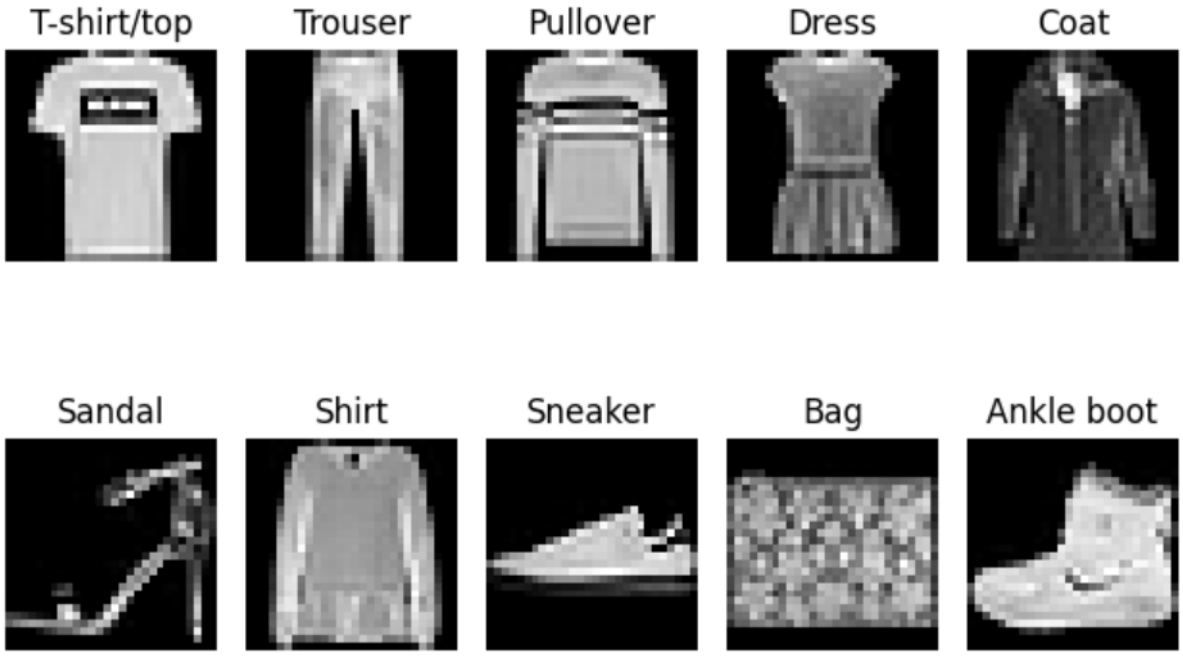
Classificazione di immagini di articoli di moda utilizzando reti neurali convolutive

La classificazione delle immagini è uno dei problemi fondamentali nella computer vision. Sebbene il problema di identificare un'entità visiva da un'immagine sia molto banale per un essere umano, è molto impegnativo per un algoritmo informatico fare lo stesso con una precisione uguale al livello umano. Le reti neurali convolutive hanno dimostrato ottimi risultati nella classificazione delle immagini.

I dati utilizzati per la creazione delle reti neurali, esposte successivamente, sono immagini di articoli di moda di Zalando contenuti nel dataset Fashion-MNIST. Il dataset è già suddiviso in training set e test set contenenti 60.000 e 10.000 osservazioni, rispettivamente. Ogni esempio è un'immagine in scala di grigi 28x28 e a ogni immagine è associata un'etichetta. Le classi sono 10 e le rispettive descrizioni sono le seguenti:

