

Few Shot Learning for cancer detection

Marco Russo

Il problema

- Il cancro rimane una delle sfide sanitarie più importanti al mondo, responsabile di più 8 milioni di morti l'anno secondo la World Health Organization.
- Una diagnosi tempestiva è fondamentale per migliorare la prognosi e aumentare le possibilità di successo del trattamento.
- Molto spesso non è possibile identificare un punto di origine per via di metastasi molto estese

Lavori Correlati

PCAWG Consortium

- Il Pan-Cancer Analysis of Whole Genome è un gruppo di ricerca internazionale i cui obiettivi è capire le basi genetiche dei tumori.
- L'obiettivo principale del PCAWG è di identificare pattern di mutazioni genetiche e alterazioni in cellule tumorali che possono essere usate per nuove cure e strumenti di diagnosi.
- Sono riusciti a generare un patrimonio di informazioni genetiche riguardo ai tumori e sono state rese disponibili a tutti i ricercatori interessati

Lavori Correlati

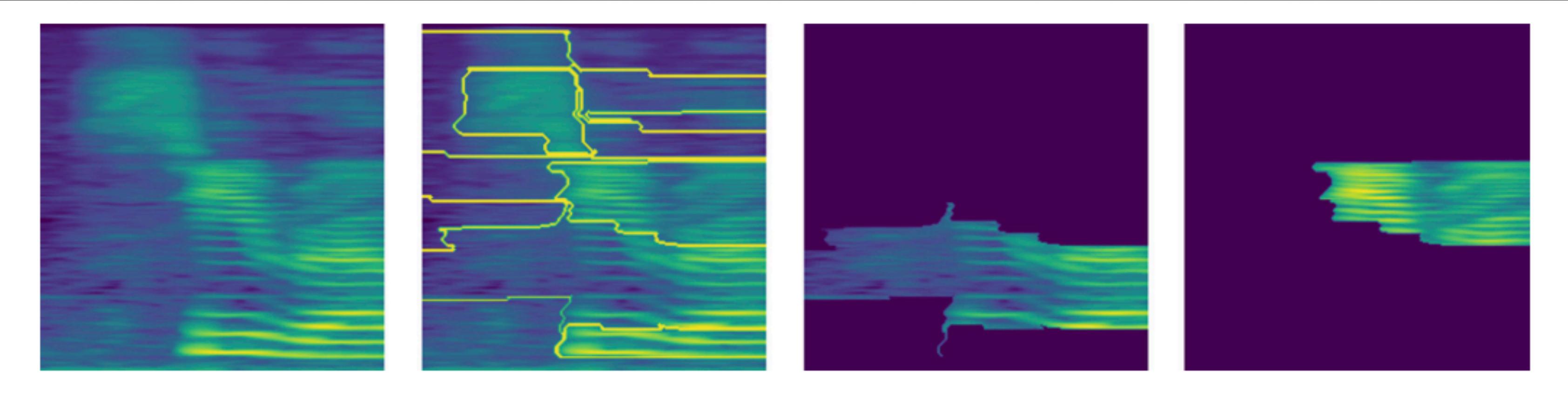
PCAWG Consortium

- Utilizzando i dati a loro disposizione, hanno addestrato un random forest classifier ed una rete neurale per identificare l'origine del tumore
- I dati usati sono sequenziamenti genetici di cellule tumorali provenienti dalla metastasi del cancro.
- Entrambi i modelli sono molto rudimentali poiché in questo lavoro è presente una grossa enfasi sull'acquisizione ed il preprocessing dei dati

Lavori Correlati

Few Shot Learning for Audio Data

- In questo lavoro dei colleghi di Pisa, è stata utilizzata una rete Siamese per evidenziare le differenze tra due tracce audio



Rete Siamese

- Una rete siamese è un tipo di rete neurale artificiale che è stata progettata per confrontare e valutare la somiglianza tra due oggetti di ingresso
- Questi rami condividono gli stessi pesi e architettura, permettendo loro di estrarre le stesse caratteristiche dall'input

Rete Siamese

- La rete siamese è spesso utilizzata per risolvere problemi di riconoscimento di pattern, di identificazione delle impronte digitali, di riconoscimento facciale, di matching di testo e di molte altre applicazioni di apprendimento automatico
- Una delle caratteristiche principali della rete siamese è che, invece di produrre un'uscita binaria (es. "match" o "non-match") come altre reti di classificazione, essa produce una distanza o una similarità tra i due input
- Questa distanza può poi essere utilizzata per la classificazione, il clustering o altre attività di elaborazione dei dati.

