

Redes Neuronales Artificiales

Implementación teórica softmax y entropía cruzada

Implementación teórica de una red neuronal artificial usando la función sigmoide como función de activación de las capas ocultas, softmax para la capa de salida, y la función de coste/pérdida la entropía cruzada

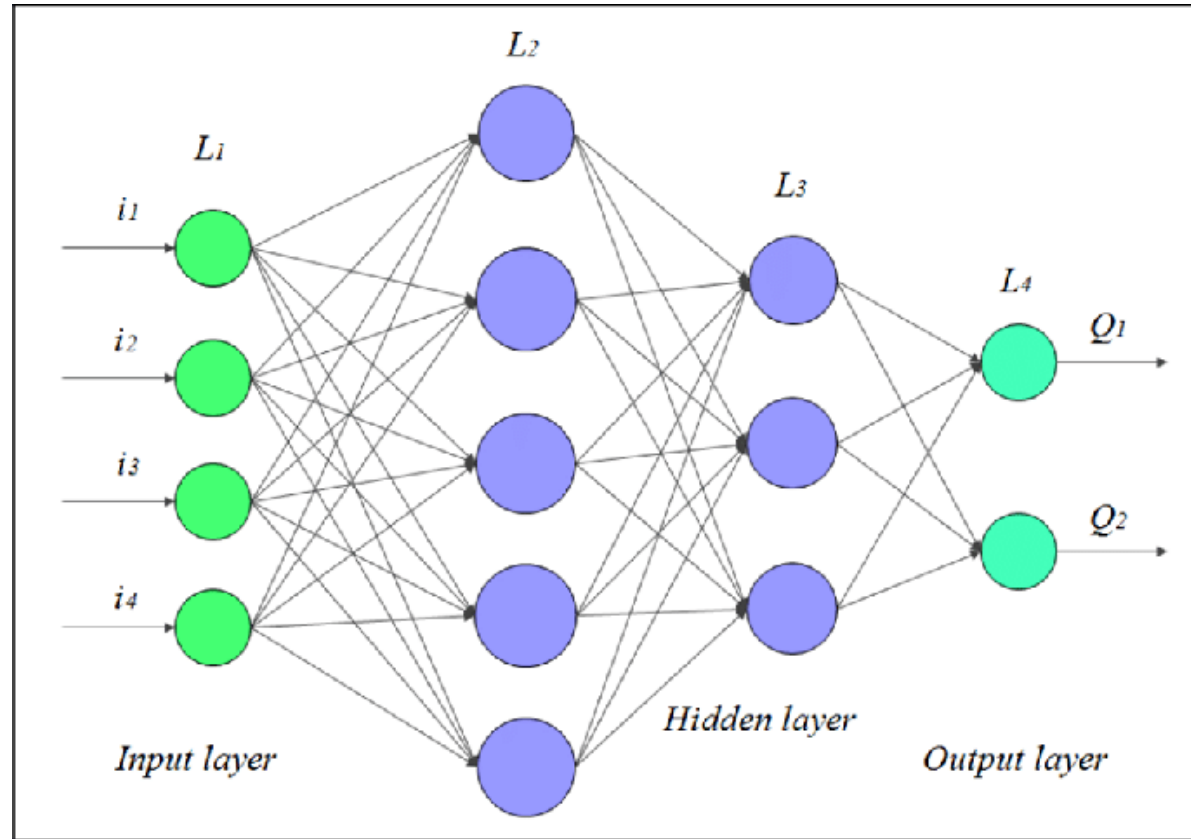
Antes de empezar

La entropía cruzada como función de pérdida junto con la función softmax sirven para problemas de clasificación. Los valores de salida con los que se entrenan estas redes son con un vector siendo 0 todos sus elementos de la clase que no pertenece al ejemplo, y 1 uno de esos elementos cuya clase si pertenece a ese ejemplo

Por ejemplo, tenemos las clases perro, gato y pájaro en este orden, el vector de salida de perro esperado es $[1, 0, 0]$, el de gato $[0, 1, 0]$, y del pájaro $[0, 0, 1]$

En teoría, debería ser posible clasificar cada ejemplo como varias clases, es decir, que pueda haber más de un elemento con un 1, pero lo normal es utilizarlo para estos casos

Nos basaremos en:



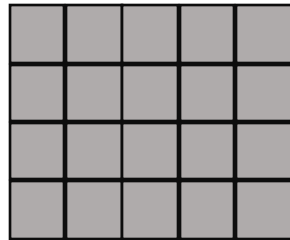
Una red neuronal artificial con esta estructura. Tiene **4 capas**.
Las capas ocultas tienen la función de activación sigmoide, y la capa de salida, softmax

Representación red matricial:

AL(L1) 4



W(L1,L2) 4x5



ZL(L2) 5



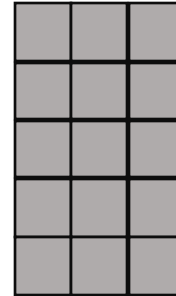
AL(L2) 5



B(L2) 5



W(L2,L3) 5x3



ZL(L3) 3



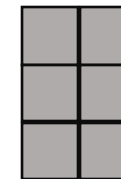
AL(L3) 3



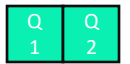
B(L3) 3



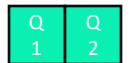
W(L3,L4) 3x2



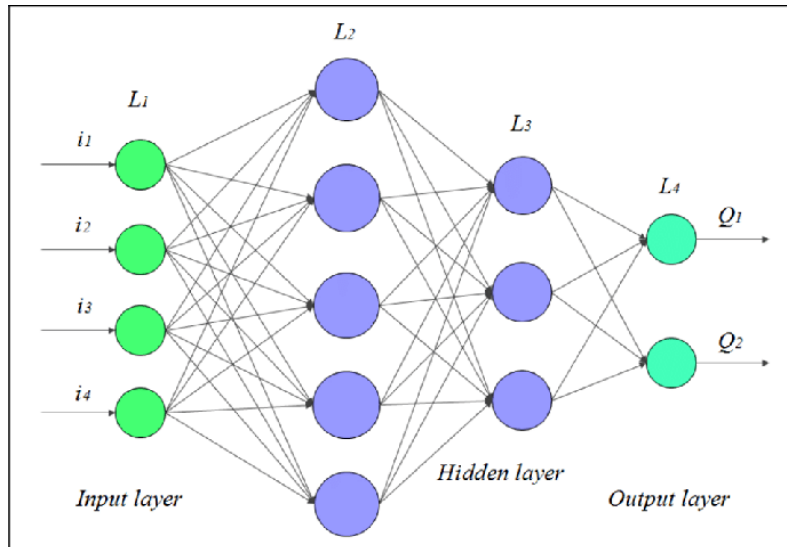
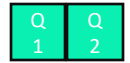
ZL(L4) 2



AL(L4) 2



B(L4) 2



B: vector donde se guardan los valores del bias de la neurona en la capa actual. Se suele inicializar con todos sus valores iguales a 0

W: matriz con los valores de los pesos de las aristas. Se inicializa con valores aleatorios, hay distintas formas de inicializarlos, pero supongamos números entre 3.0 y -3.0

ZL: vector donde se almacena o los valores de entrada (sólo primera capa) o el resultado para cada neurona de la capa actual el calcular el sumatorio del valor de cada neurona de la capa anterior por el peso de la arista que conecta la neurona de la capa actual, más el bias de dicha neurona (resto de capas). No importan sus valores iniciales

AL: vector donde se almacena el resultado de aplicar a cada elemento en ZL la función de activación de dicha capa. No importan sus valores iniciales

Representación red como listas

```
Red.numero_capas = 4
```

```
Red.zl = [
```

```
    [ vector 5 elementos sin importar sus valores ],
```

```
    [ vector 3 elementos sin importar sus valores ],
```

```
    [ vector 2 elementos sin importar sus valores ]
```

```
]
```

```
Red.al = [
```

```
    [ vector 4 elementos sin importar sus valores ],
```

```
    [ vector 5 elementos sin importar sus valores ],
```

```
    [ vector 3 elementos sin importar sus valores ],
```

```
    [ vector 2 elementos sin importar sus valores ]
```

```
]
```

```
Red.bias = [
```

```
    [ vector 5 elementos todos sus valores 0 ],
```

```
    [ vector 3 elementos todos sus valores 0 ],
```

```
    [ vector 2 elementos todos sus valores 0 ]
```

```
]
```

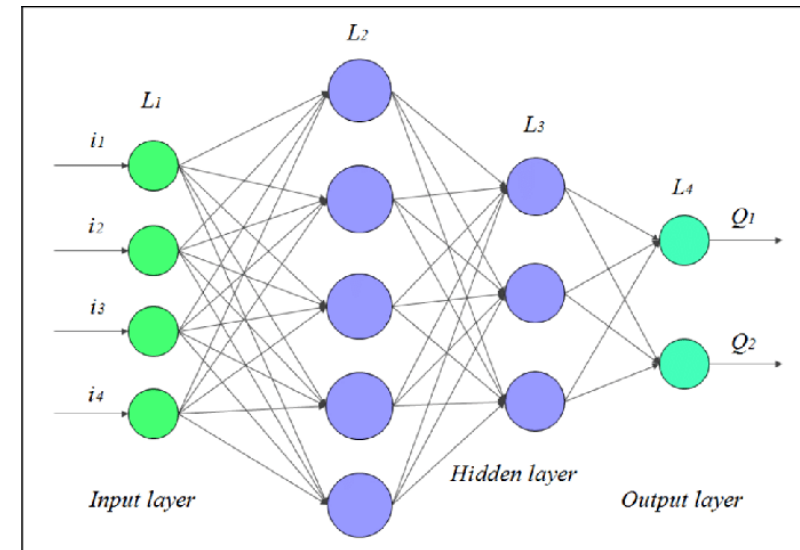
```
Red.pesos = [
```

```
    [ matriz 4x5 elementos todos sus valores aleatorios entre (3.0, -3.0) ],
```

```
    [ matriz 5x3 elementos todos sus valores aleatorios entre (3.0, -3.0) ],
```

```
    [ matriz 3x2 elementos todos sus valores aleatorios entre (3.0, -3.0) ]
```

```
]
```



Propagación hacia detrás

Acción que permite a la red pasar unos valores de entrada que deseamos evaluar en su extremo izquierdo y, en su extremo derecho, la capa de salida, esperamos recibir unos valores concretos en cada uno de los nodos de esta capa de salida

Recordemos que la red tiene 4 capas, aunque se consideran como neuronas solamente los nodos de las capas intermedias (ocultas) y la capa a la derecha (salida), ya que los nodos de la capa a la izquierda (entrada) no modifican los valores de entrada