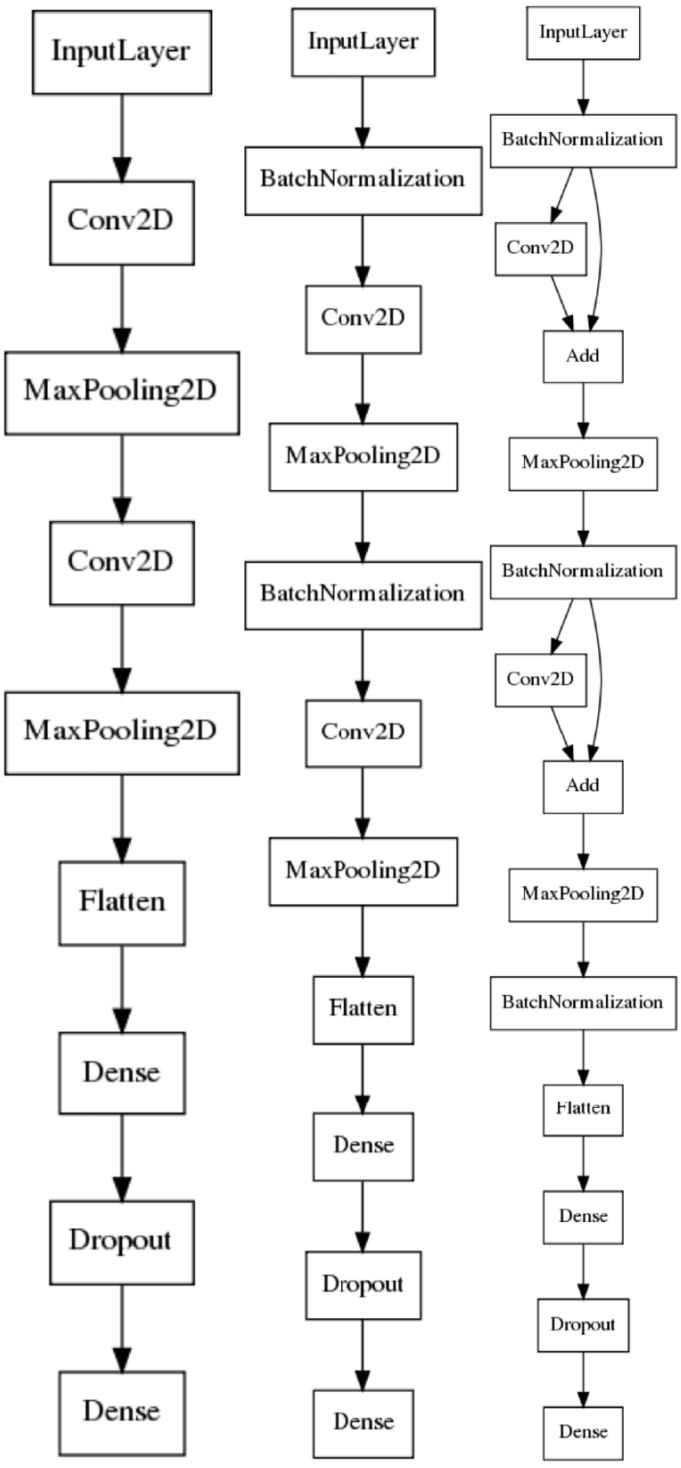
Etichette	Descrizioni
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot

Nella figura è mostrato un esempio di immagine per ogni classe:



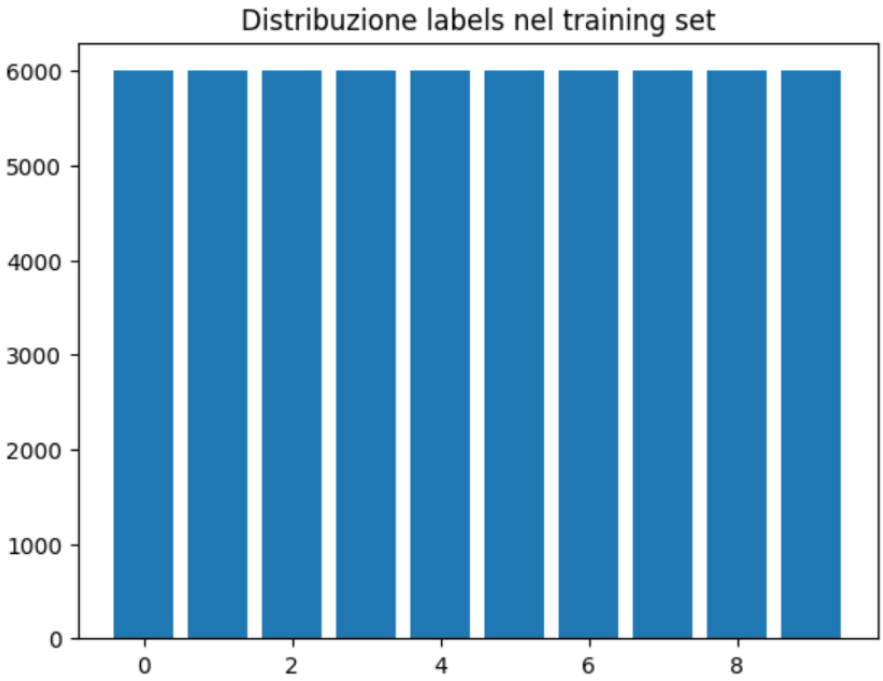
L’obiettivo è proporre tre diverse architetture di rete neurale convolutiva per la classificazione delle immagini di articoli di moda del dataset Fashion-MNIST, usando batch normalization e residual skip connections per facilitare e accelerare il processo di apprendimento. Le tre architetture sono state apprese tramite un set di validazione pari al 20%, ampiezza del minibatch pari a 64 e un numero di epoche pari a 50. Dopo aver scelto la migliore rete fra le tre tramite misure di validazione come accuratezza ed F1 score, si procede cambiando i parametri definiti precedentemente per comprendere come variano le misure di validazione al variare di tali parametri.

