



Università degli Studi di Verona

DIPARTIMENTO DI INFORMATICA
Corso di Laurea in Ingegneria e scienze informatiche

PROGETTO DI TEORIE E TECNICHE DEL RICONOSCIMENTO

Food Recognition

Candidato:

Alessia Bodini

Matricola VR451051

Anno Accademico 2019–2020

Indice

1	Motivazioni	2
2	Stato dell'arte	2
3	Obbiettivi	2
4	Metodologia	3
4.1	Ricerca del dataset	3
4.2	Estrazione delle features	3
4.3	Metodi di riconoscimento usati	3
5	Esperimenti e risultati raggiunti	6
5.1	KNN	6
5.2	SVM	6
5.3	DNN	7
6	Conclusioni	7

1 Motivazioni

Il seguente progetto si pone lo scopo di identificare una serie di cibi facendo uso di modelli visti durante il corso di studio. Tale tipo di riconoscimento può risultare molto utile per quanto riguarda la classificazione di piatti in tutto il mondo. Si pensi ad esempio a viaggiatori o stranieri che vogliono avere maggiori informazioni su alcuni piatti o per coloro che sono interessati a conoscere i valori nutrizionali del cibo proposto, il tutto con una sola foto.

2 Stato dell'arte

L'applicazione maggiormente conosciuta per quanto riguarda il riconoscimento di cibi è al momento *Calorie Mama* [4]. Tale applicazione è disponibile per Apple e Android e permette non solo di riconoscere i cibi ma anche di mostrarne i valori nutrizionali e di far gestire all'utente le calorie assunte giornalmente. La funzione di *instant food recognition* viene alimentata da *Food AI API* [9] basata sulle ultime innovazioni in campo di deep learning e in grado di riconoscere ad oggi 756 cibi diversi (gran parte cibi tipici di Singapore). Ogni piatto viene poi legato a specifici valori nutrizionali che l'utente utilizza per controllare le proprie diete direttamente dall'app.

3 Obbiettivi

Il mio progetto non si pone di superare i risultati già raggiunti dall'applicazione nè da *Food AI API*, ma di eseguire un'analisi sulle migliori tecniche di classificazione conosciute e decretare la più efficiente tra queste. In particolare il mio lavoro si è concentrato sull'analisi di tre principali metodi per la classificazione supervisionata: KNN (*K-Nearest Neighbors*), SVM (*Support Vector Machine*) e reti neurali.

4 Metodologia

Il lavoro in sé si è suddiviso nella ricerca di un dataset, con relativa estrazione dei dati e delle features, e nell'implementazione di alcuni dei modelli di riconoscimento visti durante il corso. Si spiegano ora nei dettagli queste tre diverse fasi.

4.1 Ricerca del dataset

Il dataset scelto, denominato *Food-101* [5], è disponibile sul sito [kaggle.com](https://www.kaggle.com) e presenta un totale di 101000 fotografie di piatti e cibi diversi. In particolare, l'insieme è suddiviso in 101 categorie di cibi, ognuna composta da 1000 foto. La classe di appartenenza è deducibile dalla cartella in cui essa è contenuta.

Per tutti e tre i metodi di riconoscimento usati si è fatto uso di un campione di sole 10 classi, prendendo 100 immagini ciascuna per la fase di training e 10 per quella di testing, ridimensionate in un formato 64x64. La scelta di questo insieme ridotto di immagini ha permesso di svolgere le operazioni in tempi relativamente brevi. Solo nel caso delle reti neurali si sono presi in considerazione per ogni categoria anche un training set di 750 immagini e un testing set di 250, prendendo le foto in input con due formati diversi, 64x64 e 128x128. Negli altri due casi invece si è visto fin da subito che l'utilizzo di un campione più grande non portava alcun tipo di miglioramento e si sono evitati ulteriori test.

Le immagini del training e testing set sono state estratte attraverso l'uso della libreria Pandas [12] e della sua funzione *read_csv* per leggere i file di testo in cui erano presenti i nomi delle immagini da importare. Ognuna di queste foto è stata poi introdotta nel codice tramite la libreria OpenCV [1], utilizzando rispettivamente le funzioni *imread* e *imresize* per leggere e ridimensionare le foto (nel formato specificato sopra).

4.2 Estrazione delle features

Per l'estrazione delle features si è fatto uso di una rete neurale disponibile tra i modelli di Torchvision [7] e già richiamata tramite il file *resnet.py* rilasciato per questo progetto. Tale modello è il ResNet-50 ed è stato definito a partire dalla ricerca *Deep Residual Learning for Image Recognition* [2].

ResNet-50 è stato usato per i primi due metodi di riconoscimento usati (KNN e SVM), mentre alla rete neurale definita successivamente sono state date direttamente in pasto le immagini del dataset (nel formato specificato sopra).

