờ được học thì nay lại được học bình thường. Điều này giúp chúng ta làm giảm đi sự phụ thuộc vào giá trị khởi tạo của các tham số.
  + Thứ ba: **Batch Normalization** còn có vai trò như một dạng của **regularization** giúp cho việc giảm thiểu overfiting. Sử dụng batch normalization, chúng ta sẽ không cần phải sử dụng quá nhiều dropput và điều này rất có ý nghĩa vì chúng ta sẽ không cần phải lo lắng vì bị mất quá nhiều thông tin khi dropout weigths của mạng. Tuy nhiên vẫn nên sử dụng kết hợp cả hai kĩ thuật này

**Câu 8: Vẽ và mô tả lại dữ liệu sau khi đi qua từng lớp từng câu lệnh**

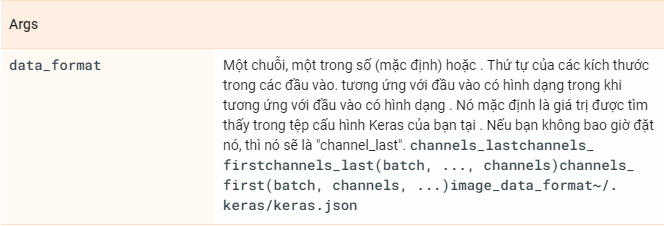
**Câu 9: Tại sao khi training model người ta lại phải tráo ảnh lộn xộn lên chứ ko để thành các phần riêng biệt như ban đầu?**

* Thường thì các bộ dữ liệu chia ra thành từng phần ví dụ: tập hợp những ảnh có cảm xúc buồn bã, … thì khi training model theo từng phần một mà các ảnh giống nhau => điều này làm cho ảnh hưởng đến tính tổng quát của ảnh thậm chí tráo hết nhãn

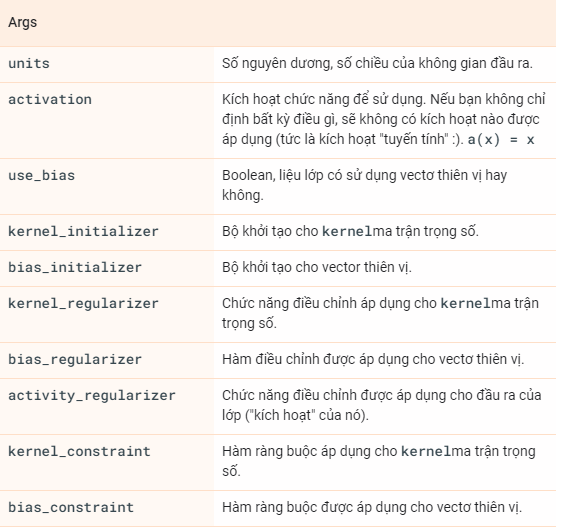
**Câu 10: Tìm hiểu về Dropout?**

* Hiểu đơn giản là, trong mạng neural network, kỹ thuật dropout là việc chúng ta sẽ bỏ qua một vài unit trong suốt quá trình train trong mô hình, những unit bị bỏ qua được lựa chọn ngẫu nhiên. Ở đây, chúng ta hiểu “bỏ qua - ignoring” là unit đó sẽ không tham gia và đóng góp vào quá trình huấn luyện (lan truyền tiến và lan truyền ngược).
* Về mặt kỹ thuật, tại mỗi giai đoạn huấn luyện, mỗi node có xác suất bị bỏ qua là 1-p và xác suất được chọn là p
* Vậy tại sao cần dropout tại sao cần loại đi một vài unit
* => Câu trả lời cho câu hỏi này là **để chống over-fitting**
* Trước đây trong ML, người ta thường sử dụng regularization để ngăn chặn overfiting, regularization làm giảm overfiting bằng cách thêm vào những loss function yếu tố **Laplacian** để features weights giảm đi sự phụ thuộc lẫn nhau

