

Multiple Manipulations on Clothes Images with MANN

Francesco Fantechi

Relatore: Alberto Del Bimbo
Corelatore: Federico Becattini

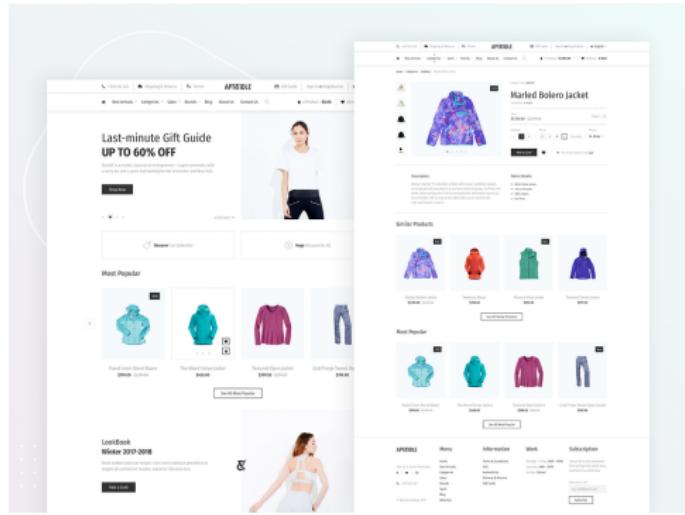


UNIVERSITA' DEGLI STUDI DI FIRENZE
Facolta di Ingegneria
Corso di Laurea Magistrale in Ingegneria Informatica

A.A. 2022-2023

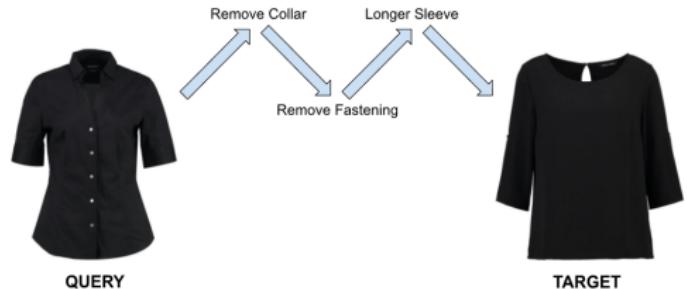
Introduzione

- Compiere manipolazioni su immagini di vestiti per un negozio online di moda



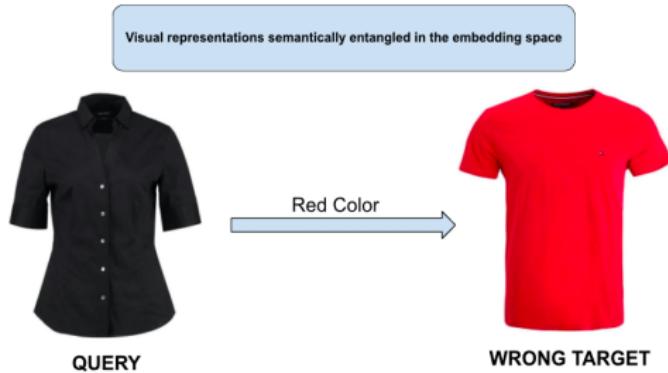
Introduzione

- Compire manipolazioni su immagini di vestiti per un negozio online di moda
 - Esempio: rimuovere il colletto, la chiusura e allungare le maniche dal vestito di sinistra



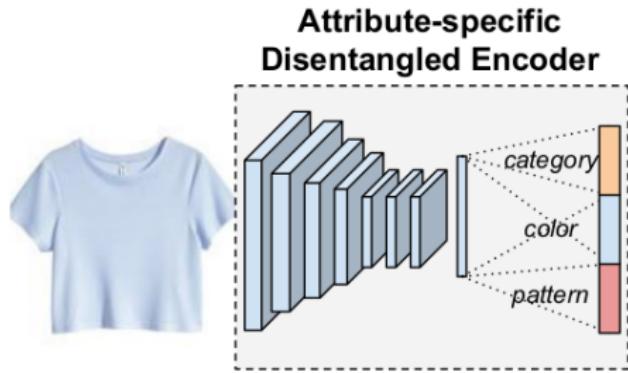
Introduzione

- Compiere manipolazioni su immagini di vestiti per un negozio online di moda
 - Esempio: rimuovere il colletto, la chiusura e allungare le maniche dal vestito di sinistra
- Un'interazione può impattare in modo eccessivo anche altri aspetti dell'immagine stravolgendola più del dovuto



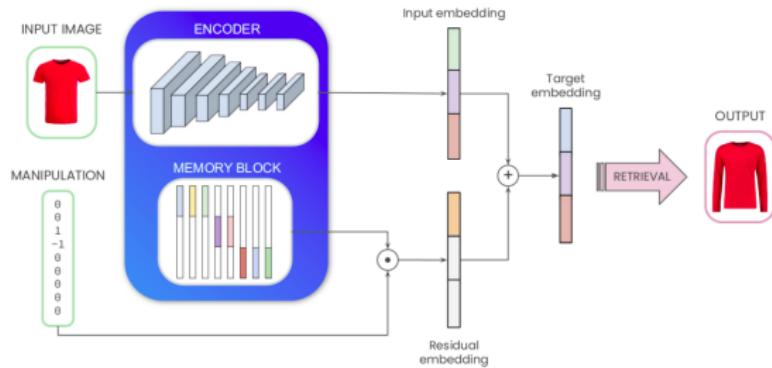
Soluzione proposta da Amazon

- Attribute-Driven Disentangled Encoder (ADDE) per ottenere una rappresentazione visuale meno correlata nello spazio degli attributi



Soluzione proposta da Amazon

- Attribute-Driven Disentangled Encoder (ADDE) per ottenere una rappresentazione visuale meno correlata nello spazio degli attributi
- ADDE-M, ADDE + Memory Block