Few Shot Learning

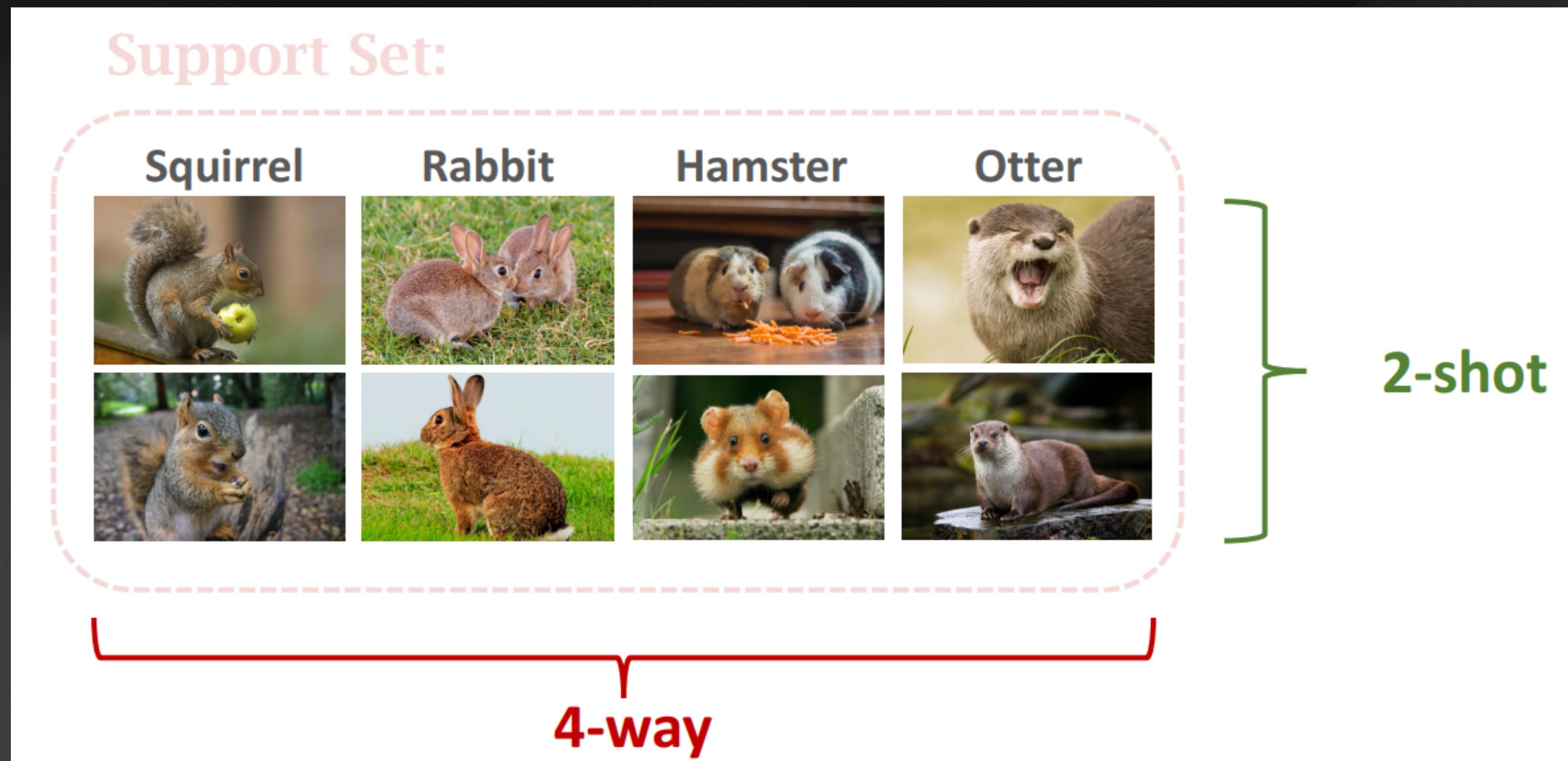
- Il few-shot learning è una tecnica di apprendimento automatico che mira a risolvere il problema di apprendere da pochi esempi di dati di training.
- In altre parole, invece di avere un grande set di dati di training disponibili, si assume di avere solo pochi esempi per addestrare il modello (support set)
- Il few-shot learning si basa sulla capacità del modello di estrarre conoscenza generale dai dati di training limitati e di applicarla a nuovi dati.
- Ciò significa che il modello deve essere in grado di generalizzare oltre i dati di training e di imparare ad adattarsi a nuove situazioni in modo efficiente.

