Università degli studi di
Milano-Bicocca

Advanced Machine Learning

Progetto Finale

Overfitting

Autori:
Lorenzo Mammana - 807391- l.mammana@campus.unimib.it
Eric Nisoli - 807147- e.nisoli1@campus.unimib.it

2 gennaio 2020

Sommario

The ABSTRACT is not a part of the body of the report itself. Rather, the abstract is a brief summary of the report contents that is often separately circulated so potential readers can decide whether to read the report. The abstract should very concisely summarize the whole report: why it was written, what was discovered or developed, and what is claimed to be the significance of the effort. The abstract does not include figures or tables, and only the most significant numerical values or results should be given.

1 Introduzione

L'obiettivo del progetto è quello di costruire un modello di machine learning per la classificazione di immagini in grado di operare bene su un dataset di piccole dimensioni comprendente un alto numero di classi simili tra di loro. Come è facilmente intuibile, il problema principale di operare su un dataset di questo tipo è l'overfitting del modello sui dati di training con conseguente fallimento del modello su dati nuovi. Per risolvere questo problema ci siamo concentrati sulla ricerca di modelli, iperparametri e tecniche di generalizzazione che permettessero al modello di essere in grado di operare su un dataset di test molto più ampio di quello utilizzato per l'addestramento. Il modello utilizzato per questo tipo di analisi è una rete neurale convoluzionale; questo tipo di modello si è dimostrato negli ultimi anni il migliore nel classificare immagini in moltissimi campi di applicazione.

The introduction should provide a clear statement of the problem posed by the project, and why the problem is of interest. It should reflect the scenario, if available. If needed, the introduction also needs to present background information so that the reader can understand the significance of the problem. A brief summary of the hypotheses and the approach your group used to solve the problem should be given, possibly also including a concise introduction to theory or concepts used later to analyze and to discuss the results.

2 Dataset

Il dataset utilizzato è il *102 Category Flower Dataset* creato dai ricercatori del *Visual Geometry Group* di *Oxford*. Il dataset è composto da 8189 immagini

RGB di dimensione variabile, ogni immagine contiene uno o più fiori su sfondo neutro ed è etichettata con una singola categoria estratta da un insieme di centodue possibili. E' stata mantenuta la suddivisione del dataset definita nella pubblicazione originale, in particolare si hanno:

- 1020 immagini nel training set
- 1020 immagini nel validation set
- 6149 immagini nel test set

Ogni categoria è rappresentata da 10 immagini nel training set e nel validation set, mentre la proporzione di immagini per ogni categoria varia nel test set. La difficoltà di operare su un dataset di questo tipo è evidente, il numero di immagini è limitato mentre il test set è ampio. Mettere immagini dataset

In this section the available data sets must be presented. The term dataset refers to any type of information source, for example web services for geolocation fall into this category. In addition, all necessary data manipulation processes, such as cleaning and enrichment with external sources, must be presented and discussed.

3 The Methodological Approach

Per classificare le immagini nelle classi corrette si è deciso di utilizzare una rete neurale convoluzionale (CNN). Come è noto addestrare questo tipo di modello da zero richiede un numero di immagini dell'ordine di milioni di elementi, non avendo a disposizione così tante immagini è necessario eseguire la tecnica nota come *finetuning* in cui la CNN viene inizializzata con pesi calcolati su un dataset di milioni di immagini e modificati in accordo con i propri dati disponibili. In particolare si è deciso di utilizzare reti preaddestrate sul dataset Imagenet contenente più di 10 milioni di immagini di cui più di trecentomila rappresentanti fiori.

3.1 Esperimento baseline

Considerando il fatto che il numero di immagini di training è limitato è stato inizialmente deciso di definire un modello baseline non particolarmente complesso. Il modello scelto è una *Resnet-18* preaddestrata su *Imagenet*. Inizialmente si è provato ad utilizzare la rete come estrattore di features

per valutare quanto il preaddestramento influisca sulla classificazione finale. Le features estratte vengono classificate utilizzando un semplice percettrone multistrato:

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 256)	6422784
dense_2 (Dense)	(None, 128)	32896
dense_3 (Dense)	(None, 102)	13158

Tabella 1: Architettura del percettrone

3.1.1 Preprocessing e Data augmentation

Prima di procedere con l'estrazione delle features le immagini vengono normalizzate per avere media nulla e varianza unitaria utilizzando la formula:

$$x' = \frac{x - \overline{x}}{\sigma} \tag{1}$$

Dove $\overline{x}$ e σ rappresentano rispettivamente la media e la varianza calcolate su *Imagenet*. Questo permette di passare alla rete delle immagini più simili a quelle su cui è stata addestrata ottenendo potenzialmente delle features più discriminative.