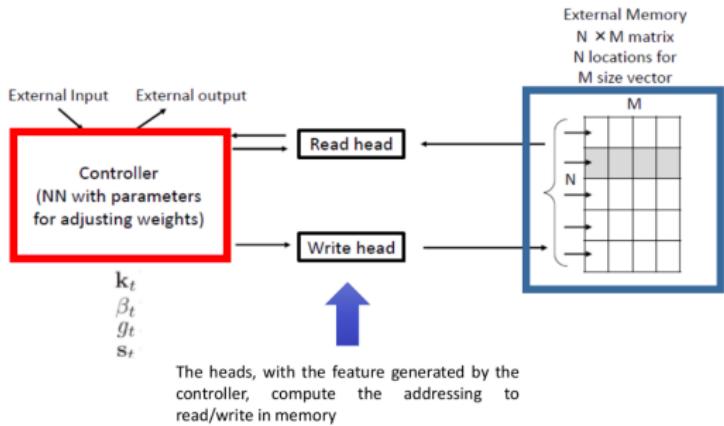
Soluzione proposta da Amazon

- Attribute-Driven Disentangled Encoder (ADDE) per ottenere una rappresentazione visuale meno correlata nello spazio degli attributi
- ADDE-M, ADDE + Memory Block
- Supporta una sola manipolazione alla volta



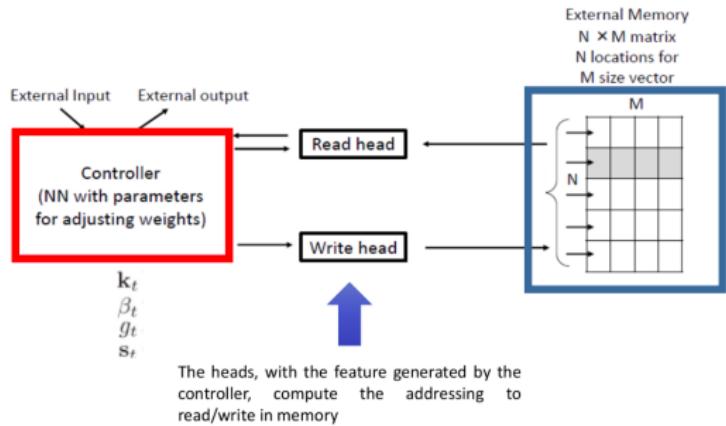
La nostra proposta

- Memory Augmented Neural Network (MANN)



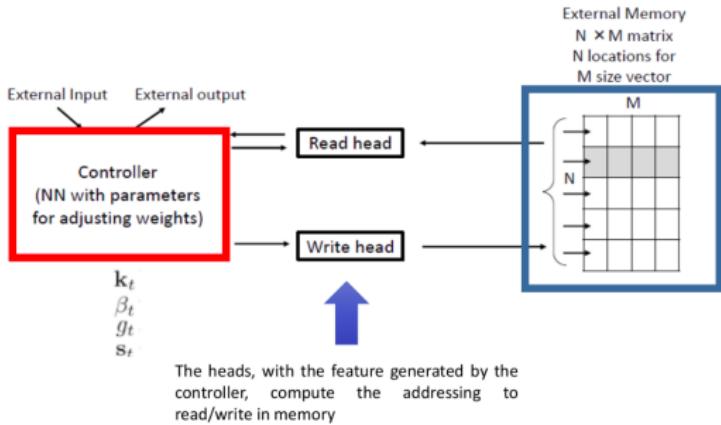
La nostra proposta

- Memory Augmented Neural Network (MANN)
 - Memoria esterna indipendente con il ruolo di base di conoscenza



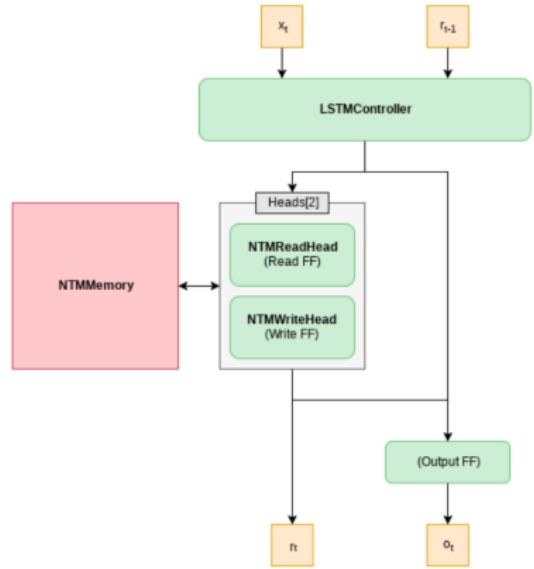
La nostra proposta

- Memory Augmented Neural Network (MANN)
 - Memoria esterna indipendente con il ruolo di base di conoscenza
 - Controller che impara ad interagire con essa leggendo e scrivendo per produrre l'output



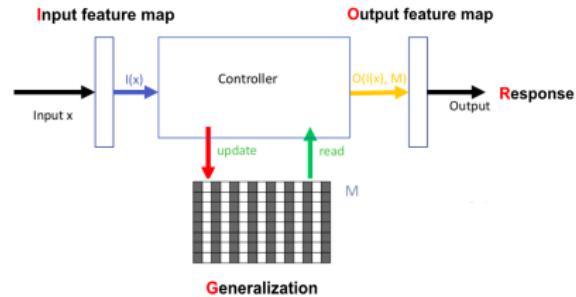
La nostra proposta

- Neural Turing Machine (NTM)