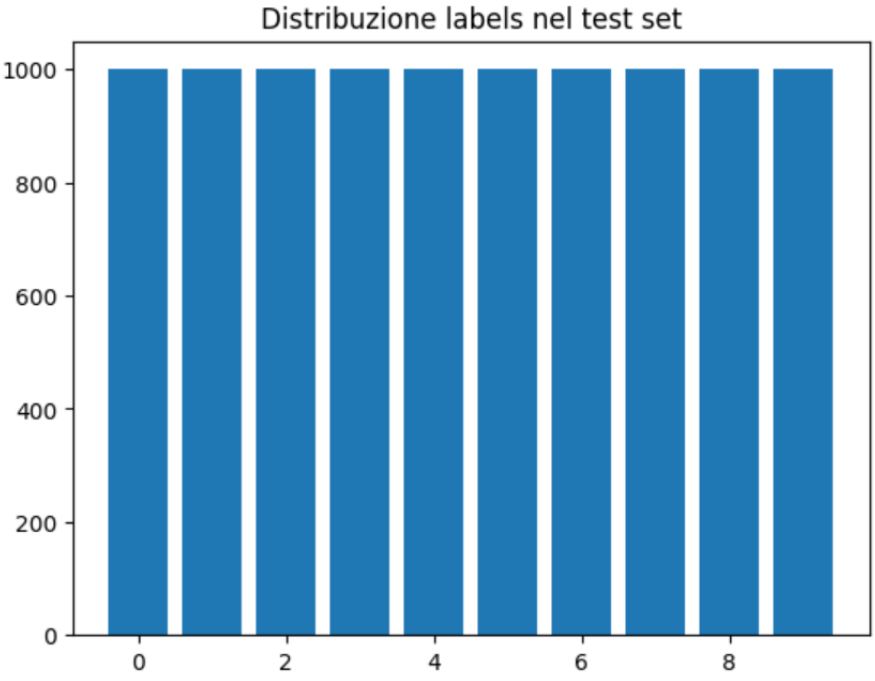
Uno schema delle tre architetture proposte è mostrato in figura:



Analisi esplorativa

Prima di procedere con la costruzione delle reti neurali, è importante osservare la distribuzione delle classi sia nel training set che nel test set. Dai grafici a barre si nota che i due set sono perfettamente bilanciati, infatti, ogni classe contiene esattamente un decimo delle osservazioni; quindi, per quanto riguarda il training set ogni classe contiene 6.000 esempi e per quanto riguarda il test set ogni classe contiene 1.000 esempi. Il bilanciamento delle classi permette di addestrare il modello in modo equilibrato su tutte le classi, di conseguenza, consente una migliore generalizzazione e di ottenere una valutazione più accurata delle prestazioni del modello.