L'estrazione nel codice è avvenuta attraverso la creazione di una classe *FeatureExtractor* (definita in *resnet.py*) e di un successivo richiamo alla funzione *getFeatures* per ottenere le caratteristiche specifiche di ogni immagine. Per minimizzare ulteriormente le tempistiche, una volta estratte le features per il dataset, quest'ultime vengono salvate in file .csv così da essere recuperate in modo più veloce nelle esecuzioni successive. Per fare ciò ci si è avvalsi delle librerie *csv* e *Numpy* [6] [10] rispettivamente per la scrittura e la lettura dei dati.

4.3 Metodi di riconoscimento usati

I metodi di riconoscimento implementati sono i seguenti.

KNN Il metodo dei *K-Nearest Neighbors* è stato costruito utilizzando diversi tipi di metriche e un diverso numero di vicini (*K*) per l'attribuzione delle immagini a una certa categoria. In particolare si è fatto uso delle seguenti metriche per il calcolo delle distanze tra le features:

- distanza euclidea: $\|u - v\|$;

- distanza di Minkowski: $\|u - v\|_p$;
- distanza del coseno: $1 - \frac{u \cdot v}{\|u\|_2 \cdot \|v\|_2}$;
- correlazione: $1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{\|(u - \bar{u})\|_2 \cdot \|(v - \bar{v})\|_2}$;

Per ognuna delle precedenti se ne è calcolata l'efficienza per K pari a 1, 3 e 7.

Dal punto di vista dell'implementazione, si è fatto uso delle seguenti funzioni e classi:

- `sklearn.preprocessing.StandardScaler` [8] per standardizzare le features in ingresso prima di darle in pasto al sistema;
- `scipy.spatial.distance.cdist` [11] per il calcolo della distanza tra una coppia di features;
- `scipy.stats.mode` [11] per trovare tra i K vicini la classe più frequente.

SVM Per l'implementazione della *Support Vector Machine* si sono presi in considerazione anche in questo caso alcuni kernel diversi:

- lineare;
- polinomiale, con grado pari a 3, 5 e 7;
- RBF (*Radial Basis Function*), usando diversi valori per *gamma* ($\frac{1}{n_features}$ se posta uguale a *auto* o $\frac{1}{n_features \times set_variance}$ se posta su *scale*) e di *C* (0.1 e 1).

Per tutti i casi si è testato il modello su 10, 100 e 1000 iterazioni totali.

Dal punto di vista dell'implementazione, si è fatto uso delle seguenti funzioni e classi:

- `sklearn.preprocessing.StandardScaler` [8] (come in precedenza);
- `sklearn.svm.SVC` [8]: classe che definisce un modello per la classificazione con SVM;
- `sklearn.svm.SVC.fit`: funzione che adatta il modello definito in base ai dati di training;
- `sklearn.svm.SVC.predict_proba`: probabilità di appartenenza di un campione al modello sopra definito.

DNN La rete neurale è stata creata ad hoc per il dataset e comprende 6 diversi strati:

1. convoluzione 2D con kernel di dimensione 3x3, passando da 3 canali in input (RGB) a 6 finali;
2. *max-pooling* di dimensione 2x2;
3. seconda convoluzione 2D con uguale kernel (3x3), passando da 6 canali in input a 16 in output;
4. trasformazione lineare che prende come features in input l'insieme dei valori che riguardano l'immagine come finora è stata modificata, cioè con 16×14 (o 30×14 o 30×30) (*numero_di_canali* \times *altezza* \times *larghezza*), che confluiscono in 2048 features in output;
5. seconda trasformazione lineare che riduce le features in 1024;
6. terza e ultima trasformazione lineare che da 1024 features passa a sole 10 che rappresentano il numero di classi finali di appartenenza. La feature tra le 10 finali che presenta il valore più alto sarà identificata come la classe di appartenenza.

Tale configurazione è stata ispirata da quella presente in nel tutorial di PyTorch: [Training a classifier](#). Il tasso di apprendimento è stato impostato tra 0.001 e 0.01 e il numero di batch a 4. La fase di addestramento è stata fatta proseguire per 2, 5 e 10 epoche.

Per la costruzione della rete neurale è stato utilizzato il framework PyTorch [7] e in particolare:

- *torch.utils.data.DataLoader*: produce il dataset sotto forma di tensore su cui iterare;
- *torch.true_divide* per normalizzare i valori in input dei pixel (da 0-255 a 0-1);
- *torch.nn.Module*, *torch.nn.Conv2d*, *torch.nn.MaxPool2d*, *torch.nn.Linear*, *torch.nn.functional.relu* per definire la topologia della rete neurale;
- *torch.nn.CrossEntropyLoss* per definire la funzione di perdita e la relativa funzione di *backwards*. In generale, una *Cross-entropy loss* si occupa di misurare le performance di un classificatore i cui output (per ogni classe) sono rappresentati da un valore di probabilità tra 0 e 1. Il valore di perdita ritornato incrementa col divergere del valore predetto da quello dell'etichetta effettiva;
- *torch.optim.SGD* per stabilire l'ottimizzatore (basato sulla discesa stocastica del gradiente) e le sue funzioni *step* e *zero_grad*. In questo modo i pesi vengono aggiornati secondo la regola: $weight = weight - learning_rate \times gradient$.