Aparte, recordar que las capas que se consideran neuronas aplican a los resultados que computan una función concreta, llamada función de activación, y nosotros consideraremos que dicha función es la sigmoide. Esta función solamente toma un valor como entrada y lo modifica realizando esta operación:

sigmoid(x):
devolver $1 / (1 + e^{-x})$

Aparte, en la capa de salida, tenemos otra función de activación, softmax, cuya fórmula es:
Tal y como veremos, se calcula con todos los valores de las neuronas de la capa de salida

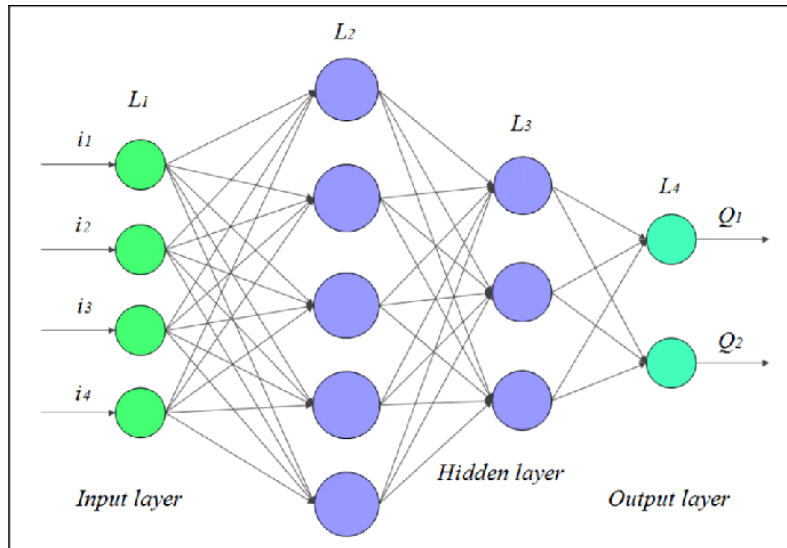
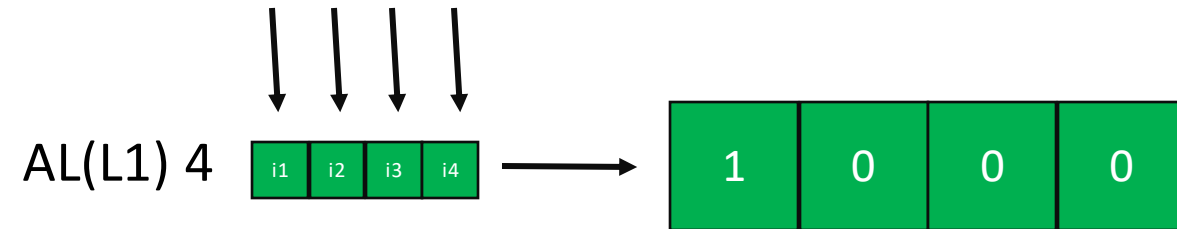
Softmax :

\mathbf{s} = vector of scores

$$\sigma(\mathbf{s})_c = \frac{\exp(s_c)}{\sum_{c'} \exp(s_{c'})}$$

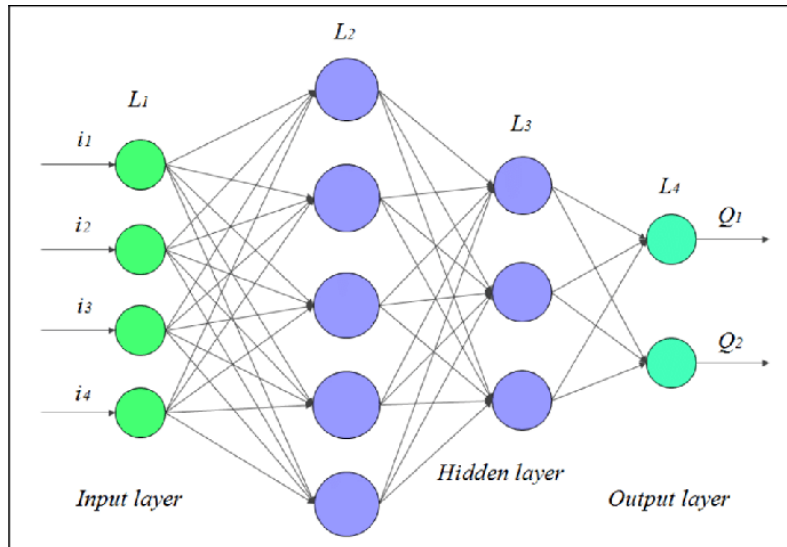
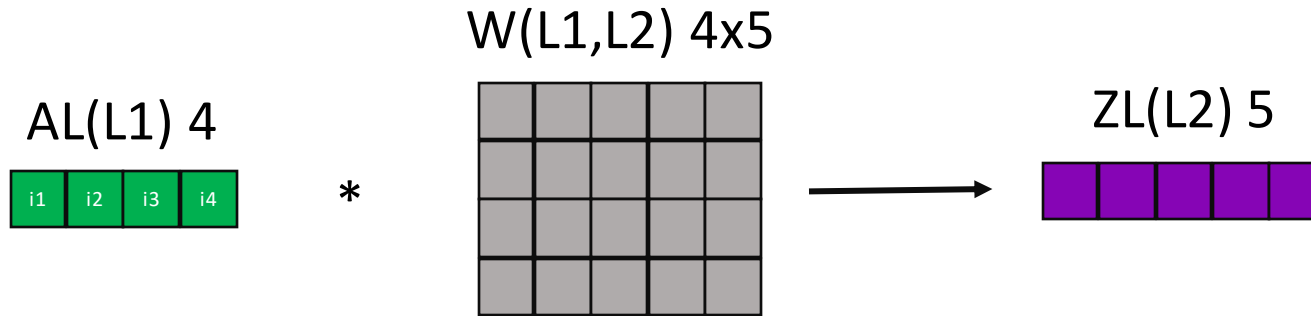
Propagación hacia delante

VALORES DE ENTRADA(LISTA) = [1, 0, 0, 0]



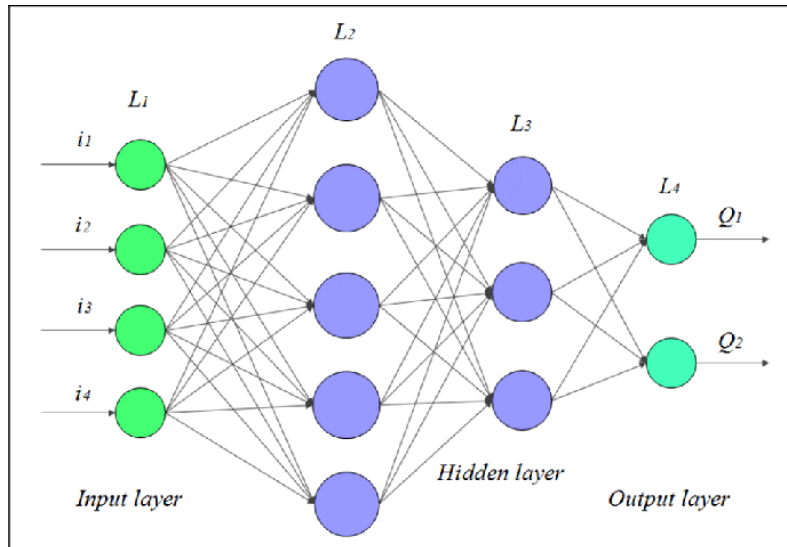
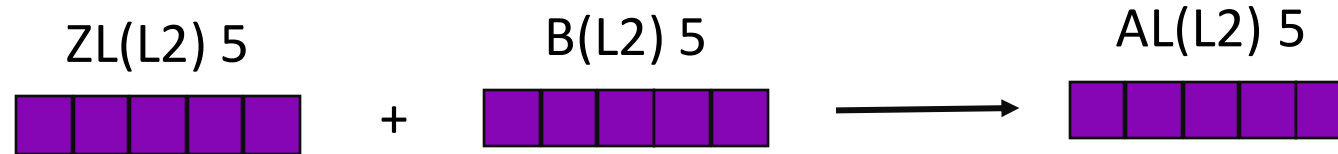
Pasamos los valores de entrada deseados a la primera lista al de la red

Propagación hacia delante



Hasta que no lleguemos a la capa final, calculamos el producto del vector al de la capa actual como matriz de 1 fila por la matriz de pesos de la capa actual. El resultado lo pasamos al vector z_l de la capa siguiente. Empezamos desde la capa de entrada

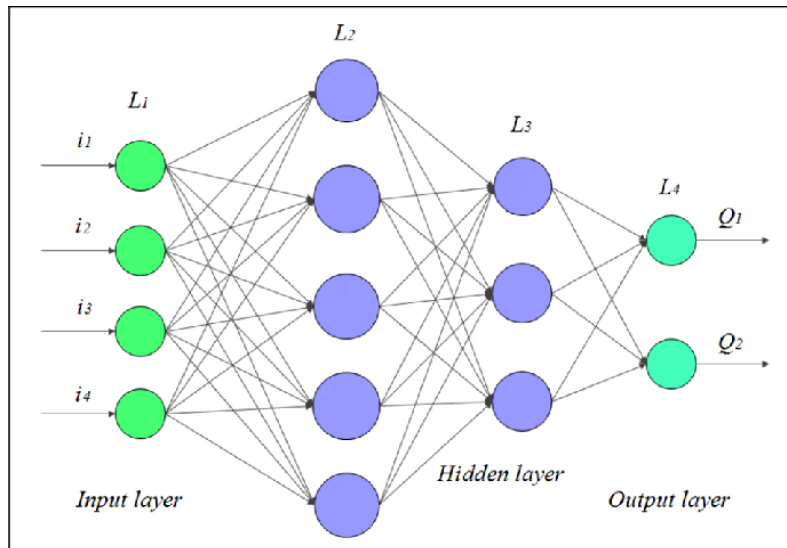
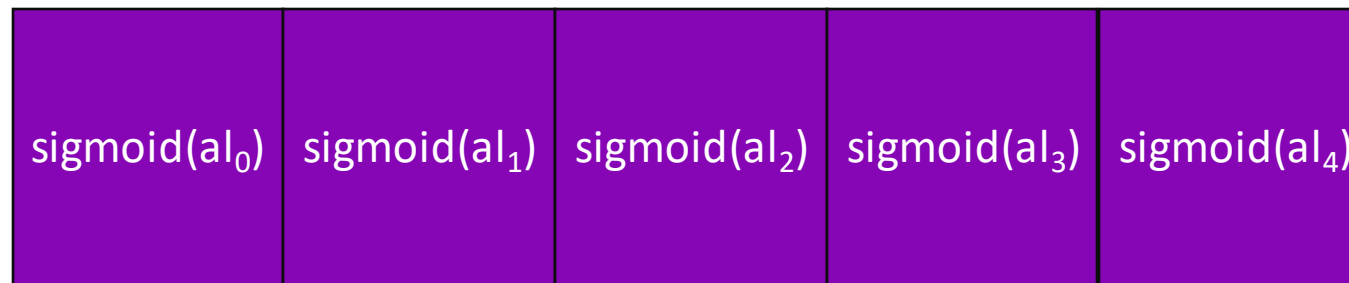
Propagación hacia delante