Few Shot Learning

- Per raggiungere questo obiettivo, le tecniche di few-shot learning spesso utilizzano reti neurali che sono state pre-addestrate su un set di dati più grande (pretraining)
- Questo aiuta a catturare le caratteristiche generali dei dati di input e ad adattarsi rapidamente a nuovi dati di input

Few Shot Learning

Support Set



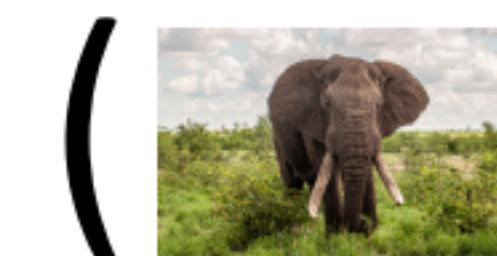
Few Shot Learning

Training set

Positive Samples

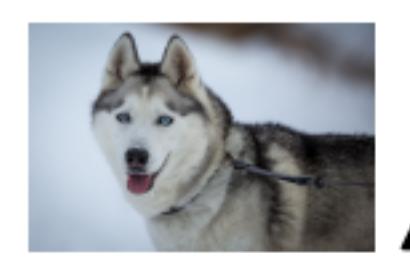
(, , 1)

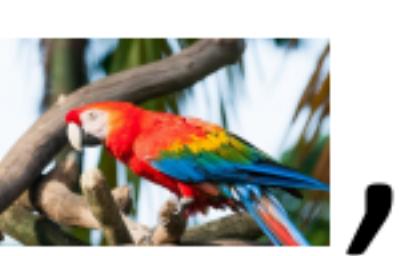
(, , 1)

(, , 1)