Convolutional neural network

Un tipico strato di una rete convolutiva è costituito da tre stadi:

1. Nel primo stadio si utilizzano un certo numero kernel di dimensioni normalmente molto piccole che si fanno scorrere sull'immagine di ingresso per creare una feature map. Nelle reti proposte, ogni layer convolutivo ha 32 kernel di dimensione 3x3;
2. Nella fase successiva, si applica una funzione di attivazione non lineare sul risultato della convoluzione al passo precedente. Nelle reti proposte la funzione di attivazione è la ReLU;
3. Infine, viene applicata una statistica riassuntiva agli output vicini. Questo stadio è chiamato pooling e nelle reti proposte è stato eseguito il max pooling su ogni porzione di dimensione 2x2 pixel.

Prima architettura

La prima architettura considerata è composta da due layers convolutivi e due layers max pooling uno dopo l'altro. Dopo aver appiattito l'uscita dell'ultimo layer di max-pooling, è stata utilizzata una rete MLP (Multilayer Perceptron) per classificare le immagini attraverso un hidden layer di 128 neuroni e un output layer di 10 neuroni per le 10 diverse classi. La funzione di attivazione utilizzata nell'hidden layer è la "ReLU", mentre la funzione di attivazione utilizzata nell'output layer è la softmax. Inoltre, è stato utilizzato anche il 50% di dropout come misura di regolarizzazione. Infine, è necessario specificare la funzione di costo e l'algoritmo di ottimizzazione. In questo caso la funzione di costo è la categorical cross entropy e per l'ottimizzazione dei parametri è stato utilizzato l'algoritmo Adam.

Seconda architettura

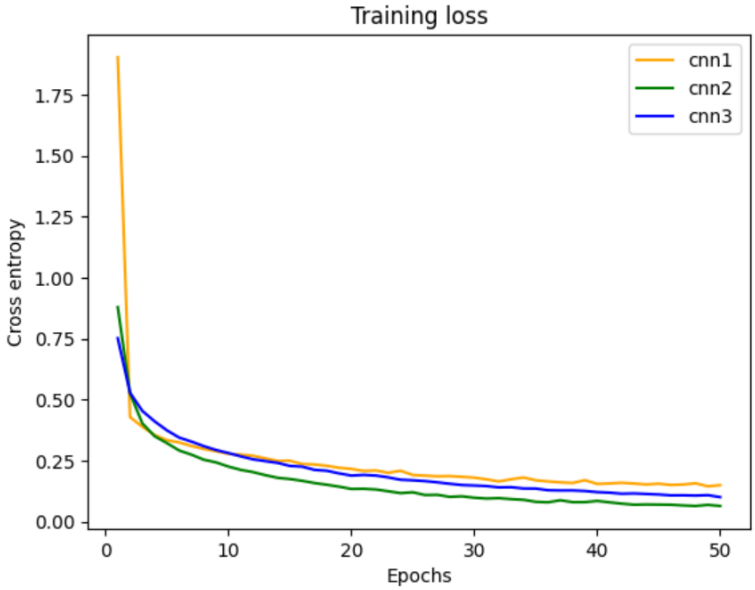
Questo modello aggiunge la batch normalization all'architettura precedente. La batch normalization è stata effettuata prima di ogni strato convolutivo per migliorare la velocità di addestramento del modello. Durante il processo di addestramento, gli ingressi di ogni layer convolutivo vengono normalizzati utilizzando la media e la varianza dei valori del minibatch attualmente in elaborazione. L'ingresso del minibatch è generalmente modificato in modo da avere media zero e varianza unitaria. Questa tecnica ci permette di addestrare la rete più velocemente con tassi di apprendimento più elevati.