Ahora, sumamos el vector z_l de la capa posterior con el vector b de la capa posterior, y copiamos el resultado al vector al de la capa posterior

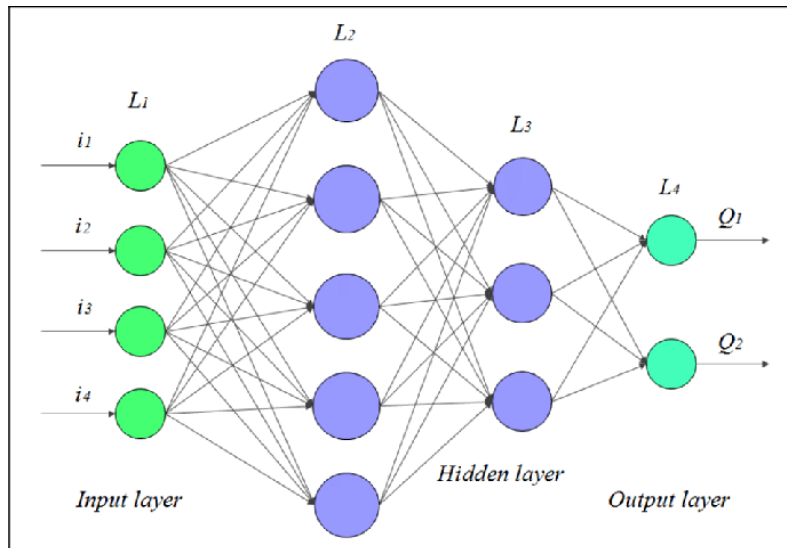
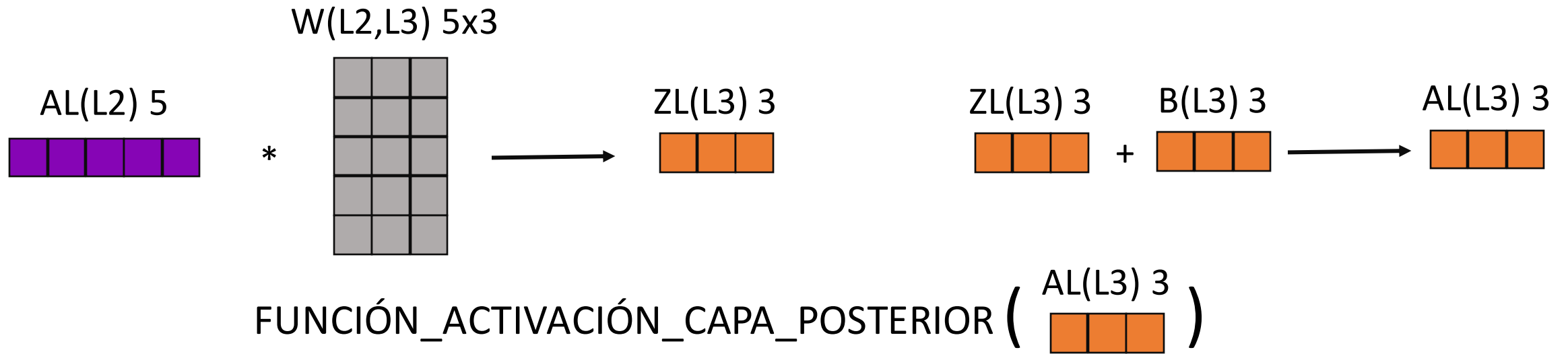
Propagación hacia delante

FUNCIÓN_ACTIVACIÓN_CAPA_POSTERIOR ($\overset{\text{AL}(L2) \ 5}{\begin{array}{|c|c|c|c|c|}\hline & & & & \end{array}}$)



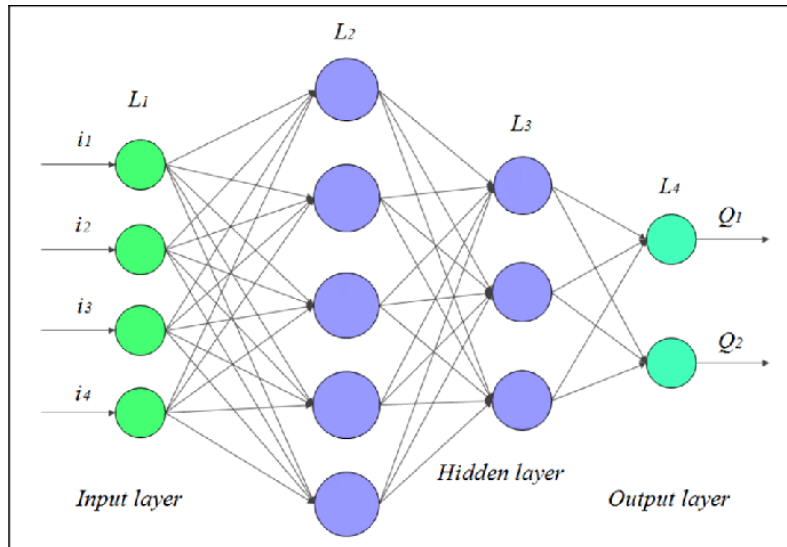
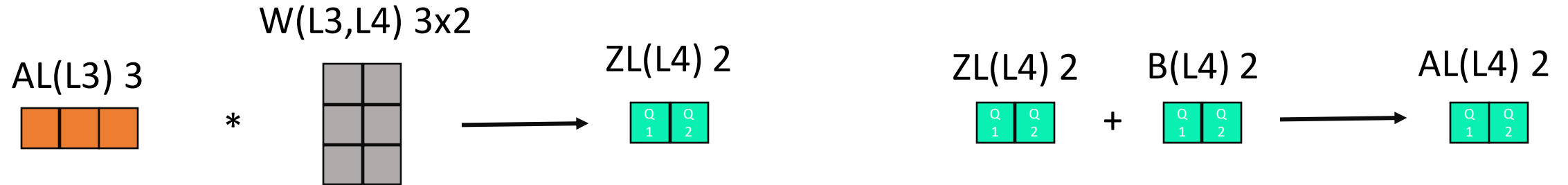
Ahora, aplicamos la función correspondiente a la capa posterior a cada elemento del vector a_l de la capa posterior. En este ejemplo, consideramos utilizar la función sigmoide para todas las capas, y denotamos cada elemento de a_l como a_i , siendo i la posición de cada elemento en el vector.

Propagación hacia delante



Nos movemos a la siguiente capa y repetimos el proceso

Propagación hacia delante



Al final, llegamos a la capa de salida, repetimos el mismo proceso que antes. Aplicaremos la función softmax en la siguiente diapositiva

Propagación hacia delante

Softmax :

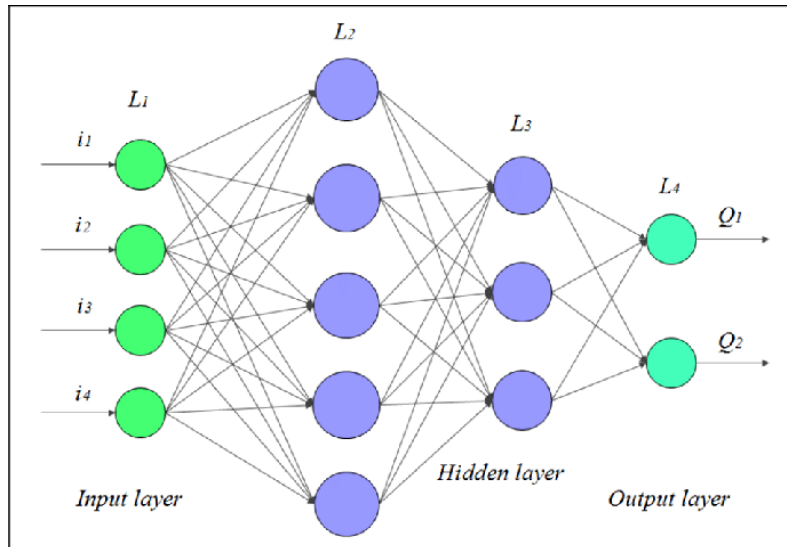
s = vector of scores

$$\sigma(s)_c = \frac{\exp(s_c)}{\sum_{c'} \exp(s_{c'})}$$

$$\sum_{c'} \exp(s_{c'}) = e^{Q_1} + e^{Q_2}$$

$$\text{FUNCIÓN_ACTIVACIÓN_CAPA_SALIDA} \left(\begin{matrix} \text{AL(L4) 2} \\ Q_1 \quad Q_2 \end{matrix} \right) =$$

$$\frac{e^{Q_1}}{\sum_{c'} \exp(s_{c'})} \quad \frac{e^{Q_2}}{\sum_{c'} \exp(s_{c'})}$$



Primero calculamos el denominador de la operación, que es común a todas las neuronas. Simplemente es el sumatorio de e (número de Euler) elevado al valor de salida de la propia neurona

Luego, para cada neurona, cambiamos su valor a e (número de Euler) elevado al valor de salida de la propia neurona, dividido entre el sumatorio anteriormente calculado

Ojo! El valor para cada neurona debe ser un número decimal acotado entre 0 y 1, y el sumatorio de los valores de las neuronas debe siempre ser igual a 1

Propagación hacia delante

Vector_a_matriz_1_fila: recibe como parámetro 1 vector, y devuelve 1 matriz de una fila y tantas columnas como nº elementos del vector

Matriz_1_fila_a_vector: recibe como parámetro 1 matriz de 1 fila, y devuelve un vector equivalente a la fila de dicha matriz

Producto_matrices: recibe como parámetros dos matrices, en orden, m1 y m2, y devuelve otra matriz como el producto $m1 * m2$

Reemplazar_valores_vectores: recibe como parámetros dos vectores, en orden, v1 y v2, copia los valores de v2 en v1

Suma_valores_vectores: recibe como parámetros dos vectores, en orden, v1 y v2, suma los valores de v2 a v1

Aplicar_funcion_vector: recibe como parámetros, en orden, un vector v1 y una función f, y pasa el valor de cada elemento x de v1 a f(x)

Propagación hacia delante

sigmoid(x):

devolver $1 / (1 + e^{-x})$

softmax(x, all_output_values):

denominador = sumatorio([e^y para y cada valor de all_output_values])

devolver $e^x / \text{sumatorio}$

propagacion_hacia_delante(vector_valores_entrada, red, funcion_capas_ocultas, funcion_capa_final):