3.1.2 Scelta degli iperparametri

Per addestrare il percettrone sulle features estratte si è deciso di utilizzare il semplice mini-batch stochastic gradient descent (SGD), con dimensione del batch pari a 64 e momentum pari a 0.9. Il learning rate viene calcolato in accordo con l'algoritmo descritto in *Cyclical Learning Rates for Training Neural Networks*. Questo algoritmo prevede che il learning rate oscilli avanti ed indietro in un range prefissato di valori, questo ha un duplice scopo:

- Uscire da punti di sella o da minimi locali che potrebbero bloccare la corretta convergenza dell'ottimizzatore.
- Ridurre il bias introdotto dalla scelta iniziale di un learning rate scorretto.

Per calcolare il range su cui far variare il learning rate viene utilizzata una tecnica automatizzata così funzionante:

1. Si definiscono un estremo inferiore piccolo ed un estremo superiore grande su cui far variare il learning rate, ad esempio [1e-10, 1e-1].
2. Si addestra la rete per un numero ridotto di epoche partendo dall'estremo inferiore ed incrementando esponenzialmente il learning rate ad ogni batch.
3. Il training continua fino a quando il learning rate non raggiunge l'estremo superiore
4. Viene plottato un grafico che mostra quando il learning rate è troppo basso o troppo alto.

L'algoritmo sul percettrone multistrato produce il seguente grafico:

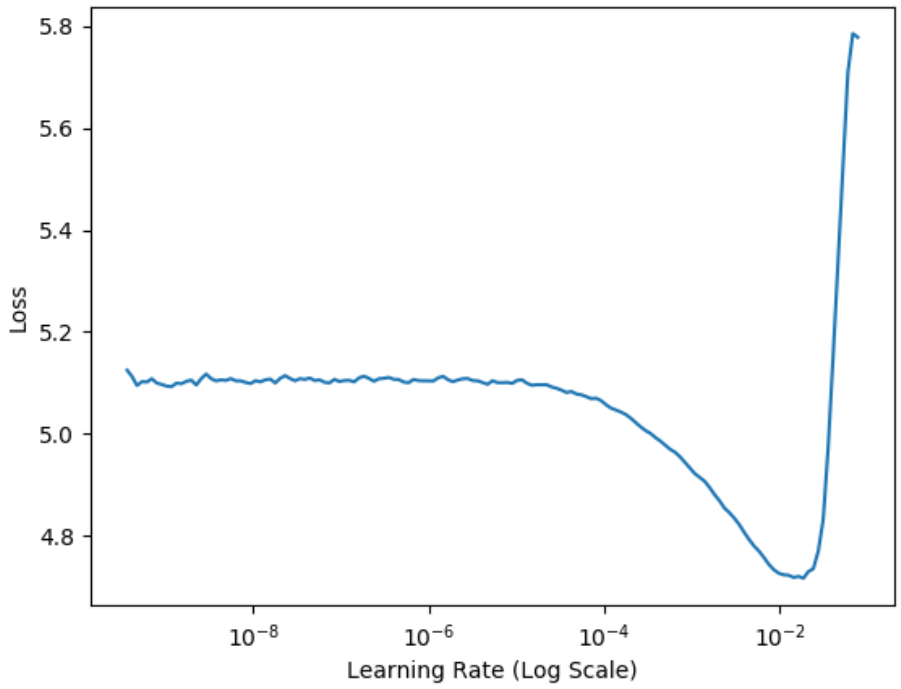

Figura 1: Output dell'algoritmo di ricerca del learning rate

Dal grafico si vede come la rete cominci ad apprendere circa a partire da 1e-5 e diverga una volta superato 1e-2, questi due valori verranno quindi utilizzati come estremi per l'algoritmo di cyclical learning rate. Questo algoritmo ha due ulteriori iperparametri, il primo è definito *Step size* ed indica il numero di iterazioni richiesto per passare dal minimo learning rate al massimo, il secondo è il metodo con cui viene modificato il learning rate. La rete viene addestrata per 50 epoche con diverse coppie di iperparametri; i risultati sono riportati in tabella.