La nostra proposta

- Neural Turing Machine (NTM)
- Modello IGOR

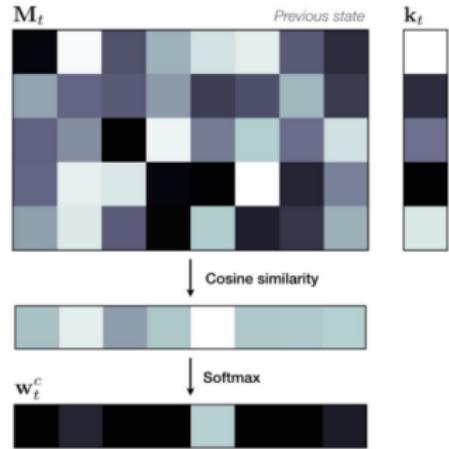


La nostra proposta

- Neural Turing Machine (NTM)
- Modello IGOR
- Content Based Addressing

Content addressing

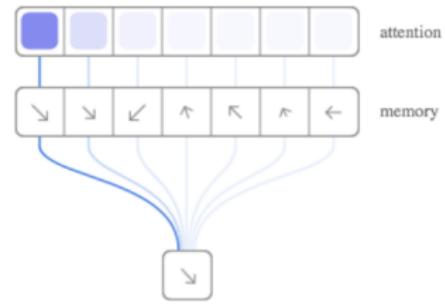
$$w_t^c(i) \leftarrow \text{softmax}(\beta_t \cdot K[\mathbf{k}_t, \mathbf{M}_t(i)])$$



La nostra proposta

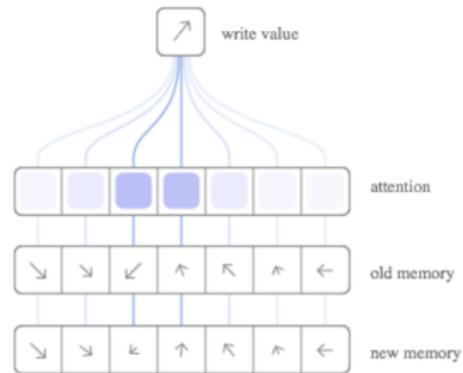
- Neural Turing Machine (NTM)
- Modello IGOR
- Content Based Addressing
- Read:

$$r_t = \sum_{i=0}^{N-1} \omega_t(i) M_t(i)$$



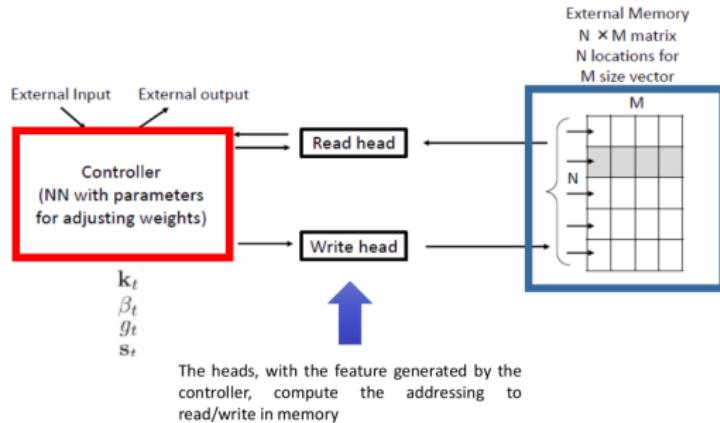
La nostra proposta

- Neural Turing Machine (NTM)
- Modello IGOR
- Content Based Addressing
- Read:
$$r_t = \sum_{i=0}^{N-1} \omega_t(i) M_t(i)$$
- Write:
$$M_t(i) = M_{t-1}(i)(1 - \omega_t(i)e_t) + \omega_t(i)a_t$$



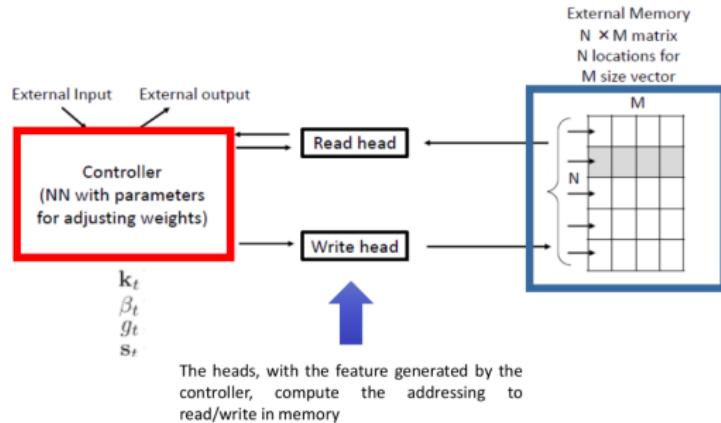
Vantaggi

- Grande memoria indipendente e indirizzabile



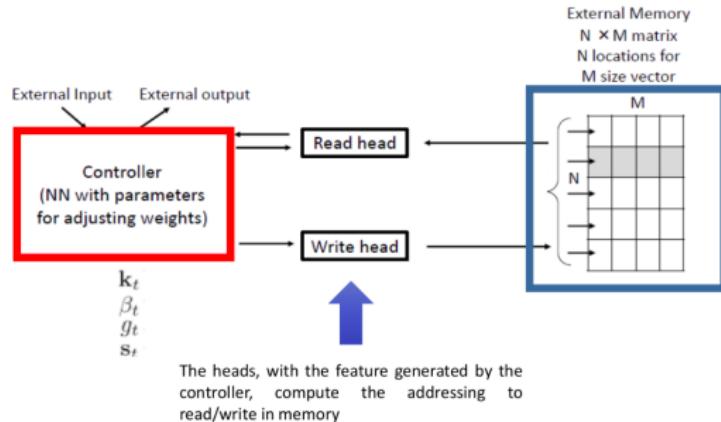
Vantaggi

- Grande memoria indipendente e indirizzabile
 - Supporta $N \geq 1$ manipolazioni in sequenza



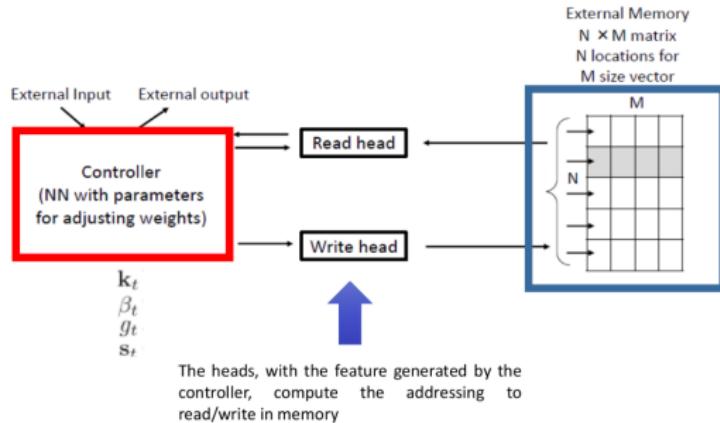
Vantaggi

- Grande memoria indipendente e indirizzabile
 - Supporta $N \geq 1$ manipolazioni in sequenza
 - Maggior supervisione sul comportamento interno della rete