Reemplazar_valores_vectores(red.al[0], vector_valores_entrada)

Índice i desde 0 hasta Red.numero_capas-1 (incluidos):

Reemplazar_valores_vectores(red.zl[i], Matriz_1_fila_a_vector(Producto_matrices(Vector_a_matriz_1_fila(red.al[i]), red.pesos[i])))

Suma_valores_vectores(red.zl[i], red.bias[i])

Reemplazar_valores_vectores(red.al[i], red.zl[i])

Si $i < \text{Red.numero_capas}-2$:

Aplicar_funcion_vector(red.al[i], funcion_capas ocultas)

Sino:

Aplicar_funcion_vector(red.al[i], funcion_capa final(all_output_values = red.al[i]))

Ejemplo de ejecución:

Para cada ejemplo en conjunto_entrenamiento:

propagacion_hacia_delante(ejemplo.entrada, Red, sigmoid, softmax) #el primero por ejemplo sería el [1, 0, 0, 0] visto antes

Propagación hacia atrás

Acción que permite a la red aprender de los errores que comete al realizar predicciones, modificando los valores de los pesos y biases

Existen algunas diferencias entre algunas redes y otras (a veces no se aplica la función de activación en la capa de salida de la red y las derivadas parciales en dicha capa no la incluyen, en otros casos se calcula el coste como un sumatorio de todos los errores de todos los ejemplos, mientras que en otros casos se calcula el vector gradiente separado para cada uno y luego se suma una media aritmética de dichos errores, en otros casos, se calculan las derivadas parciales del coste con respecto al valor esperado en vez de con respecto el valor predicho, etc.), pero en esencia, comparten las mismas ideas de cómo se calcula el vector gradiente

Aparte, existen diferentes funciones de pérdida para calcular el error cometido por un nodo en la capa de salida, acompañadas de una función de coste para evaluar el error total de la red (media aritmética de todas las funciones de pérdida). Nosotros utilizaremos **la entropía cruzada junto con la salida de softmax como función de coste y de pérdida**.

En esta diapositiva, se siguen los pasos de este vídeo: <https://www.youtube.com/watch?v=tleHLnjs5U8&t=3s>

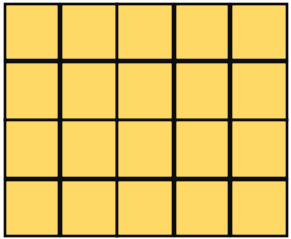
Y las derivadas parciales de la entropía cruzada con softmax en la capa de salida (15:45): <https://www.youtube.com/watch?v=xBEh66V9gZo>

Básicamente, la idea es primero calcular un vector que incluye todos los valores que servirán para ajustar los valores de los pesos y los biases de la red, llamado vector gradiente, uno para cada uno de los ejemplos del conjunto de entrenamiento. Después, se sumarán los valores en las mismas posiciones de los vectores gradiente calculados para todos los ejemplos del conjunto de entrenamiento.

Finalmente, cada valor se dividirá entre el número de ejemplos del conjunto de entrenamiento, y se multiplicará por un hiperparámetro del entrenamiento llamado tasa de aprendizaje, el cual suele ser un número decimal entre 1 y valores cercanos al 0, que permite controlar la velocidad a la que nos aproximamos a un punto dentro de la función de error

Representación v. gradiente matricial:

EW(L1,L2) 4x5



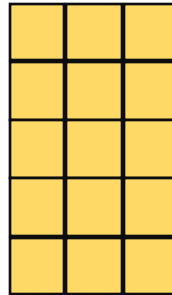
EB(L2) 5



EC(L2) 5



EW(L2,L3) 5x3



EB(L3) 3



EC(L3) 3



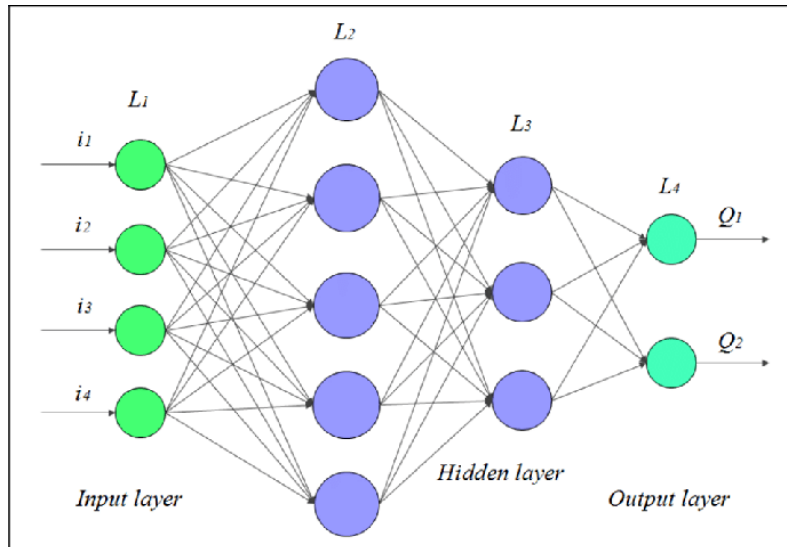
EW(L3,L4) 3x2



EB(L4) 2



EC(L4) 2



EB: vector donde se guardan los valores de error del bias de la neurona en la capa actual. Sus valores se calcularán al realizar 1 vez la propagación hacia atrás

EC: vector donde se guardan los valores de error de dicha capa con respecto a la función de coste, calculado en la iteración anterior. Sus valores se calcularán al realizar 1 vez la propagación hacia atrás

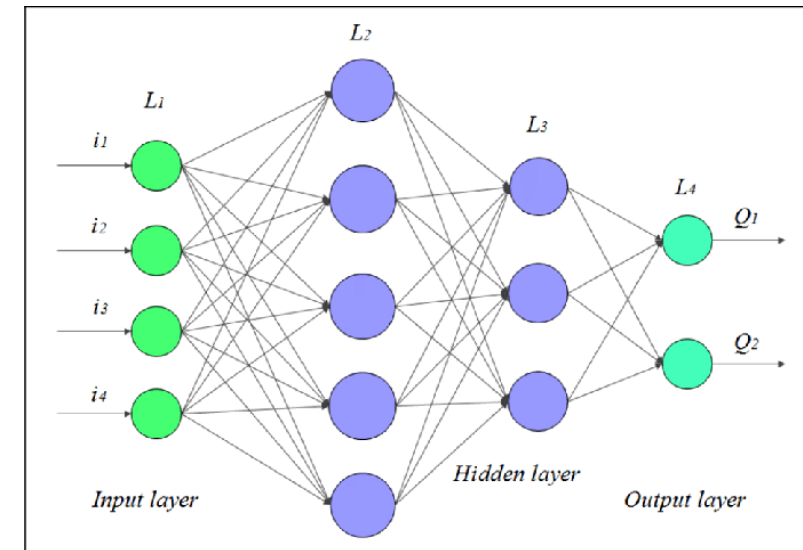
EW: matriz con los valores de los errores de los pesos de las aristas. Sus valores se calcularán al realizar 1 vez la propagación hacia atrás

Nota: no se almacena ni el error del bias ni el error de la capa, para la capa de entrada de la red

Representación v. gradiente como listas

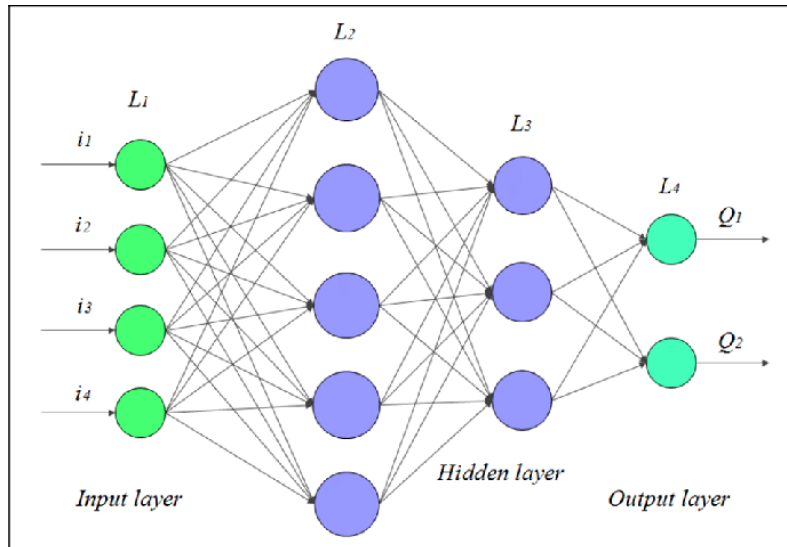
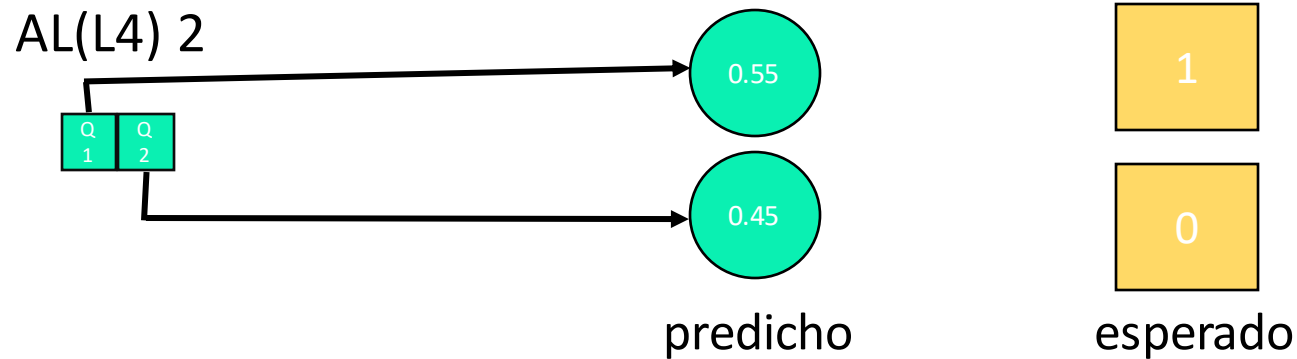
```
VGrad.err_bias = [  
    [ vector 5 elementos todos sus valores iguales a 0 ],  
    [ vector 3 elementos todos sus valores iguales a 0 ],  
    [ vector 2 elementos todos sus valores iguales a 0 ]  
]  
VGrad.err_capa = [  
    [ vector 5 elementos todos sus valores iguales a 0 ],  
    [ vector 3 elementos todos sus valores iguales a 0 ]  
]  
VGrad.err_pesos = [  
    [ matriz 4x5 elementos todos sus valores iguales a 0 ],  
    [ matriz 5x3 elementos todos sus valores iguales a 0 ],  
    [ matriz 3x2 elementos todos sus valores iguales a 0 ]  
]
```

Nota: no se almacena el error del bias para la capa de entrada



Propagación hacia atrás

`propagacion_hacia_delante([1, 0, 0, 0], Red, sigmoid, sigmoid)`



Empezamos calculando el vector gradiente para un ejemplo del conjunto de entrenamiento. Supongamos que sus valores son $[1, 0, 0, 0]$.