Metodo	Step size	Test accuracy
Triangular	2	0.550
Triangular	4	0.550
Triangular	6	0.558
Triangular	8	0.547
Triangular2	2	0.510
Triangular2	4	0.527
Triangular2	6	0.550
Triangular2	8	0.533
Non cyclical	-	0.478

Tabella 2: Risultati dei diversi iperparametri sul test set

L'algoritmo non cyclical utilizza SGD con learning rate iniziale pari a 1e-2 e decay del learning rate di un fattore 0.1 in caso di non decrescita della loss sul set di validazione. E' evidente come l'algoritmo ciclico converga ad una soluzione decisamente migliore e soprattuto lo riesca a fare riducendo al massimo il tempo necessario per cercare il miglior learning rate per la rete. In tabella vengono mostrati due metodi uno definito *Triangular* e l'altro definito *Triangular2*, la differenza è che il secondo metodo dimezza l'ampiezza dell'intervallo di ricerca del learning rate ad ogni completamento del ciclo, questo è ben visibile plottando la variazione del learning rate nel tempo.

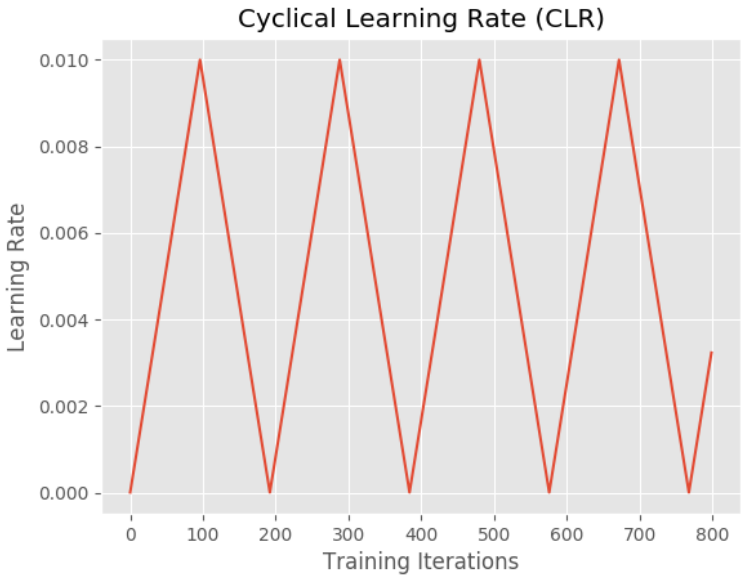

Figura 2: Andamento del metodo Triangular con step size 6

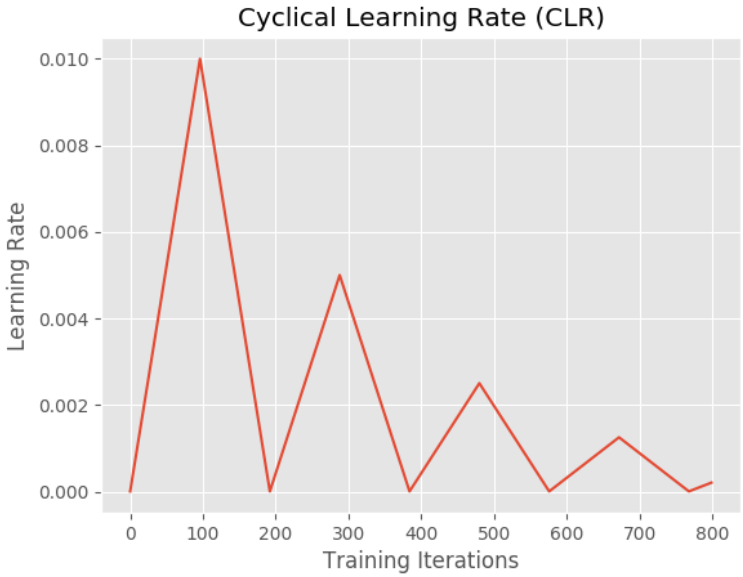

Figura 3: Andamento del metodo Triangular2 con step size 6

Da questa prima valutazione si è in grado di notare come il preaddestramento sia già in grado di discriminare con una buona accuratezza le diverse classi da cui è composto il dataset.