Vantaggi

- Grande memoria indipendente e indirizzabile
 - Supporta $N \geq 1$ manipolazioni in sequenza
 - Maggior supervisione sul comportamento interno della rete
 - Modello semplice e numero di parametri contenuto



Addestramento

- Dataset Shopping100k

Addestramento

- Dataset Shopping100k
 - 80586 immagini di addestramento

Addestramento

- Dataset Shopping100k
 - 80586 immagini di addestramento
 - 20000 immagini di test

- Dataset Shopping100k

- 80586 immagini di addestramento
- 20000 immagini di test
- Ogni immagine è caratterizzata attraverso 12 attributi a ognuno dei quali è associato un certo numero di etichette

Addestramento

- Dataset Shopping100k
 - 80586 immagini di addestramento
 - 20000 immagini di test
 - Ogni immagine è caratterizzata attraverso 12 attributi a ognuno dei quali è associato un certo numero di etichette
- La rete è stata addestrata su migliaia di esempi

Addestramento

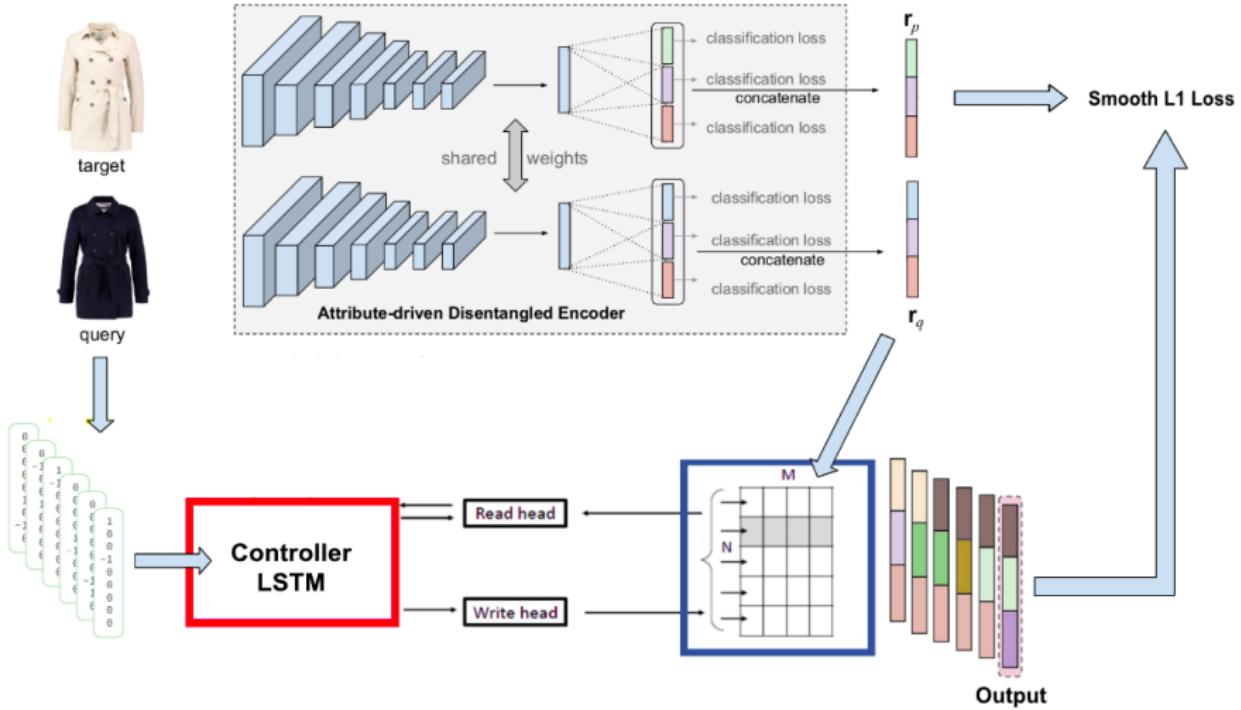
- Dataset Shopping100k
 - 80586 immagini di addestramento
 - 20000 immagini di test
 - Ogni immagine è caratterizzata attraverso 12 attributi a ognuno dei quali è associato un certo numero di etichette
- La rete è stata addestrata su migliaia di esempi
- Ogni esempio (immagine query e target a distanza ≤ 8) è stato generato randomicamente real time di modo da ridurre la possibilità di overfitting

- Dataset Shopping100k
 - 80586 immagini di addestramento
 - 20000 immagini di test
 - Ogni immagine è caratterizzata attraverso 12 attributi a ognuno dei quali è associato un certo numero di etichette
- La rete è stata addestrata su migliaia di esempi
- Ogni esempio (immagine query e target a distanza ≤ 8) è stato generato randomicamente real time di modo da ridurre la possibilità di overfitting
- Ad ogni nuovo esempio la memoria esterna viene inizializzata con il vettore "disentangled" dell'immagine query

- Dataset Shopping100k
 - 80586 immagini di addestramento
 - 20000 immagini di test
 - Ogni immagine è caratterizzata attraverso 12 attributi a ognuno dei quali è associato un certo numero di etichette
- La rete è stata addestrata su migliaia di esempi
- Ogni esempio (immagine query e target a distanza ≤ 8) è stato generato randomicamente real time di modo da ridurre la possibilità di overfitting
- Ad ogni nuovo esempio la memoria esterna viene inizializzata con il vettore "disentangled" dell'immagine query
- Dopo aver avuto in input $N \leq 8$ manipolazioni la memoria contiene il vettore da confrontare con quello dell'immagine target