Negative Samples

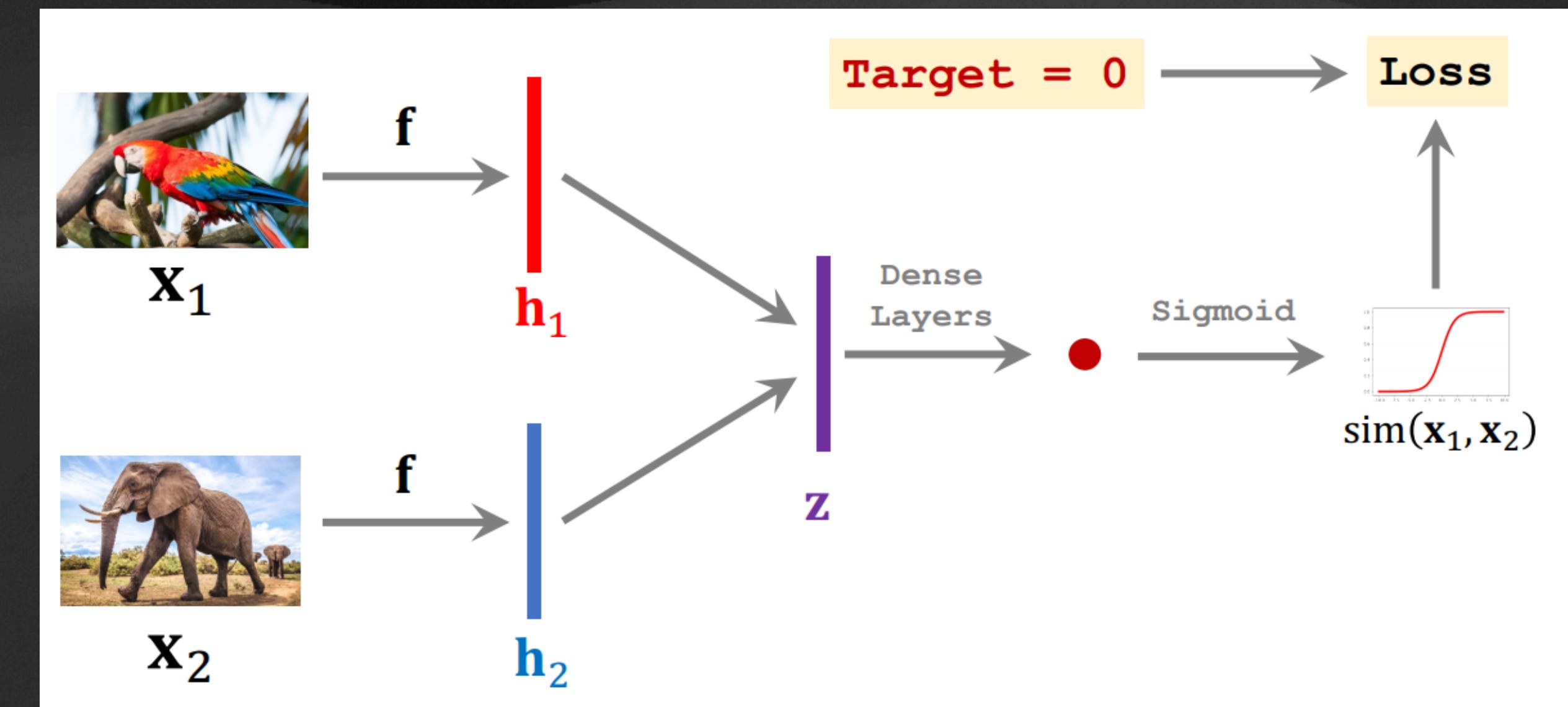
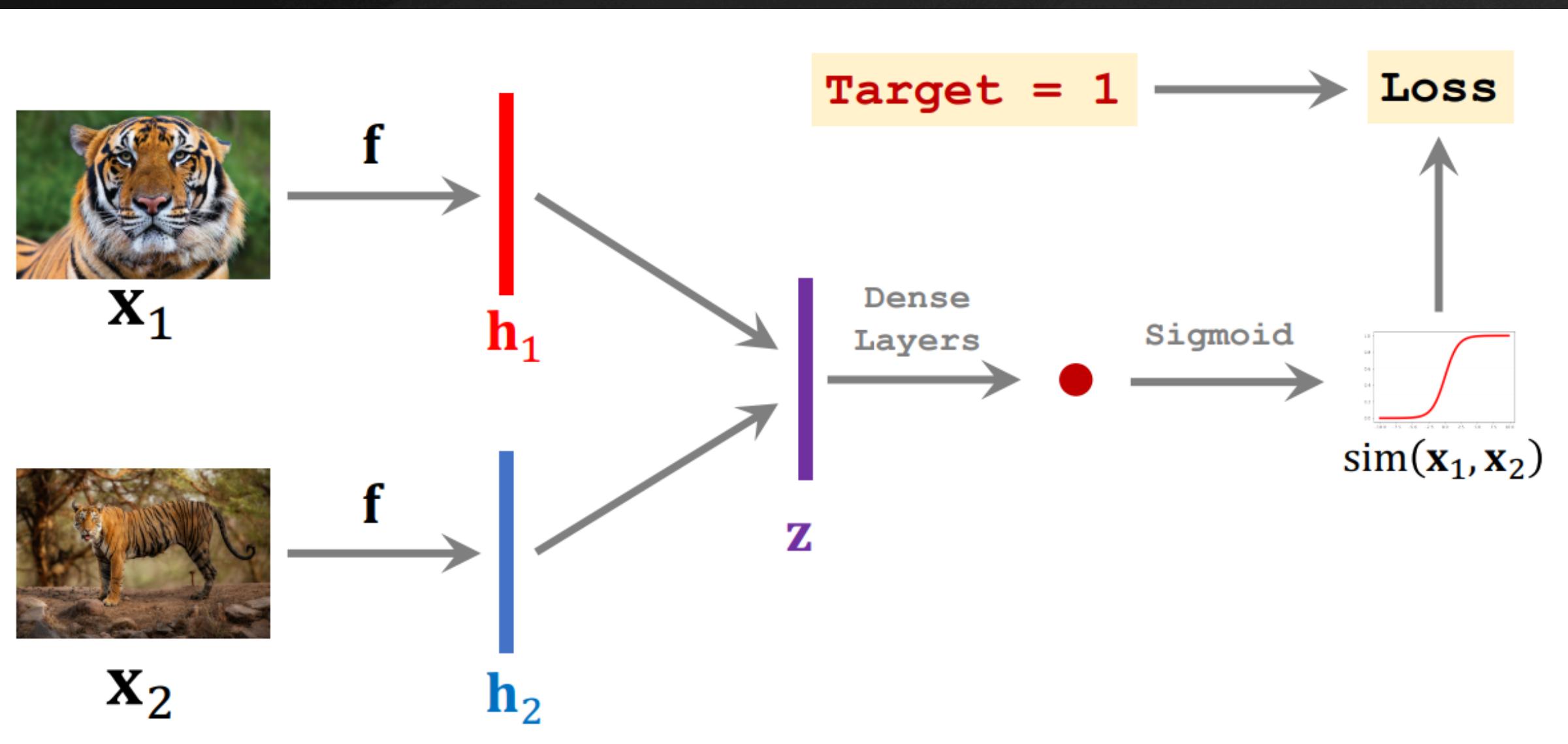
(, , 0)