**Câu 11: Tìm hiểu về Flatter layer**

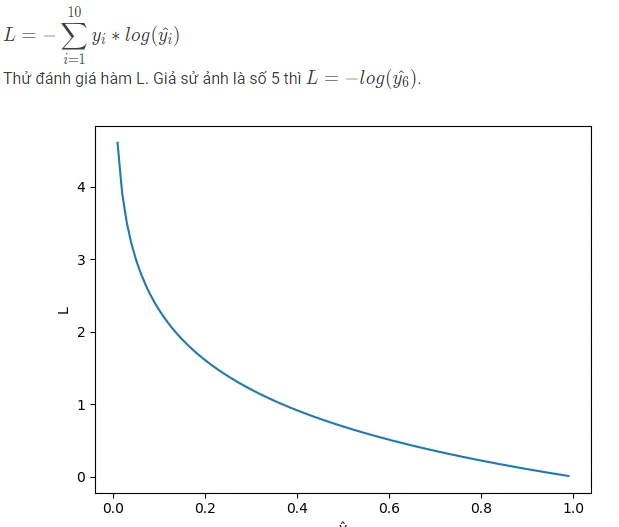
* Flatter đơn giản chỉ là một lớp dùng để làm phẳng dữ liệu rồi cung cấp cho mạng neural để xử lý

**Câu 12: Tìm hiểu về lớp Dense**

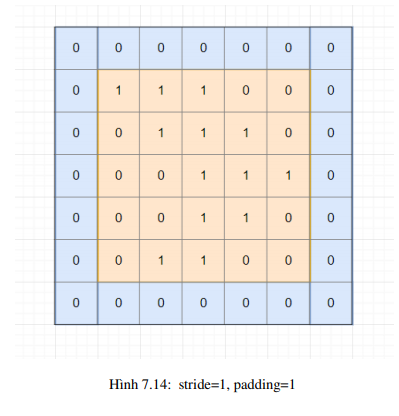
* Khái niệm: dense là một lớp trong cnn được kết nối sâu, có nghĩa là: mỗi một neuron trong lớp dense sẽ nhận input đầu vào từ tất cả các lớp trước đó
  + lớp này sẽ thực hiện nhân các vector ma trận, các giá trị sử dụng trong ma trận thật ra chỉ là các tham số truyền vào có thể training và cập nhật được với sự trợ giúp của lan truyền ngược
  + đầu ra của dense layer là một vecto m chiều. Do đó, về cơ bản thì dense làm thay đổi kích thước của một vector. Các lớp dense cũng hoạt động theo các cách như xoay, chia tỷ lệ, dịch trên vector.
* Thông số truyền vào trong tensorflow

**Câu 13: Compile**

* Ý nghĩa: cấu hình mô hình để tiến hành đào tạo nhóm các lớp thành một đối tượng với các tính năng huấn luyện và suy luận. Để đào tạo một mô hình với fit() , bạn cần chỉ định một hàm mất mát, một trình tối ưu hóa và tùy chọn, một số chỉ số cần theo dõi. Bạn chuyển chúng vào mô hình dưới dạng đối số cho phương thức compile() :
* Các tham số truyền vào:
  + Loss: loss function, ví dụ trong bài đang sử dụng là categorical\_crossentropy, giải thích về categorical\_crossentropy: Loss function giống như một hình thức để bắt model đóng phạt mỗi lần nó dự đoán sai, và số mức phạt tỉ lệ thuận với độ trầm trọng của sai sót, thì có rất nhiều cách xây dựng hàm loss function ví dụ như categorical\_crossentropy thì sẽ có cách xây dựng sau hàm L nhỏ khi giá trị model dự đoán gần đúng với giá trị thật và rất lớn khi model dự đoán sai, hay nói cách khác L càng nhỏ thì model dự đoán càng gần với giá trị thật => bài toán tìm L nhỏ nhất
  + Optimizer: (trả lời riêng ở câu hơi bên dưới)
  + Metrics: là một hàm chức năng đánh giá hiệu suất của mình hình, các tham số tương tự như loss function



**Câu 14: Stride là gì**

* Khi thực hiện padding ta thu đc ma trận mới có cùng kích thước, ta gọi stride là 1
* Tuy nhiên nếu stride=k (k > 1) thì ta chỉ thực hiện phép tính convolution trên các phần tử x1+i\*k, 1+j\*k. Ví dụ k = 2. => hiểu đơn giản là từ bắt đầu vị trí x11 thì nó sẽ nhảy k bước theo chiều dọc và theo chiều ngang cho đến hết ma trận
* Hiểu đơn giản là bắt đầu từ vị trí x11 sau đó nhảy k bước theo chiều dọc và ngang cho đến hết ma trận X. Kích thước của ma trận Y là 3\*3 đã giảm đi đáng kể so với ma trận X.