- Dataset Shopping100k
 - 80586 immagini di addestramento
 - 20000 immagini di test
 - Ogni immagine è caratterizzata attraverso 12 attributi a ognuno dei quali è associato un certo numero di etichette
- La rete è stata addestrata su migliaia di esempi
- Ogni esempio (immagine query e target a distanza ≤ 8) è stato generato randomicamente real time di modo da ridurre la possibilità di overfitting
- Ad ogni nuovo esempio la memoria esterna viene inizializzata con il vettore "disentangled" dell'immagine query
- Dopo aver avuto in input $N \leq 8$ manipolazioni la memoria contiene il vettore da confrontare con quello dell'immagine target
- Utilizzo della libreria Tensorboard per monitorare l'andamento dell'addestramento

Addestramento



Risultati

- Test di Amazon: ≈ 1400 coppie a distanza = 1

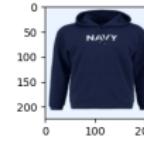
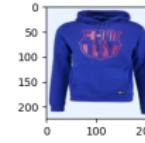
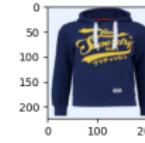
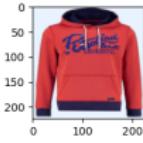
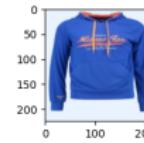
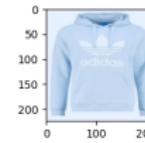
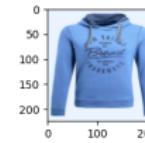
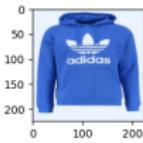
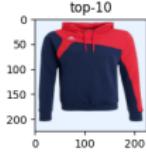
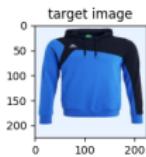
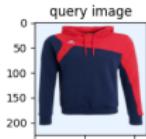
Table: Top-k retrieval accuracies on Shopping100k for attribute manipulation

Shopping100k					
	Top-10	Top-20	Top-30	Top-40	Top-50
AMNet	25.62	36.13	42.94	47.71	51.64
ADDE-M	41.17	52.93	59.81	64.10	67.29
MANN	33.58	44.96	52.00	56.72	60.18

Risultati

- Confronto visuale MANN e ADDE-M con 1 manipolazione

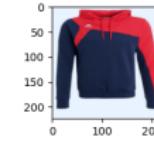
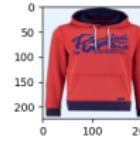
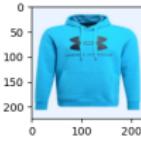
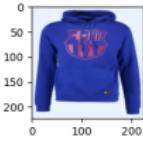
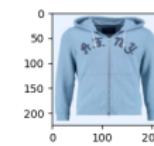
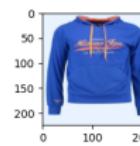
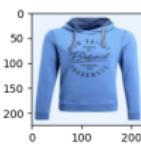
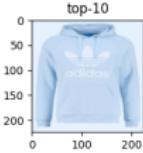
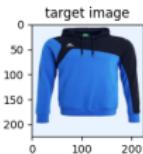
Visual Evaluation MAN - Manipulation: 1
Manipulation in Color attribute, - Navy + Blue



Risultati

- Confronto visuale MANN e ADDE-M con 1 manipolazione

Visual Evaluation ADDE-M - Manipulation: 1
Manipulation in Color attribute, - Navy + Blue



Risultati

- Test di Amazon: ≈ 1400 coppie a distanza = 1

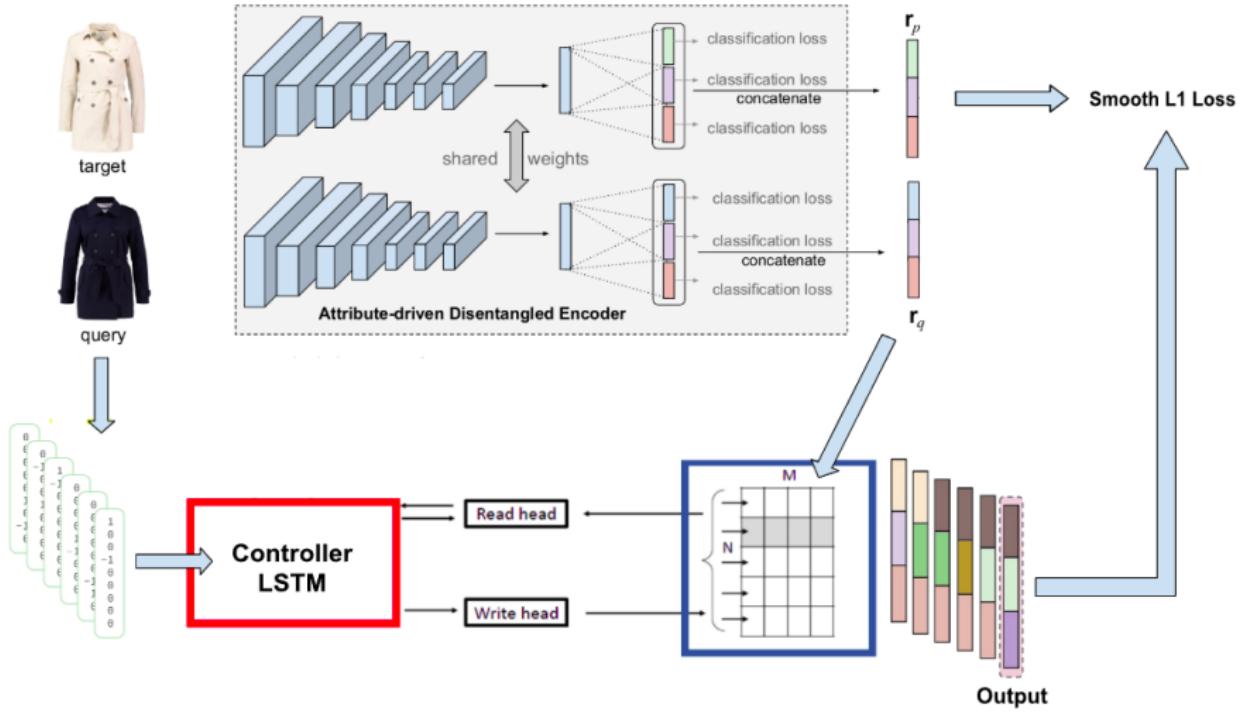
Table: Normalized Discounted Cumulative Gain metrics