(, , 0)

(, , 0)

Few Shot Learning

Training



Few Shot Learning

One-Shot Prediction

Query:



sim = 0.2

Fox



sim = 0.9

Squirrel



sim = 0.7

Rabbit



sim = 0.5

Hamster



sim = 0.3

Otter



sim = 0.4

Beaver



Lavoro Svolto

Dataset

- Per il lavoro svolto è stato utilizzato il dataset MSK-Impact
- MSK-IMPACT (Integrated Mutation Profiling of Actionable Cancer Targets) è un test genetico sviluppato dal Memorial Sloan Kettering Cancer Center (MSK) per l'analisi dei tumori
- Contiene i dati del profilo molecolare di oltre 30.000 pazienti affetti da tumori di diversi tipi di cancro, tra cui, il cancro al seno, ai polmoni, alla prostata e al colon-retto.

Preprocessing

Il dataset

- Il dataset è l'unione di 3 file di MSK-Impact:
 - Clinical Data (contiene i dati clinici dei pazienti ed il tipo di cancro diagnosticato)
 - SV (variazioni strutturali riscontrate durante il sequenziamento del campione di cellule tumorali estratto dal paziente)
 - CNA (I valori delle Copy Number Alterations del paziente)

Preprocessing

- Rimozione righe “spurie” (righe con almeno una colonna vuota)
- Rimozione righe appartenenti a tumori outliers (n campioni < 30)
- Cancellazione di colonne non pertinenti
- One hot encoding delle colonne con valori testuali
- Riduzione della dimensionalità con la Principal Component Analysis (PCA)
- One hot encoding dei label

La Proposta

- Modello DNN preaddestrato sui dati per l'estrazione delle features dai dati
- Rete Siamese con 2 Shot Learning per valutare se due pazienti hanno lo stesso tipo di cancro o no
- Il modello è stato sviluppato utilizzando tensorflow e keras