Lo primero que hay que hacer es realizar la propagación hacia delante, para que los valores de los vectores al y zl de la red tengan los valores de la acción de predicción de la red

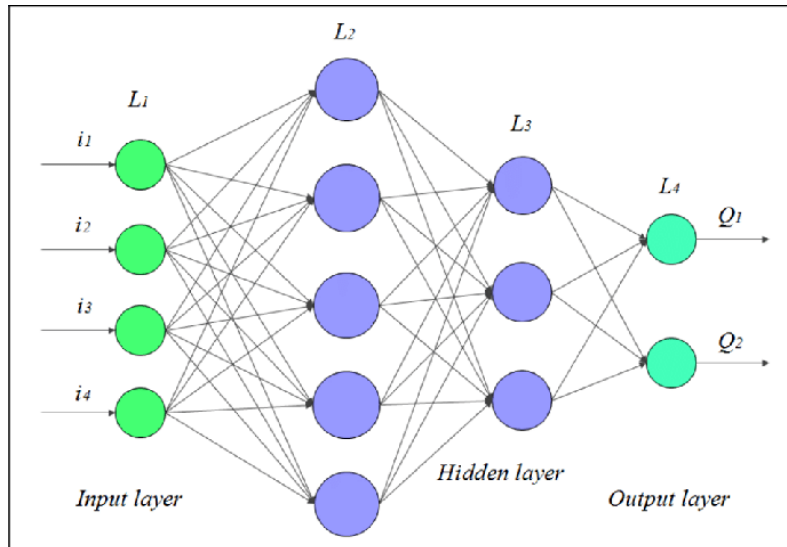
Ahora supongamos que en la capa de salida esperábamos obtener los valores $[1, 0]$, pero nuestra red, sin ajustar, predice los valores $[0.55, 0.45]$

Propagación hacia detrás

$$[(0.55 - 1), (0.45 - 0)] = [-0.45, 0.45]$$

$$[-0.45, 0.45] \xrightarrow{\text{EC}(L4) \ 2} \begin{bmatrix} Q & Q \\ 1 & 2 \end{bmatrix}$$

$$\text{FC} = 1 * -\log(0,55) + 0 * -\log(0,45) = 0,259637311$$



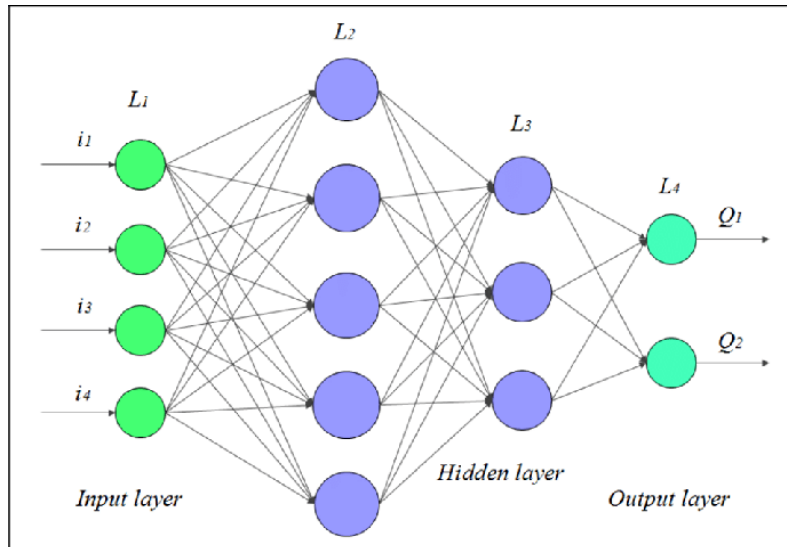
Primero, permaneciendo en la capa de salida, calcularemos el error de los biases de sus neuronas. EL primer paso es guardar en `VGrad.err_capa` en la capa de salida los valores de esta derivada de la función de pérdida

La fórmula de la función de pérdida, recordando que escogimos el error cuadrático, es (x es el valor predicho de una neurona, e y es el valor esperado para dicha neurona): $(x-y)^2$

La derivada de dicha fórmula con respecto a x queda: $2*(x-y) * 1 = 2*(x-y)$

La función de coste (FC) es sumatorio de la salida esperada por $-\log$ de la salida predicha 20

Propagación hacia atrás



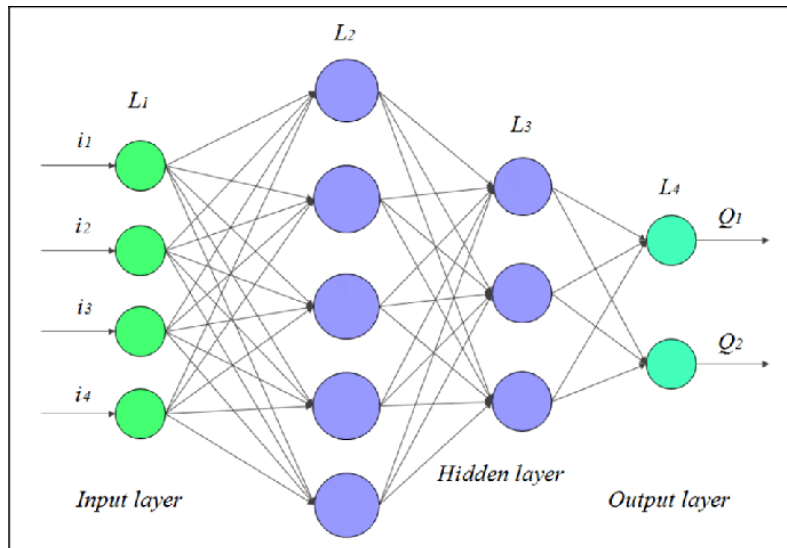
Seguimos en la capa de salida, ahora calcularemos por fin el valor del bias para las neuronas de la capa de salida. Solamente habrá que copiar los valores del error de la capa da salida al vector del bias de la capa de salida

Recordemos que en esta fase se multiplicaba cada valor del error por la derivada parcial de la función de activación, pero en el caso de Softmax esa operación no es necesario realizarlo.

Propagación hacia atrás

$$\text{transponer}(\text{vector_matriz1fila}(\text{AL(L3) } 3)) * \text{vector_matriz1fila}(\text{EB(L4) } 2) \longrightarrow \text{EW(L3,L4) } 3 \times 2$$

$$\begin{matrix} \begin{matrix} \square \\ \square \\ \square \end{matrix} & * & \begin{matrix} Q & Q \\ 1 & 2 \end{matrix} & = & \begin{matrix} \square & \square \\ \square & \square \\ \square & \square \end{matrix} \\ 3 \times 1 & & 1 \times 2 & & 3 \times 2 \end{matrix}$$



Ahora calcularemos el error de los pesos de las aristas de la capa actual. Para ello, simplemente, calcularemos este producto de matrices. EL resultado lo pasaremos a `VGrad.err_pesos` en la capa anterior a la capa de salida

La función `vector_matriz1fila` convierte un vector en una matriz de 1 fila igual al vector

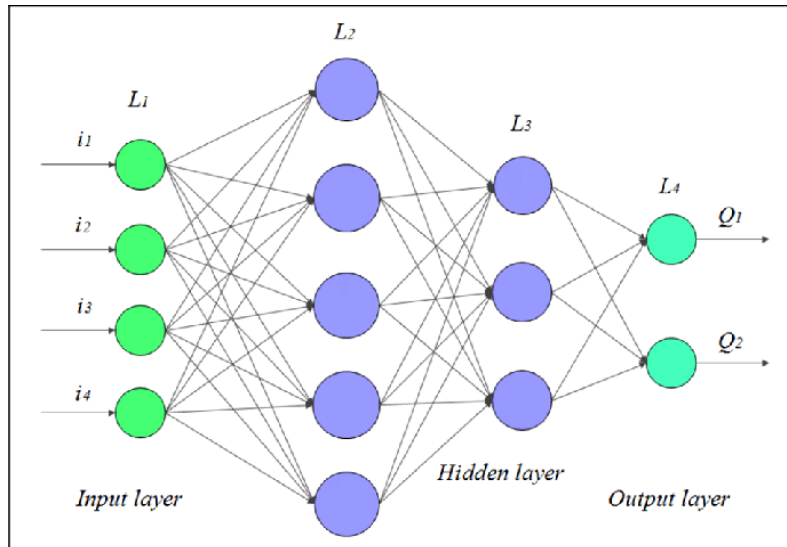
Aparte, la función `transponer`, transpone la matriz pasada como argumento

Propagación hacia atrás

$$\text{vector_matriz1fila} \left(\begin{matrix} Q & Q \\ 1 & 2 \end{matrix} \right) * \text{transponer} \left(\begin{matrix} & & \\ & & \\ & & \end{matrix} \right) \longrightarrow \text{EC}(L3) \ 3$$

$$\begin{matrix} Q & Q \\ 1 & 2 \end{matrix} \begin{matrix} * \\ \\ \end{matrix} \begin{matrix} & & \\ & & \\ & & \end{matrix} = \begin{matrix} & & \\ & & \\ & & \end{matrix}$$