Shopping100k		
	ADDE-M	MANN
$NDCG@30$	0.7367	0.7448
$NDCG_t@30$	0.4305	0.3259
$NDCG_o@30$	0.7779	0.7987

- $NDCG = \frac{1}{Z} \sum_{j=1}^k \frac{2^{rel(j)-1}}{\log(j+1)}$
- $rel(j)$ è il punteggio di rilevanza sugli attributi della i-esima immagine (numero di attributi che matchano con il ground-truth diviso per il numero totale di attributi)
- $NDCG_t@30$, $rel(j)$ calcolato solo su attributi manipolati
- $NDCG_o@30$, $rel(j)$ calcolato solo su attributi non manipolati

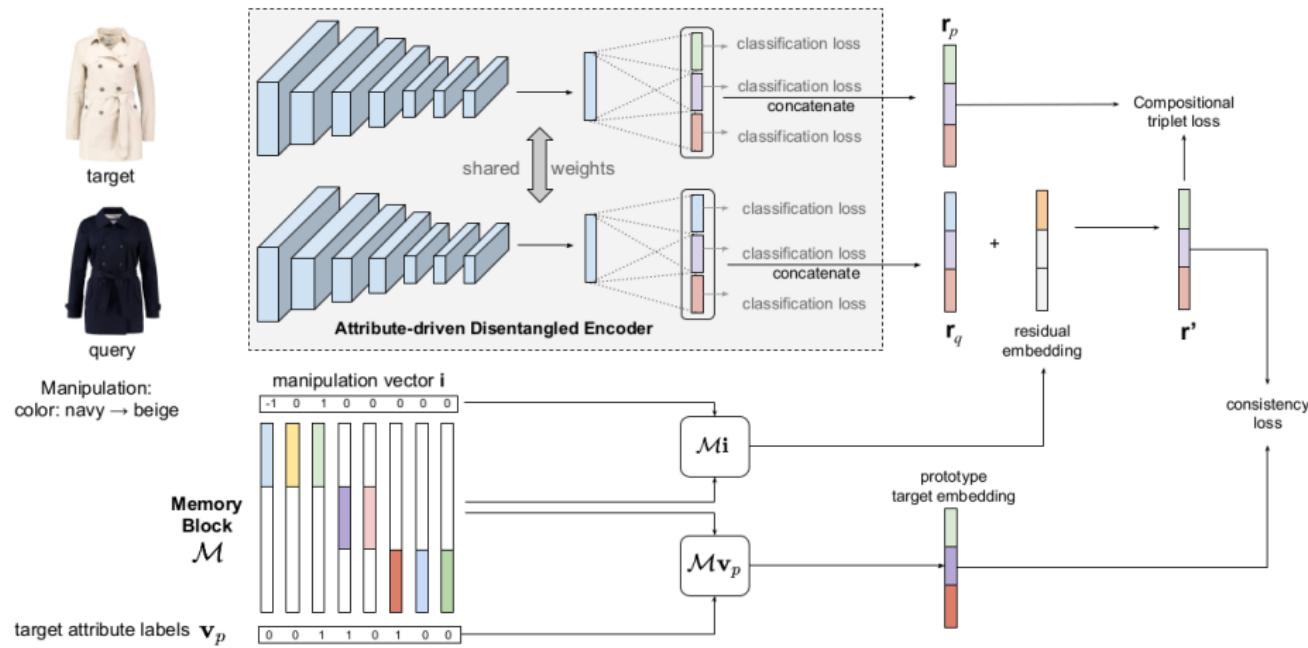
Risultati

- 1 Loss: Smooth L1 Loss



Risultati

- 3 Loss: Triplet Loss, Consistency Loss e Label Triplet Loss



Risultati

- Il nostro test: 1000 coppie a distanza ≤ 8

Table: Top-k retrieval accuracies on Shopping100k for attribute manipulation

Shopping100k					
	Top-10	Top-20	Top-30	Top-40	Top-50
ADDE-M	2.0	4.0	5.0	6.0	7.0
MANN	33.0	45.0	52.00	56.0	60.0

Risultati

- Confronto visuale MANN e ADDE-M con 3 manipolazioni

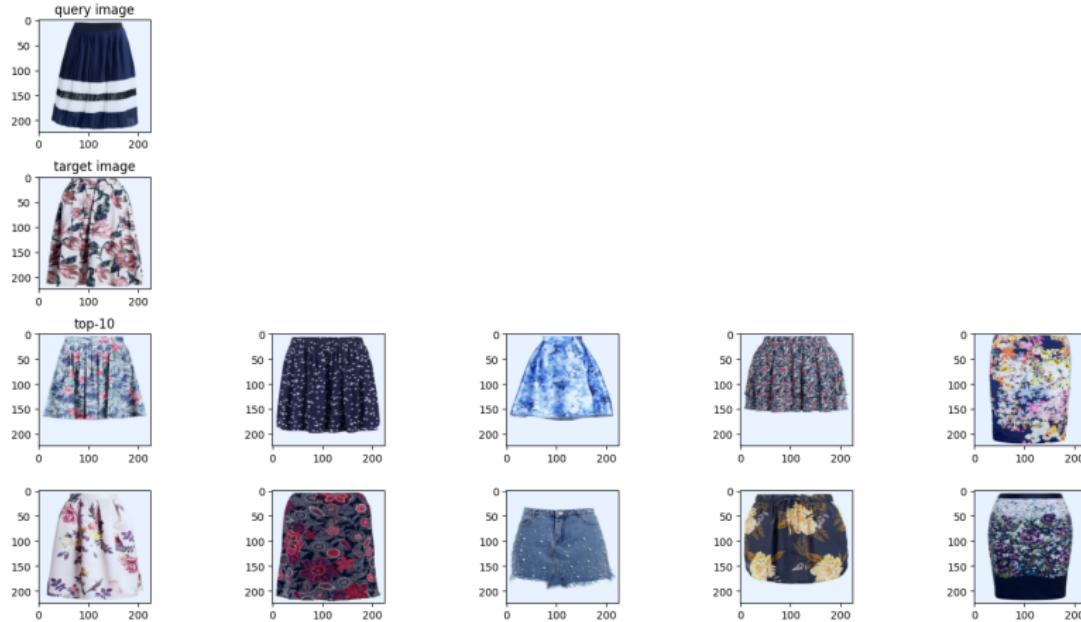
Visual Evaluation MAN - Manipulation: 1
Manipulation in Fabric attribute, + Rib