3.2 Esperimento 1 - Finetuning Resnet-18

La sola estrazione delle feature non è ovviamente abbastanza per ottenere delle performance soddisfacenti sul test set, si è proceduto quindi ad eseguire il finetuning dell'intera rete in modo tale da aggiornare i pesi per renderli più compatibili con il nuovo dataset. Adesso invece di accodare un percettrone multistrato all'ultimo layer della rete viene invece applicato un average pooling seguito da un layer di dropout ed un layer completamente connesso contenente 102 neuroni.

3.2.1 Preprocessing e Data augmentation

Per massimizzare la generalizzazione del modello si è deciso di "aumentare" le immagini di training applicando i seguenti effetti durante la fase di training:

- Flip orizzontale
- Aumento/riduzione della luminosità
- Rotazione in un angolo di $\pm$ 10°

This is the central and most important section of the report. Its objective must be to show, with linearity and clarity, the steps that have led to the definition of a decision model. The description of the working hypotheses, confirmed or denied, can be found in this section together with the description of the subsequent refining processes of the models. Comparisons between different models (e.g. heuristics vs. optimal models) in terms of quality of solutions, their explainability and execution times are welcome.

Do not attempt to describe all the code in the system, and do not include large pieces of code in this section, use pseudo-code where necessary. Complete source code should be provided separately (in Appendixes, as separated material or as a link to an on-line repo). Instead pick out and describe just the pieces of code which, for example:

- are especially critical to the operation of the system;

- you feel might be of particular interest to the reader for some reason;
- illustrate a non-standard or innovative way of implementing an algorithm, data structure, etc..

You should also mention any unforeseen problems you encountered when implementing the system and how and to what extent you overcame them. Common problems are: difficulties involving existing software.

4 Esperimento 0

4.0.1 Preprocessing e Data augmentation

Per massimizzare la generalizzazione del modello si è deciso di "aumentare" le immagini di training applicando i seguenti effetti durante la fase di training:

- Flip orizzontale
- Aumento/riduzione della luminosità
- Rotazione in un angolo di $\pm$ 10°

Le immagini vengono poi normalizzate per avere media nulla e varianza unitaria utilizzando la formula:

$$x' = \frac{x - \overline{x}}{\sigma} \tag{2}$$

Dove $\overline{x}$ e σ rappresentano rispettivamente la media e la varianza calcolate su *Imagenet*

5 Results and Evaluation

The Results section is dedicated to presenting the actual results (i.e. measured and calculated quantities), not to discussing their meaning or interpretation. The results should be summarized using appropriate Tables and Figures (graphs or schematics). Every Figure and Table should have a legend that describes concisely what is contained or shown. Figure legends go below the figure, table legends above the table. Throughout the report, but especially in this section, pay attention to reporting numbers with an appropriate number of significant figures.

6 Discussion

The discussion section aims at interpreting the results in light of the project's objectives. The most important goal of this section is to interpret the results so that the reader is informed of the insight or answers that the results provide. This section should also present an evaluation of the particular approach taken by the group. For example: Based on the results, how could the experimental procedure be improved? What additional, future work may be warranted? What recommendations can be drawn?

7 Conclusions

Conclusions should summarize the central points made in the Discussion section, reinforcing for the reader the value and implications of the work. If the results were not definitive, specific future work that may be needed can be (briefly) described. The conclusions should never contain "surprises". Therefore, any conclusions should be based on observations and data already discussed. It is considered extremely bad form to introduce new data in the conclusions.

References

The references section should contain complete citations following standard form. The references should be numbered and listed in the order they were cited in the body of the report. In the text of the report, a particular reference can be cited by using a numerical number in brackets as [**?**] that corresponds to its number in the reference list. LaTeXprovides several styles to format the references

Riferimenti bibliografici