1x2 2x3 1x3

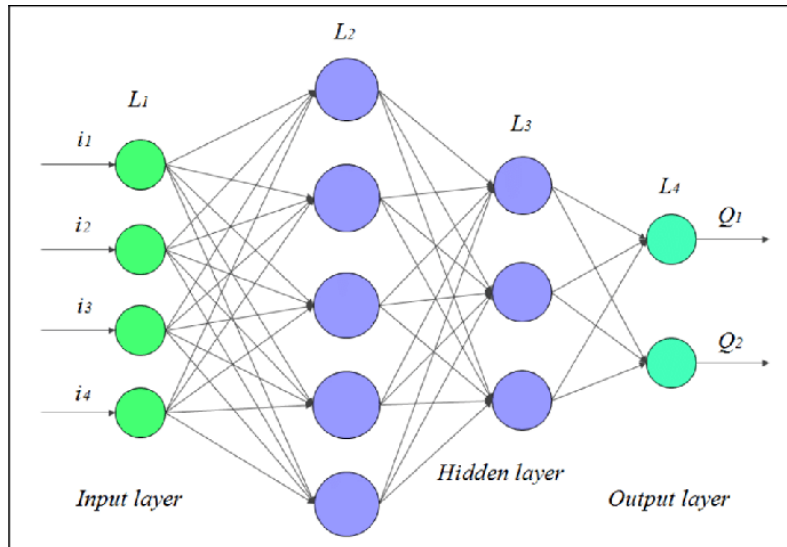
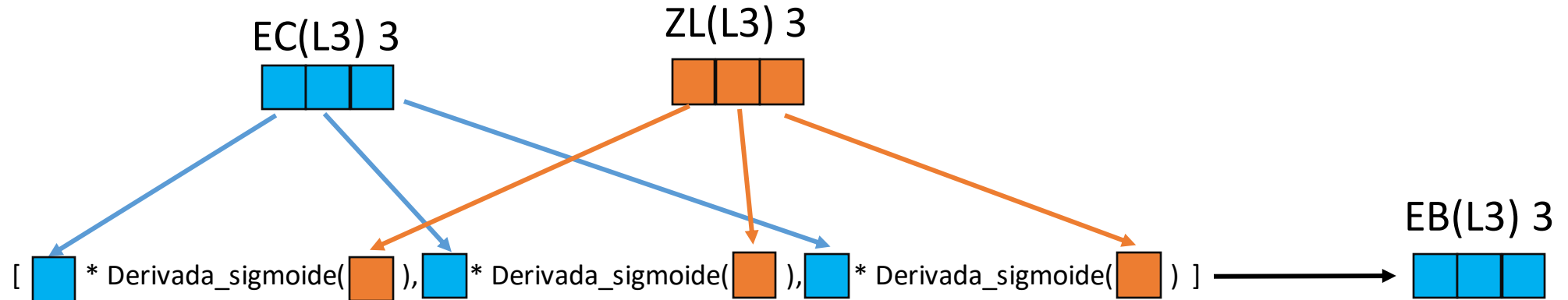


Para terminar en la capa de salida, calcularemos el vector de errores para la capa anterior. EL resultado lo pasaremos a `VGrad.err_capa` en capa anterior a la capa de salida, convirtiendo la matriz de una fila obtenida como un vector

La función `vector_matriz1fila` convierte un vector en una matriz de 1 fila igual al vector

Aparte, la función `transponer`, transpone la matriz pasada como argumento

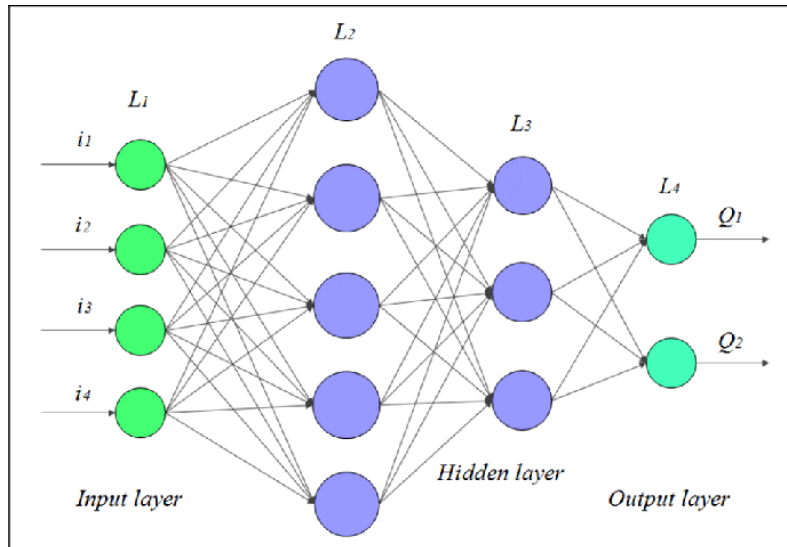
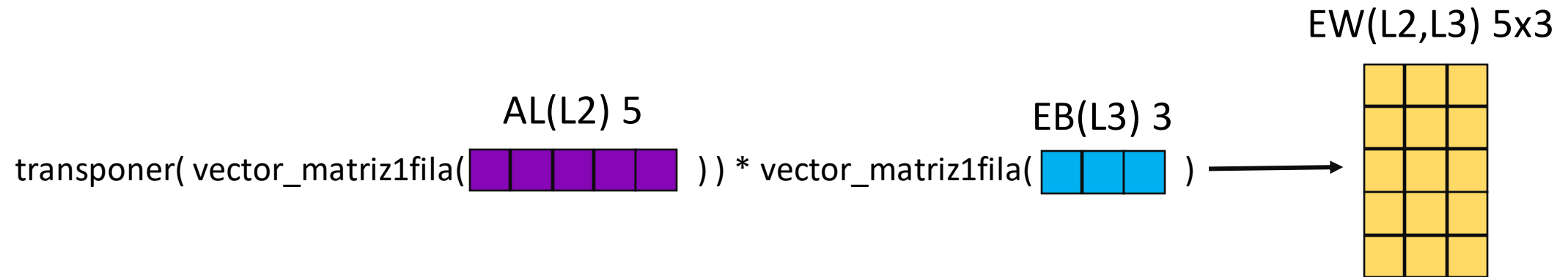
Propagación hacia atrás



Repetimos los 3 mismos pasos realizados anteriormente, pero ahora con la capa anterior, y ahora sí, multiplicando cada elemento del vector de error de la capa por la derivada parcial de cada valor de salida de la capa que ahora es la actual, sin aplicar la función de activación

Empezamos calculando los errores del bias para la que ahora es la capa actual

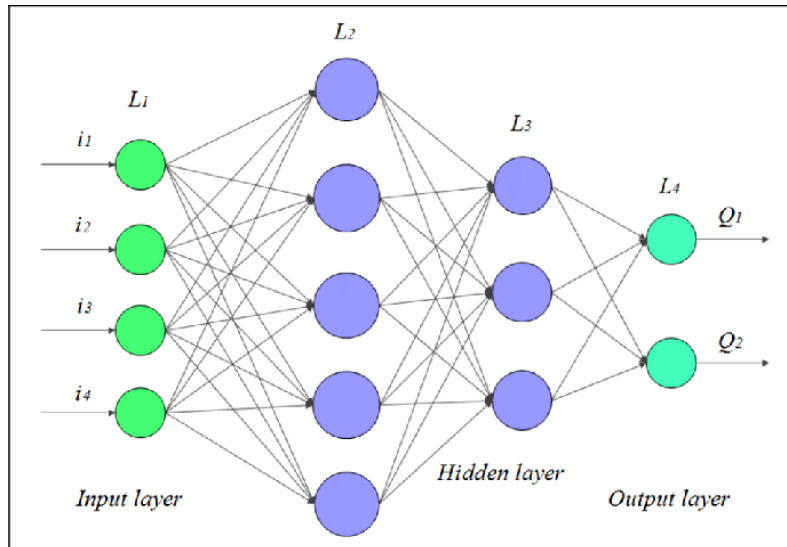
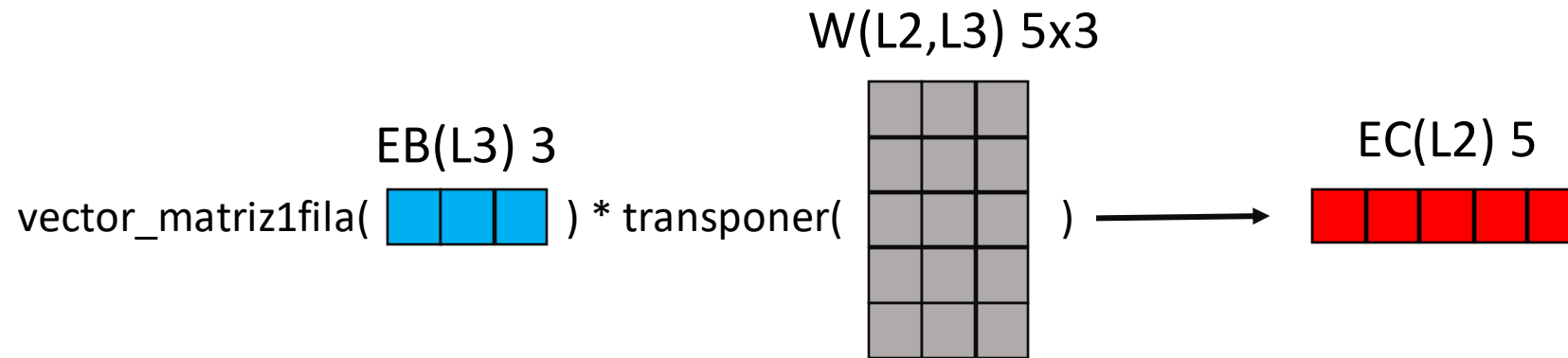
Propagación hacia atrás



Repetimos los 3 mismos pasos realizados anteriormente, pero ahora con la capa anterior

Después calculamos los errores de los pesos para la que ahora es la capa actual

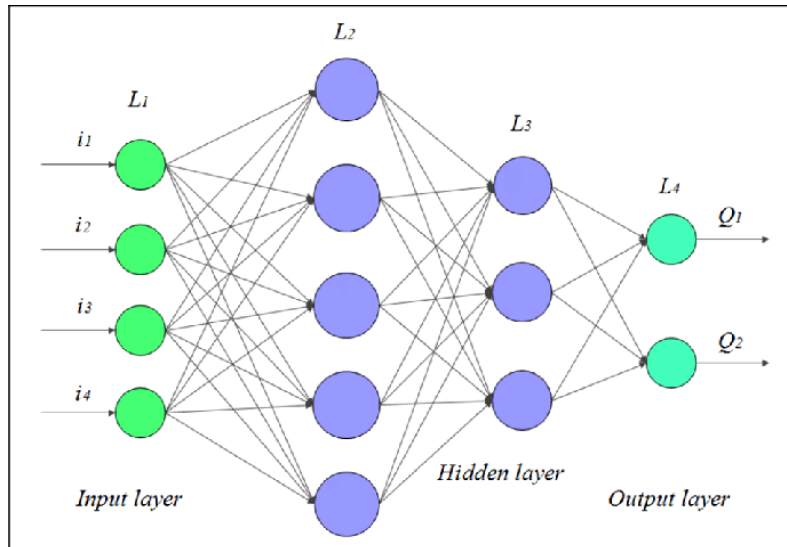
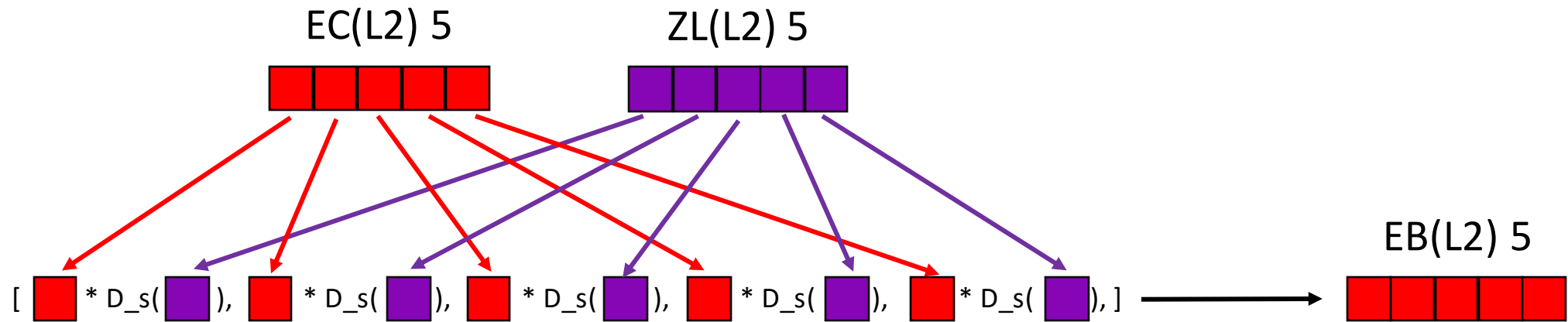
Propagación hacia atrás