Risultati

- Confronto visuale MANN e ADDE-M con 3 manipolazioni

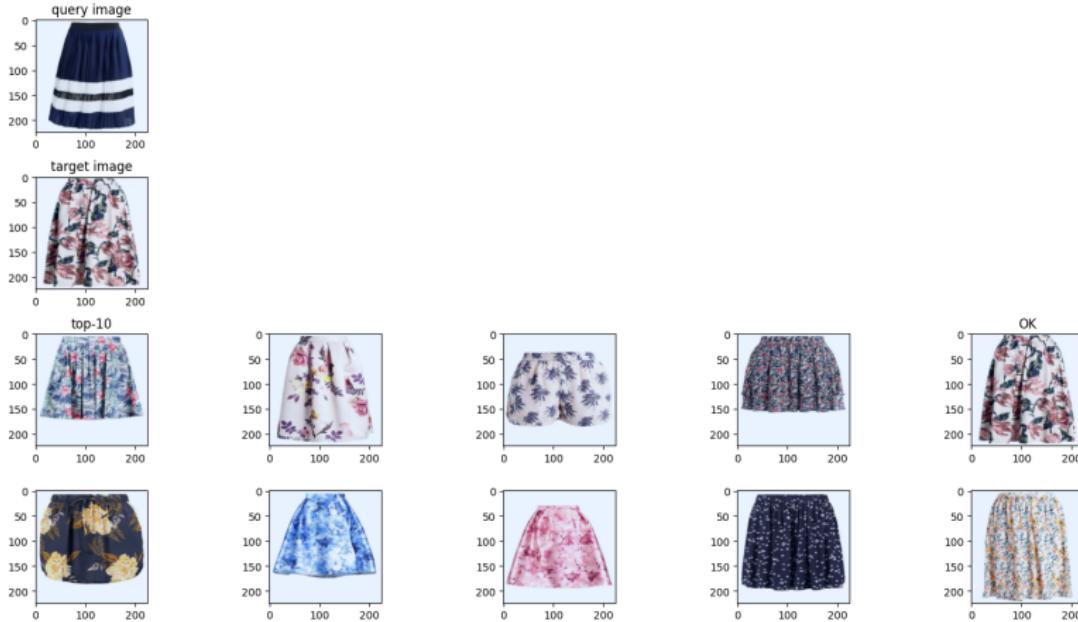
Visual Evaluation MAN - Manipulation: 2
Manipulation in Pattern attribute, - Striped + Floral



Risultati

- Confronto visuale MANN e ADDE-M con 3 manipolazioni

Visual Evaluation MAN - Manipulation: 3
Manipulation in Color attribute, - White + Pink



Risultati

- Confronto visuale MANN e ADDE-M con 3 manipolazioni

Visual Evaluation ADDE-M - Manipulation: 1
Manipulation in Fabric attribute, + Rib



Risultati

- Confronto visuale MANN e ADDE-M con 3 manipolazioni

Visual Evaluation ADDE-M - Manipulation: 2
Manipulation in Pattern attribute, - Striped + Floral



Risultati

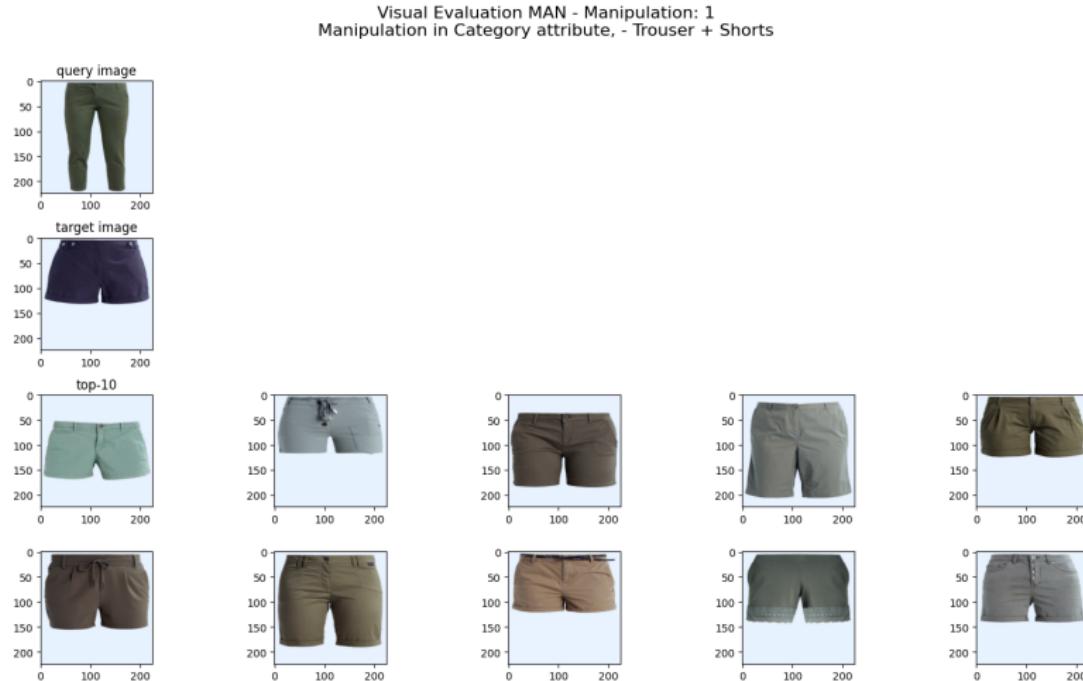
- Confronto visuale MANN e ADDE-M con 3 manipolazioni