La Proposta

Modello DNN

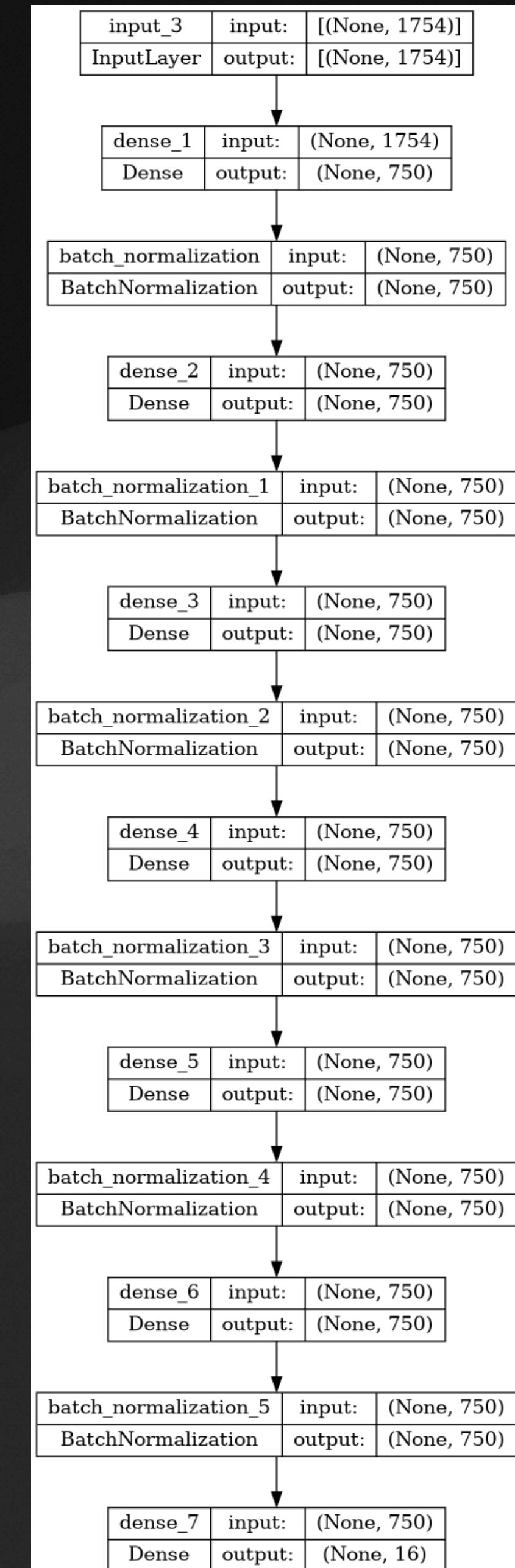
- Il modello DNN viene inizialmente addestrato sul nostro dataset
- Viene addestrato per classificare tutti i tipi di cancro del dataset (classificatore multiclass)
- Alla fine del training viene rimosso il layer di classificazione e “congelati” tutti i pesi dei neuroni, in modo da non poterlo addestrare più

La Proposta

Modello DNN

```
input_layer = Dense(1754,activation ='relu',input_shape=x_train[0].shape)(input1)

hidden_layers = Dense(750,activation='relu')(input_layer)
hidden_layers = BatchNormalization()(hidden_layers)
hidden_layers = Dense(750,activation='relu')(hidden_layers)
hidden_layers = BatchNormalization()(hidden_layers)
hidden_layers = Dense(n_classes,activation ='softmax')(hidden_layers)
```