Si è fatto inoltre uso della funzione *np.random.shuffle* [6] [10] per randomizzare l'ordine delle immagini in input all'interno della rete.

5 Esperimenti e risultati raggiunti

Sfortunatamente, i test eseguiti sui tre diversi metodi di riconoscimento non hanno condotto a ottimi risultati. Il miglior valore di accuratezza, raggiunto tramite il metodo dei *K-Nearest Neighbors*, è stato del 20%, con valori di precisione e recall rispettivamente pari a 0.33 e 0.34).

In tutti e tre i casi le performance sono state calcolate in base ad accuratezza e secondo i valori di *precision* e *recall*, estratti dalla matrice di confusione calcolata a partire dalle predizioni. Inoltre, alla fine tutte e tre le classificazioni, un set di immagini randomiche viene stampato (tramite la libreria *Matplotlib* [3]) con le relative etichette trovate dai tre classificatori e i suoi valori *ground truth*.

Da ogni metodo, in particolare, sono stati raggiunti i seguenti risultati.

5.1 KNN

Metric \ K	1	3	7
Euclidean	18%	17%	9%
Minkowski	18%	17%	9%
Cosine	20%	19%	12%
Correlation	19%	16%	14%

Tabella 1: Risultati ottenuti per il modello KNN.

Si ha quindi che la maggior accuratezza nei risultati (20%) si ha utilizzando come metrica la distanza del coseno e come 1 come numero di vicini considerati.

5.2 SVM

Nel caso della *Support Vector Machine* si distinguono i risultati ottenuti con kernel lineare e polinomiale (praticamente uguali per grado pari a 3, 5 o 7) e quelli raggiunti tramite RBF (*Radial Basis Function*), suddivisi in base ai valori scelti per γ e per C.

Kernel \ Iterations	10	100	1000
Linear	10%	14%	15%
Polynomial	10%	7%	5%

Tabella 2: Risultati ottenuti con la SVM con kernel lineare e polinomiale.

γ, C \ Iterations	10	100	1000
$\gamma = scale, C = 1$	15%	18%	16%
$\gamma = scale, C = 0.1$	8%	17%	16%
$\gamma = auto, C = 1$	8%	16%	14%
$\gamma = auto, C = 0.1$	9%	13%	16%

Tabella 3: Risultati ottenuti con la SVM con kernel RBF.

Come si può notare dalle tabelle il risultato migliore, 18% di accuratezza, viene ottenuto usando un kernel RBF, con $\gamma = scale$ e $C = 1$ (valori di default). In questo caso si ha un *precision* = 0.33 e una *recall* = 0.60.

5.3 DNN

Per addestrare la rete neurale sono state fatte molte prove differenti, in base al numero di immagini considerate (100/10 o 750/250), in base alla dimensione di quest'ultime (64x64 o 128x128), in base al numero di epoche (2, 5 e 10) e infine rispetto al *learning rate*. Per semplificare maggiormente le tabelle contenenti di valori di accuratezza risultati da questi esperimenti vengono riportati di seguito i valori maggiori ottenuti con uno tra i due *learning rate* presi in considerazione, ovvero 0.01 e 0.001.

Size \ Epochs	2	5	10
64x64	11%	11%	10%
128x128	7%	16%	16%

Tabella 4: Risultati ottenuti con il NN con un train set di 100 immagini e un test set di 10.

Size \ Epochs	2	5	10
64x64	12,60%	12,20%	10,60%
128x128	12,16%	11,16%	10,64%

Tabella 5: Risultati ottenuti con il NN con un train set di 750 immagini e un test set di 250.

Utilizzando questo tipo di classificazione quindi ci si aspetta al massimo un 16% di accuratezza, con valori di precisione e recall rispettivamente pari a 0.22 e 0.40.

6 Conclusioni

Le sperimentazioni fatte utilizzando questi tre diversi metodi di classificazione fanno capire, nonostante i bassi valori di accuratezza raggiunti, come talvolta sistemi più semplici quali i *K-Nearest Neighbors* possono condurre a risultati migliori delle più complesse reti neurali. Eseguendo diversi test sui tre classificatori si può inoltre notare come, impostati gli stessi identici input, mentre il metodo dei KNN riporta sempre gli stessi risultati, lo stesso non vale per gli altri due. La SVM infatti può ritornare dei valori di accuratezza che di volta in volta variano anche del 10%, mentre la rete neurale oscilla su un range del 5% di accuratezza rispetto ai valori riportati in tabella.

Riferimenti bibliografici

- [1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
- [3] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.
- [4] Azumio Inc. Calorie mama, 2017.
- [5] K Scott Mader. Food-101, 2018.
- [6] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.
- [7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [9] Prof. Steven HOI R&D team. Foodai.
- [10] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22, 2011.
- [11] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayоров, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- [12] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.