**Câu 15: chuyển ảnh rgb về ảnh đa mức xám (đen trắng)**

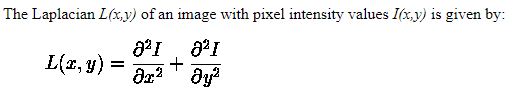
**Câu 16: Loss function và activaction là gì**

* Loss function hay còn gọi là hàm mất mát, thể hiện một mối quan hệ giữa y\* (là kết quả dự đoán của model) và y (là giá trị thực tế). Ví dụ ta có hàm loss như sau: f(y) = (y\* - y)^2. Khi đó người ta đưa vào hàm loss function này mục đích là để tối ưu model của mình sao cho tốt nhất, hay cũng dùng để đánh giá độ tốt của model , y\* (là kết quả dự đoán của model) càng gần y (là giá trị thực tế) thì càng tốt. Tức là dựa vào loss function, khi đó chúng ta có thể tính ra gradient descent để tối ưu loss function càng về gần 0 càng tốt
* Hàm kích hoạt (activation function) mô phỏng tỷ lệ truyền xung qua axon của một neuron thần kinh, là những hàm phi tuyến được áp dụng vào đầu ra của các nơ-ron trong tầng ẩn của một mô hình mạng ( có nhiệm vụ là chuẩn hoá output của neura) và được sử dụng làm input data cho tầng tiếp theo
* **Chuyện gì sẽ xảy ra nếu không có các hàm phi tuyến này ?** Hãy tưởng tượng rằng thay vì áp dụng 1 hàm phi tuyến, ta chỉ áp dụng 1 hàm tuyến tính vào đầu ra của mỗi neuron. Vì phép biến đổi không có tính chất phi tuyến, việc này không khác gì chúng ta thêm một tầng ẩn nữa vì phép biến đổi cũng chỉ đơn thuần là nhân đầu ra với các weights. Với chỉ những phép tính đơn thuần như vậy, trên thực tế mạng neural sẽ không thể phát hiện ra những quan hệ phức tạp của dữ liệu (ví dụ như: dự đoán chứng khoán, các bài toán xử lý ảnh hay các bài toán phát hiện ngữ nghĩa của các câu trong văn bản). **Nói cách khác nếu không có các activation functions, khả năng dự đoán của mạng neural sẽ bị giới hạn và giảm đi rất nhiều, sự kết hợp của các activation functions giữa các tầng ẩn là để giúp mô hình học được các quan hệ phi tuyến phức tạp tiềm ẩn trong dữ liệu.**
* Tại sao hàm activation phải non-linear? Điều gì xảy ra nếu hàm linear activation được sử dụng? Các activation function phải là nonlinear (phi tuyến), vì nếu không, nhiều layer hay một layer cũng là như nhau. Ví dụ với hai layer trong Hình 2, nếu activation function là một hàm linear (giả sử hàm f(s) = s) thì cả hai layer có thể được thay bằng một layer với ma trận hệ số W = W1xW2 (tạm bỏ qua hệ số bias)