La Proposta

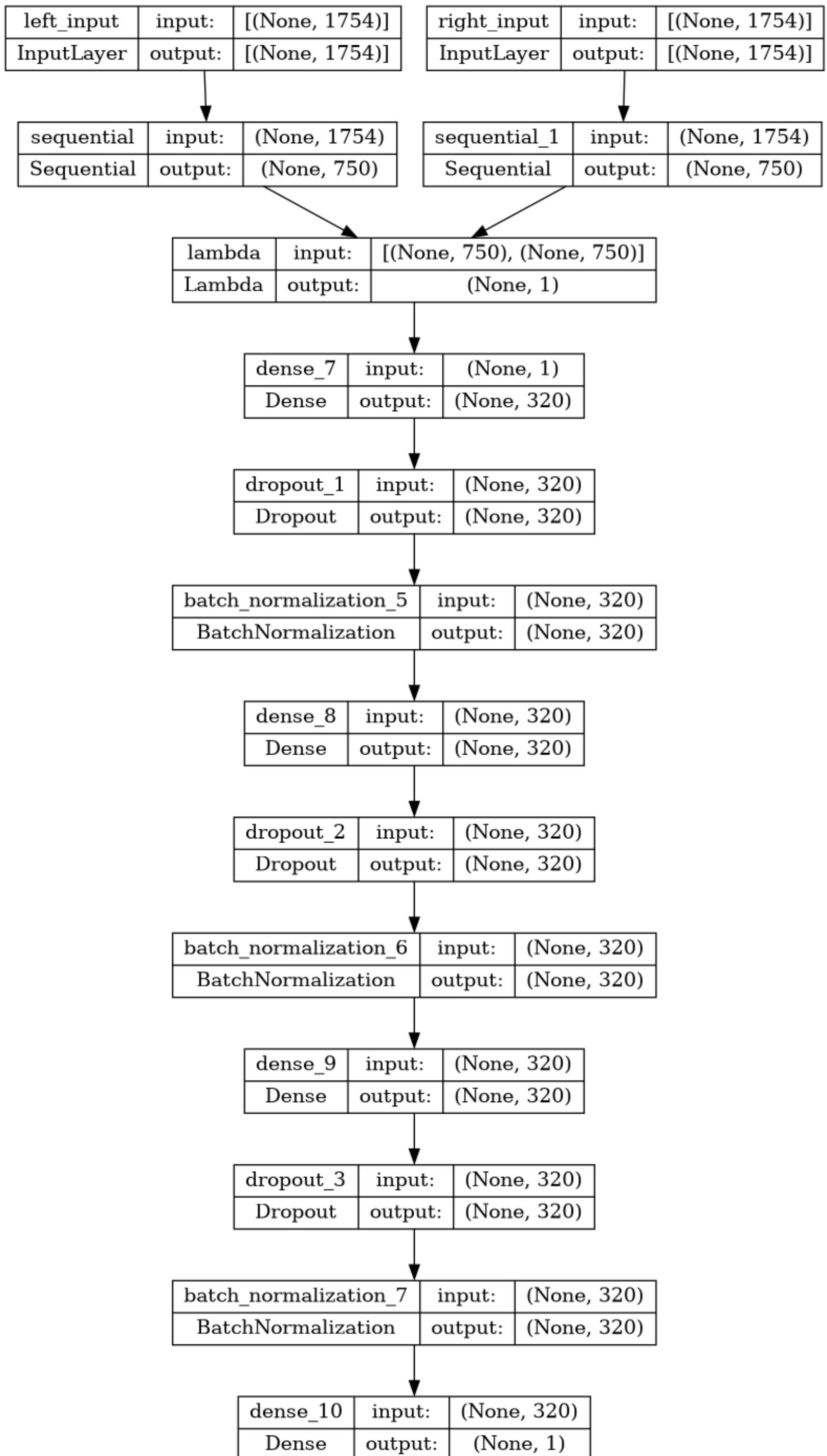
Training DNN

Cancer	Loss	Accuracy	Precision	Recall	AUC	F1 Score
Non-Small Cell Lung Cancer	0.222917	0.962963	0.962500	0.950617	0.993014	0.956522
Breast Cancer	0.403626	0.931035	0.929825	0.913793	0.972661	0.921739
Bladder Cancer	0.907647	0.833333	1.000000	0.750000	0.945833	0.857143
Head and Neck Cancer	3.557362	0.500000	0.555556	0.500000	0.747333	0.526316
Bone Cancer	0.874127	0.785714	0.846154	0.785714	0.957313	0.814815
Soft Tissue Sarcoma	2.417866	0.629630	0.680000	0.629630	0.866530	0.653846
Prostate Cancer	0.011922	1.000000	1.000000	1.000000	1.000000	1.000000
Ovarian Cancer	2.306796	0.375000	0.500000	0.375000	0.852083	0.428571
Glioma	0.017366	1.000000	1.000000	1.000000	1.000000	1.000000
Esophagogastric Cancer	1.004104	0.833333	0.833333	0.833333	0.954398	0.833333
Hepatobiliary Cancer	0.395946	0.944444	0.944444	0.944444	0.970988	0.944444
Colorectal Cancer	0.110219	0.958333	0.956522	0.916667	0.999653	0.936170
Cancer of Unknown Primary	1.058346	0.777778	0.875000	0.777778	0.930041	0.823529
Endometrial Cancer	1.526419	0.600000	0.750000	0.600000	0.948000	0.666667
Melanoma	0.315476	0.888889	0.941176	0.888889	0.997737	0.914286
Pancreatic Cancer	0.445821	0.857143	0.850000	0.809524	0.994709	0.829268

La Proposta

Rete Siamese

- I layer sequential sono il modello DNN precedente che viene usato per estrarre le features
- Dopo l'estrazione delle features viene calcolata la distanza euclidea dei due input
- Il risultato passa attraverso altri hidden layer per trovare altre correlazioni tra i dati
- Infine viene calcolata la similarità tramite un classificatore binario



La proposta

Risultati

Cancer	Loss	Accuracy	Precision	Recall	AUC	F1 Score