Repetimos los 3 mismos pasos realizados anteriormente, pero ahora con la capa anterior

Para terminar, calculamos el error de la capa anterior con respecto a la que ahora es la capa actual

Propagación hacia detrás

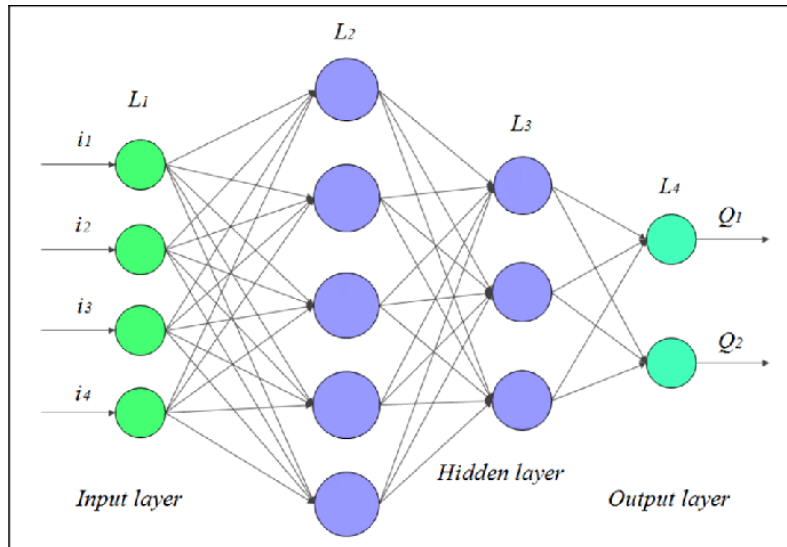
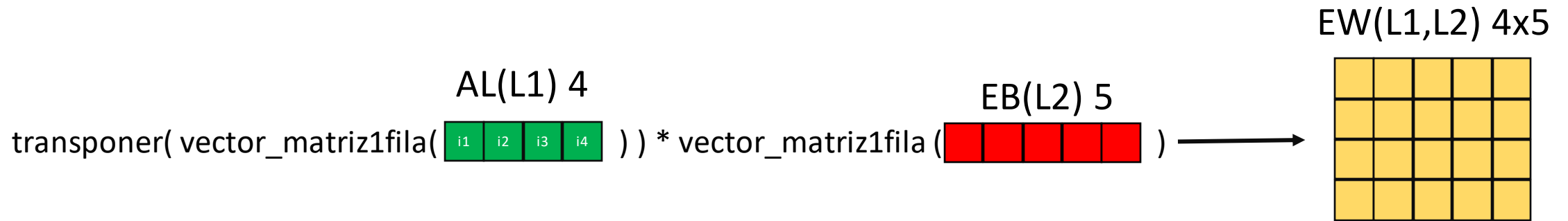


Al retroceder ahora a la capa anterior, en nuestro caso L2, podemos darnos cuenta de que la capa anterior es L1. Recordemos que la capa de entrada no se considera como una capa de neuronas, y es por ello que no hay que calcular ningún error para la capa anterior.

Por ello, solamente nos queda calcular los errores del bias y de los pesos para la capa L2

Aquí calculamos el error del bias. Acortamos Derivada_sigmoide por D_s

Propagación hacia atrás



Al retroceder ahora a la capa anterior, en nuestro caso L_2 , podemos darnos cuenta de que la capa anterior es L_1 . Recordemos que la capa de entrada no se considera como una capa de neuronas, y es por ello que no hay que calcular ningún error para la capa anterior.

Por ello, solamente nos queda calcular los errores del bias y de los pesos para la capa L_2

Aquí calculamos el error de los pesos

Funciones:

Constructor_vgrad: devuelve una lista con la estructura de la diapositiva 16, pero sin VGrad.error_capa. Recibe una red como parámetro.

Vector_mat1fila: recibe como parámetro 1 vector, y devuelve 1 matriz de una fila y tantas columnas como nº elementos del vector

Mat1fila_vector: recibe como parámetro 1 matriz de 1 fila, y devuelve un vector equivalente a la fila de dicha matriz

Producto_matrices: recibe como parámetros dos matrices, en orden, m1 y m2, y devuelve otra matriz como el producto $m1 * m2$

Transponer_matriz: toma una matriz m como argumento y devuelve su transpuesta, es decir, m^T

Reemplazar_valores_vectores: recibe como parámetros dos vectores, en orden, v1 y v2, copia los valores de v2 en v1

Reemplazar_valores_matrices: recibe como parámetros dos matrices, en orden, m1 y m2, copia los valores de m2 en m1

Se propone una implementación sin necesidad de guardar un vector de error por cada capa

Propagación hacia detrás

<code>sigmoid(x):</code> devolver $1 / (1 + e^{-x})$	<code>derivative_sigmoid(x):</code> devolver $\text{sigmoid}(x) * (1 - \text{sigmoid}(x))$	<code>softmax(x, all_output_values):</code> denominador = sumatorio(e^x para y cada valor de all_output_values) devolver $e^x / \text{sumatorio}$
---	---	---

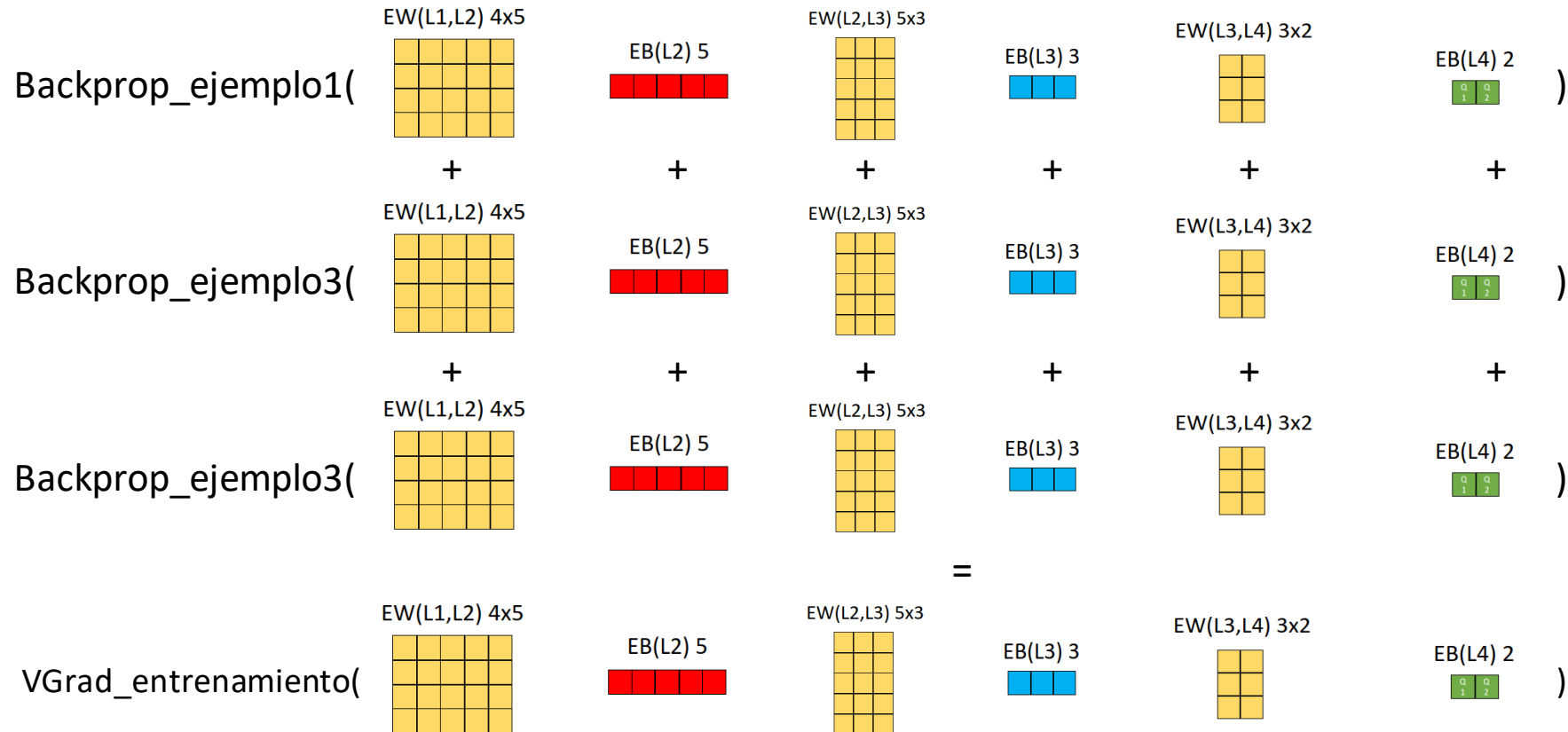
```
vector_gradiente_softmax( vector_vals_salida_esp, red, d_funcion_capas_ocultas ):
    vgrad = Constructor_vgrad(Red)
    #recordar, primero se hizo propagación hacia delante, ya se ha aplicado softmax a los vectores de la capa de salida
    bcs = [ red.al[red.numero_capas-1][i] - vector_vals_salida_esp[i] para Índice i desde 0 hasta tamaño( red.al[red.numero_capas-1] ) ]
    Índice i desde Red.numero_capas-2 hasta 0 (incluidos):
        Reemplazar_valores_vectores( vgrad.err_bias[i], bcs )
        err_pesos = Producto_matrices( Transponer_matriz( Vector_mat1fila( red.al[i] ) ), Vector_mat1fila( bcs ) )
        Reemplazar_valores_matrices( vgrad.err_pesos[i], err_pesos )
    Si i > 0:
        err_capa_ant = Mat1fila_vector( Producto_matrices( Vector_mat1fila( bcs ), Transponer_matriz( red.pesos[i] ) ) )
        bcs = [ d_funcion_capas_ocultas( red.zl[i-1][j] ) * err_capa_ant[j] para Índice j desde 0 hasta tamaño(err_capa_ant) ]
    Devolver vgrad
```