**Câu 17: Tìm hiểu về optimizer**

* Trước khi đi sâu vào vấn đề thì cần hiểu thế nào là **thuật toán tối ưu (optimizers)**. Về cơ bản, thuật toán tối ưu là cơ sở để xây dựng mô hình neural network với mục đích "học " được các features ( hay pattern) của dữ liệu đầu vào, từ đó có thể tìm 1 cặp weights và bias phù hợp để tối ưu hóa model. Nhưng vấn đề là "học" như thế nào? Cụ thể là weights và bias được tìm như thế nào! Đâu phải chỉ cần random (weights, bias) 1 số lần hữu hạn và hy vọng ở 1 bước nào đó ta có thể tìm được lời giải. Rõ ràng là không khả thi và lãng phí tài nguyên! Chúng ta phải tìm 1 thuật toán để cải thiện weight và bias theo từng bước, và đó là lý do các thuật toán optimizer ra đời.
* Các thuật toán tối ưu: Gradient Descent(GD), Adam: Adam là sự kết hợp của Momentum và RMSprop . Nếu giải thích theo hiện tượng vật lí thì Momentum giống như 1 quả cầu lao xuống dốc, còn Adam như 1 quả cầu rất nặng có ma sát, vì vậy nó dễ dàng vượt qua local minimum tới global minimum và khi tới global minimum nó không mất nhiều thời gian dao động qua lại quanh đích vì nó có ma sát nên dễ dừng lại hơn.

**Một số khái niệm liên quan**

* **Gradient** nghĩa là độ dốc. Trong xử lý ảnh, độ dốc (tức gradient) đang nói đến ở đây chính là độ dốc về mức sáng. Hay nói cách khác chính là sự thay đổi các giá trị pixel trong ảnh.
* **Laplacian** là thước đo [đẳng hướng](https://homepages.inf.ed.ac.uk/rbf/HIPR2/isotrop.htm) 2-D của [đạo hàm không gian](https://homepages.inf.ed.ac.uk/rbf/HIPR2/spatdom.htm) thứ 2 của một hình ảnh. **Laplacian** của một hình ảnh làm nổi bật các vùng thay đổi cường độ nhanh chóng và do đó thường được sử dụng để https://homepages.inf.ed.ac.uk/rbf/HIPR2/mote.gifphát hiện cạnh (xem bộ [phát hiện cạnh cắt bằng không](https://homepages.inf.ed.ac.uk/rbf/HIPR2/zeros.htm) ). **Laplacian** thường được áp dụng cho một hình ảnh lần đầu tiên được làm mịn bằng một cái gì đó xấp xỉ với [bộ lọc làm mịn **Gauss**](https://homepages.inf.ed.ac.uk/rbf/HIPR2/gsmooth.htm) để giảm độ nhạy của nó với nhiễu, và do đó hai biến thể sẽ được mô tả cùng nhau ở đây. Toán tử thường lấy một hình ảnh graylevel duy nhất làm đầu vào và tạo ra một hình ảnh graylevel khác làm đầu ra.
* Dropout là một kỹ thuật khác, một cách tiếp cận khác để regularization trong mạng neural netwoks.
* Kỹ thuật dropout được thực hiện như sau:
* **Trong pha train**: với mỗi hidden layer, với mỗi trainning sample, với mỗi lần lặp, chọn ngẫu nhiên p phần trăm số node và bỏ qua nó (bỏ qua luôn hàm kích hoạt cho các node bị bỏ qua).
* Trong pha test: Sử dụng toàn bộ activations, nhưng giảm chúng với tỷ lệ p (do chúng ta bị miss p% hàm activation trong quá trình train).
* Dropout ép mạng neural phải tìm ra nhiều robust features hơn, với đặc điểm là chúng phải hữu ích hơn, tốt hơn, ngon hơn khi kết hợp với nhiều neuron khác.
* Dropout đòi hỏi phải gấp đôi quá trình huấn luyện để đạt được sự hội tụ. Tuy nhiên, thời gian huấn luyện cho mỗi epoch sẽ ít hơn.
* Với H unit trong mô hình, mỗi unit đều có xác xuất bị bỏ qua hoặc được chọn, chúng ta sẽ có 2^H mô hình có thể có. Trong pha test, toàn bộ network được sử dụng và mỗi hàm activation được giảm đi với hệ số p
* Theo nghiên cứu và thực nghiệm chỉ ra rằng Giá trị dropout tốt nhất là 0.2 và cần epochs ít nhất là trên 20