Non-Small Cell Lung Cancer	0.074384	0.982253	0.960317	0.746914	0.960181	0.840278
Breast Cancer	0.084038	0.973060	0.883721	0.655172	0.975840	0.752475
Bladder Cancer	0.129697	0.953125	0.687500	0.458333	0.925463	0.550000
Head and Neck Cancer	0.221959	0.943750	0.666667	0.200000	0.680583	0.307692
Bone Cancer	0.091722	0.973214	0.785714	0.785714	0.960162	0.785714
Soft Tissue Sarcoma	0.185826	0.945602	0.594595	0.407407	0.791427	0.483516
Prostate Cancer	0.101098	0.969122	0.967033	0.523810	0.960138	0.679537
Ovarian Cancer	0.227238	0.917969	0.368421	0.437500	0.786068	0.400000
Glioma	0.237458	0.943257	1.000000	0.092105	0.707768	0.168675
Esophagogastric Cancer	0.104480	0.973958	0.733333	0.916667	0.971123	0.814815
Hepatobiliary Cancer	0.064282	0.987847	0.939394	0.861111	0.985597	0.898551
Colorectal Cancer	0.082385	0.971354	0.842105	0.666667	0.977836	0.744186
Cancer of Unknown Primary	0.107919	0.961806	0.733333	0.611111	0.969753	0.666667
Endometrial Cancer	0.204249	0.918750	0.333333	0.300000	0.874667	0.315789
Melanoma	0.074745	0.989583	1.000000	0.833333	0.966641	0.909091
Pancreatic Cancer	0.274922	0.936012	0.400000	0.047619	0.532445	0.085106

	Cancer Type	counts
0	Bladder Cancer	42
1	Bone Cancer	54
2	Breast Cancer	242
3	Cancer of Unknown Primary	43
4	Colorectal Cancer	100
5	Endometrial Cancer	32
6	Esophagogastric Cancer	63
7	Glioma	158
8	Head and Neck Cancer	35
9	Hepatobiliary Cancer	70
10	Melanoma	66
11	Non-Small Cell Lung Cancer	313
12	Ovarian Cancer	40
13	Pancreatic Cancer	56
14	Prostate Cancer	336
15	Soft Tissue Sarcoma	104

Possibili sviluppi futuri

- Utilizzare un dataset più completo
- Aumentare la profondità della rete
- Utilizzare support set di grandezza diversa

Grazie per l'attenzione!

Marco Russo