Terza architettura

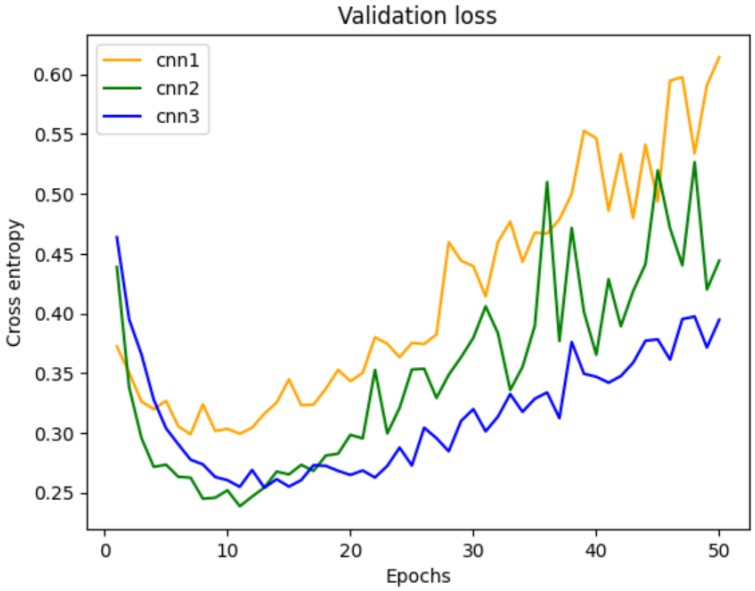
Quest'ultima architettura introduce le residual skip connections all'architettura precedente facilitando il processo di apprendimento. Quando una rete neurale profonda viene addestrata tramite backpropagation, il segnale del gradiente deve essere propagato all'indietro dallo strato di uscita più alto fino al livello di ingresso più basso per garantire che i parametri vengano aggiornati in modo appropriato. Le residual skip connections rappresentano un modo intuitivo per raggiungere questo obiettivo. In questo modello si sommano l'ingresso precedente e il valore corrente dell'output del layer convolutivo per ottenere l'output finale: $y = convolution(x) + x$

Confronto

Per quanto riguarda l'errore di addestramento calcolato sul training set, questo diminuisce al crescere del numero di epoche. Una riduzione maggiore si registra per la seconda architettura.



Tuttavia, è più importante concentrarsi sui valori dell'errore di validazione calcolati tramite il validation set. In questo caso si nota come le tre curve dalla decima epoca in poi tendono a crescere nuovamente; quindi, si riscontra un problema di overfitting con bassa capacità e bassa generalizzazione del modello.



Confrontando le reti in termini di accuratezza ed F1 score, l'architettura con batch normalization è senza residual skip connections risulta la migliore poche presenta i valori più alti. Considerando, quindi, la seconda architettura verranno modificati il numero di epoche ponendo tale valore uguale a 10, poiché è l'epoca per la quale si è registrato il minore errore di validazione, verranno confrontati

due differenti dimensioni del minibatch pari a 64 e 128 e verranno confrontati due differenti valori di validation split pari a 0.2 e 0.1.

Architettura	Accuratezza	F1 score
CNN	0.8928	0.8924
CNN + BatchNorm	0.9160	0.9161
CNN + BatchNorm + Skip	0.9105	0.9106

Scelta parametri

Considerando i diversi valori di accuratezza ed F1 score, ottenuti tramite diverse combinazioni di ampiezza del minibatch e percentuale di suddivisione per la creazione del validation set, si nota che per qualsiasi combinazione di parametri i valori delle misure di validazione risultano incrementati rispetto ai precedenti valori della seconda architettura. Inoltre, per il minibatch più ampio il modello risulta migliore probabilmente perché la stima del gradiente risulta più accurata. Ciò si nota soprattutto considerando il validation split pari a 0.1.

Parametri	Accuratezza	F1 score
batch_size=64, validation_split=0.2	0.9185	0.9185
batch_size=64, validation_split=0.1	0.9175	0.9170
batch_size=128, validation_split=0.2	0.9186	0.9183
batch_size=128, validation_split=0.1	0.9224	0.9221

In definitiva, potremmo considerare la rete con l'aggiunta della batch normalization e parametri posti uguali a 10 epoche, ampiezza del minibatch di 128 e validation split di 0.1 per prevedere la categoria appartengono le immagini di articoli di moda. La matrice di confusione di tale rete mostra che le immagini erroneamente classificate della classe 6 (Camicia) e della classe 0 (Maglietta/top) sono la maggior parte delle fonti di errore. Infatti, 86 camicie vengono predette come magliette e 70 come giubbotti. Ma anche il contrario è molto frequente poiché 90 magliette vengono classificate erroneamente come camicie. Tuttavia, l'accuratezza raggiunta del 92,24% è molto buona.

Etichette	Descrizioni
0	T-shirt/top
1	Trouser
2	Pullover
3	Dress
4	Coat
5	Sandal
6	Shirt
7	Sneaker
8	Bag
9	Ankle boot