Ejemplo de ejecución (primero se debe hacer la propagación hacia delante):

Para cada ejemplo en conjunto_entrenamiento:

```
propagacion_hacia_delante( ejemplo.entrada, Red, sigmoid, softmax ) #el primero por ejemplo sería el [ 1, 0, 0, 0 ] visto antes
vgrad_calc = vector_gradiente( ejemplo.salida, Red, derivative_sigmoid )
```

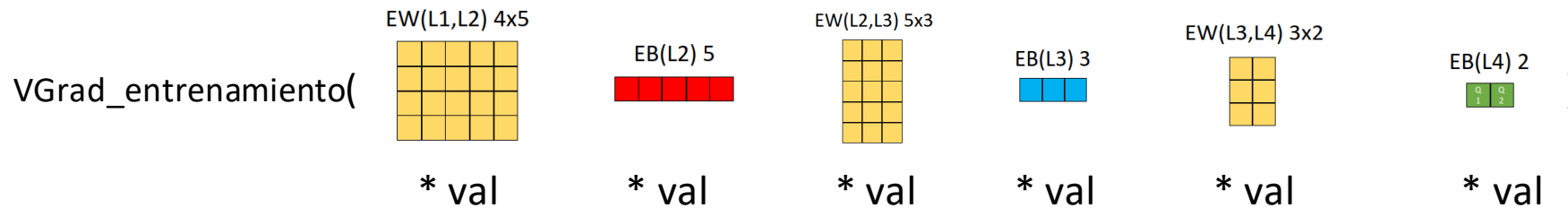
Propagación hacia atrás



Para pasar el vector gradiente a la red para que esta aprenda, teóricamente, primero se van calculando los vectores gradiente para todos los ejemplos del conjunto de entrenamiento con el que queremos entrenar nuestra red, y se suman sus valores. Supongamos que tenemos 3 ejemplos, para cada uno calculamos su vector gradiente, luego sumamos sus valores (para sus matrices de pesos y vectores de biases) y obtenemos un vector gradiente resultado

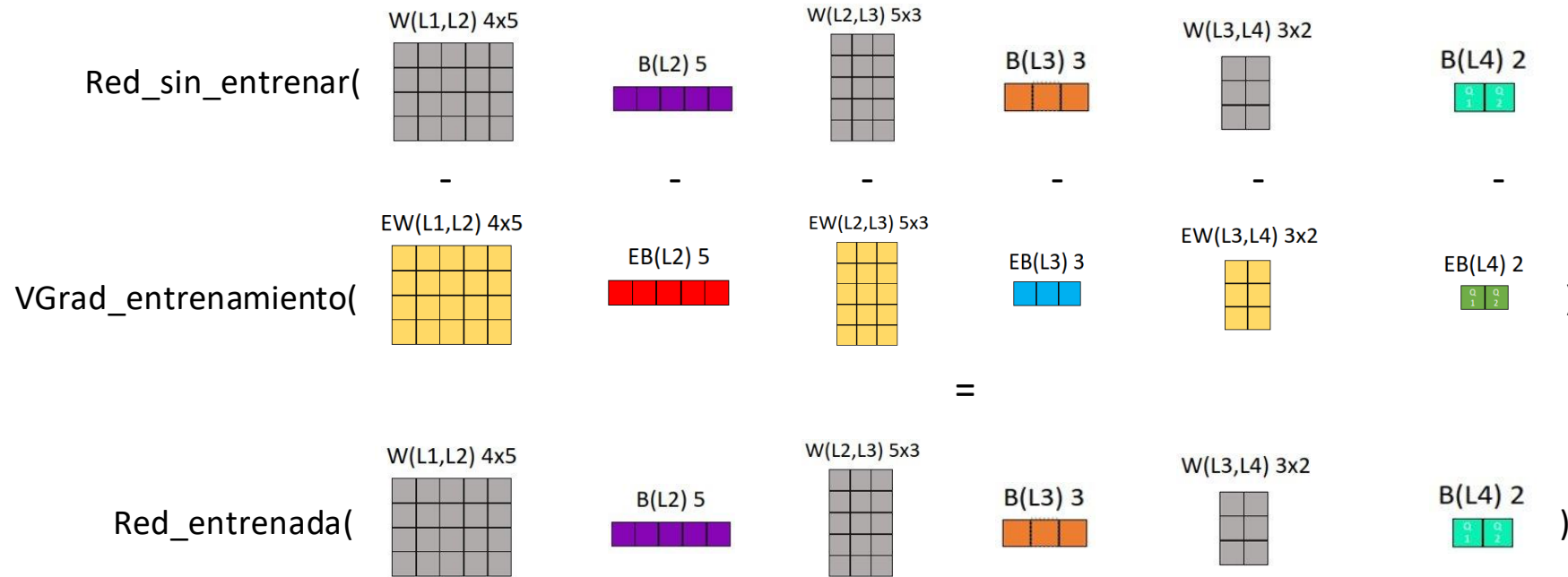
Propagación hacia atrás

$val = \text{tasa_aprendizaje} / \text{numero_ejemplos_entrenamiento} = \text{"número decimal entre 1 y cercano a 0"} / 3$



Después hay que multiplicar a cada elemento de cada matriz de pesos y vector de biases, un valor val , el cual es igual a dividir la tasa de aprendizaje escogida entre el número de ejemplos del conjunto de entrenamiento con los cuales se calculó un vector gradiente

Propagación hacia atrás



Finalmente, restamos a la red sin entrenar, para sus correspondientes matrices de pesos y vectores de biases, las matrices de error de pesos y vectores de error de biases del vector gradiente del entrenamiento ajustado del paso anterior. Con ello, ahora nuestra red se habrá ajustado para encajar en nuestro conjunto de datos.

Para que la red aprenda de manera efectiva, todo lo visto hasta ahora se repite cientos, miles o más veces.

Funciones:

Constructor_vgrad: devuelve una lista con la estructura de la diapositiva 16, pero sin VGrad.error_capa. Recibe una red como parámetro.

Suma_valores_vectores: recibe como parámetros dos vectores, en orden, v1 y v2, suma los valores de v2 a v1

Suma_valores_matrices: recibe como parámetros dos matrices, en orden, m1 y m2, suma los valores de m2 a m1

Resta_valores_vectores: recibe como parámetros dos vectores, en orden, v1 y v2, resta los valores de v2 a v1

Resta_valores_matrices: recibe como parámetros dos matrices, en orden, m1 y m2, resta los valores de m2 a m1

Multiplicar_a_vector: recibe como parámetros, en orden, un vector v y un número n, y pasa el valor de cada elemento x de v a $n \cdot x$

Multiplicar_a_matriz: recibe como parámetros, en orden, una matriz m y un número n, y pasa el valor de cada elemento x de m a $n \cdot x$

Propagación hacia atrás

```
propagacion_hacia_detrás_softmax( conjunto_entrenamiento, red, fun_act, d_fun_act ):
    vgrad = Constructor_vgrad( Red )
    para cada ejemplo en conjunto_entrenamiento:
        propagacion_hacia_delante( ejemplo.valores_entrada, Red, fun_act, softmax )
        vcalc = vector_gradiente_softmax( ejemplo.valores_salida, Red, d_fun_act )
        Para índice i desde 0 hasta tamaño( vgrad.err_bias ): Suma_valores_vectores( vgrad.err_bias[i], vcalc.err_bias[i] )
        Para índice i desde 0 hasta tamaño( vgrad.err_pesos ): Suma_valores_matrices( vgrad.err_pesos[i], vcalc.err_pesos[i] )
    Para índice i desde 0 hasta tamaño( vgrad.err_bias ):
        Multiplicar_a_vector( vgrad.err_bias[i], tasapren/tamaño(conjunto_entrenamiento) )
        Resta_valores_vectores( red.bias[i], vgrad.err_bias[i] )
    Para índice i desde 0 hasta tamaño( vgrad.err_pesos ):
        Multiplicar_a_matriz( vgrad.err_pesos[i], tasapren/tamaño(conjunto_entrenamiento) )
        Resta_valores_matrices( red.pesos[i], vgrad.err_pesos[i] )
```

Ejemplo de ejecución: `propagacion_hacia_detrás_softmax(conjunto_entrenamiento, Red, sigmoid, derivative_sigmoid)`

En teoría, como se puede ver, se calcula `vgrad` en función de todos los ejemplos del conjunto de entrenamiento, pero en la práctica, este conjunto de entrenamiento se divide en secciones o baterías (batches en inglés), y se llama a esta función tantas veces como el número de baterías en las que se haya dividido dicho conjunto

Ojo también, aquí hemos considerado utilizar todas las funciones sigmoide, no tiene por qué, por ejemplo, en la capa de salida, se podría haber escogido utilizar Softmax.

Propagación hacia atrás

Cada vez que se hacen todas las operaciones de propagación hacia atrás, hasta ajustar una vez la red al restar el vector gradiente calculado, se está realizando una época de entrenamiento. Se suelen hacer cientos, miles o más épocas para un mismo conjunto de datos, para que la red aprenda mejor, como se mencionó anteriormente.

Como mencionamos, en la práctica no se calcula un vector gradiente de todo el conjunto de entrenamiento, sino que este conjunto se separa en otros subconjuntos llamados baterías (batches en inglés), en cada uno de ellos se ordenan sus elementos de manera aleatoria, y para cada uno de ellos se ejecuta la propagación hacia atrás. Este proceso además se suele repetir varias épocas, hasta conseguir entrenar el modelo para que tenga un mejor rendimiento

Para hacer una aproximación a cómo implementar una red neuronal artificial, se ha mostrado que la tasa de aprendizaje es un hiperparámetro de la red fijo que se elige para todo el entrenamiento. Sin embargo, esto no es práctico, y se suelen utilizar algunos algoritmos, como ADAM, para ir variando esta tasa de aprendizaje en función de la función de coste y otros criterios, el cual permite llegar a un mínimo local de la función de coste antes (es decir, que permite aprender bien hasta un cierto punto sin desaprender) y además acelerar el proceso de llegada a dicho punto

Aparte, se pueden utilizar diferentes funciones de activación, y sobre todo, para la capa de salida con respecto al resto de capas ocultas, como por ejemplo Softmax, la cual utilizando otra función de coste y de pérdida llamada entropía cruzada, es posible entrenar a las redes neuronales artificiales para resolver mejor problemas de clasificación que una red neuronal como la vista en estas diapositivas (función de pérdida del error cuadrático y todas las funciones de activación sigmoide)