Visual Evaluation ADDE-M - Manipulation: 3
Manipulation in Color attribute, - White + Pink



Risultati

- Confronto visuale MANN e ADDE-M con 4 manipolazioni



Risultati

- Confronto visuale MANN e ADDE-M con 4 manipolazioni

Visual Evaluation MAN - Manipulation: 2
Manipulation in Fit attribute, - Slim + Regular



Risultati

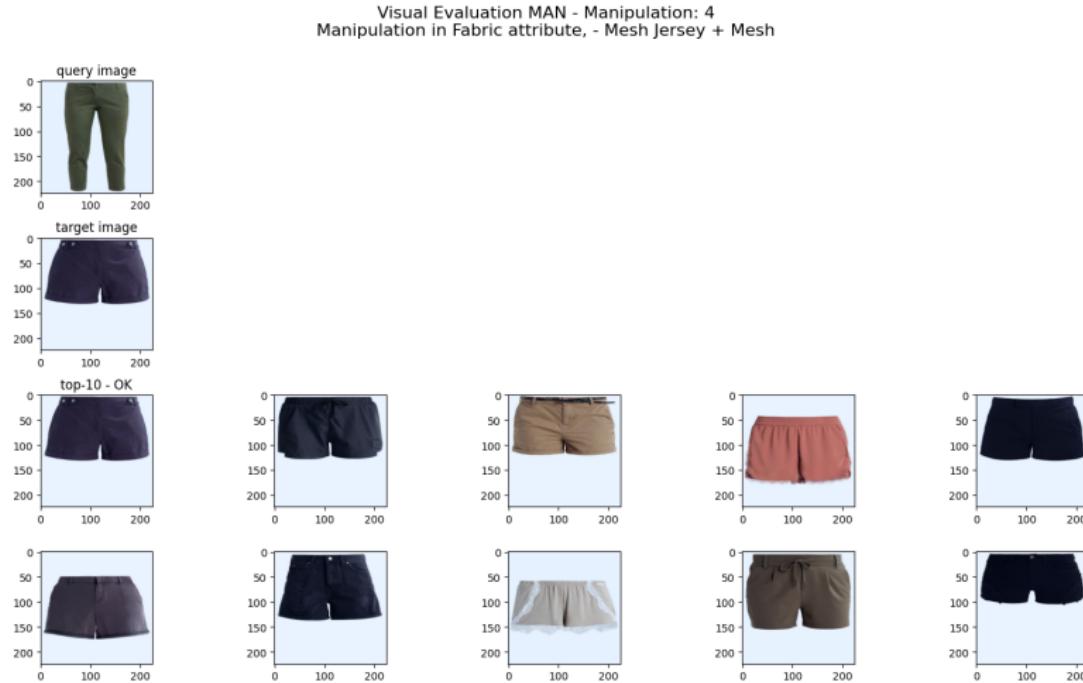
- Confronto visuale MANN e ADDE-M con 4 manipolazioni

Visual Evaluation MAN - Manipulation: 3
Manipulation in Color attribute, + Charcoal



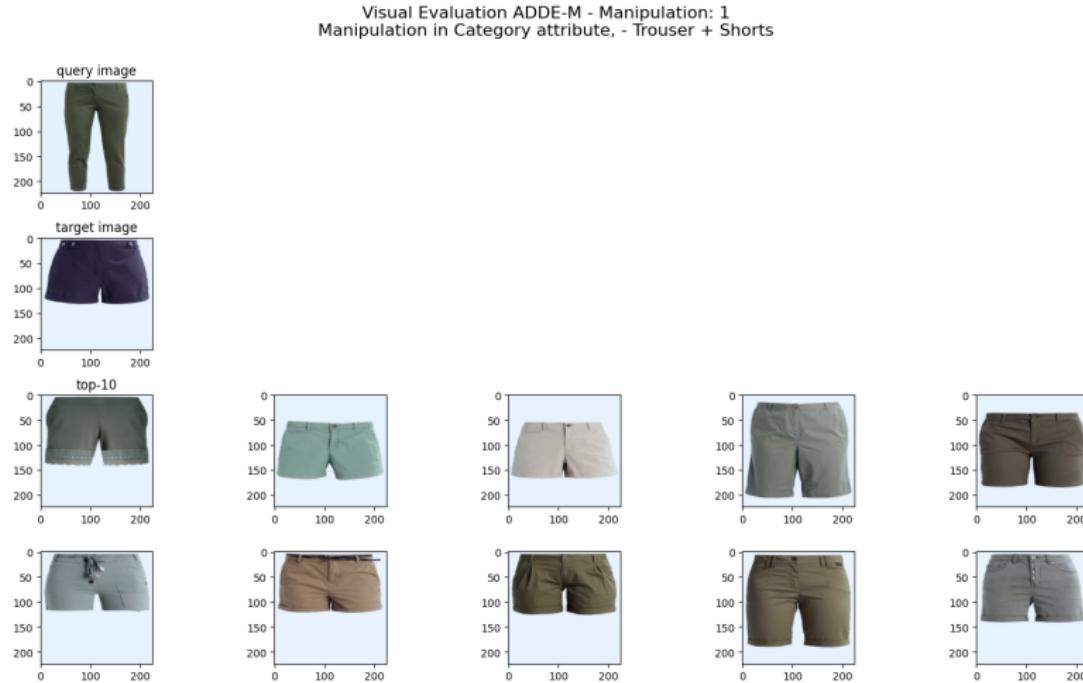
Risultati

- Confronto visuale MANN e ADDE-M con 4 manipolazioni



Risultati

- Confronto visuale MANN e ADDE-M con 4 manipolazioni



Risultati

- Confronto visuale MANN e ADDE-M con 4 manipolazioni

Visual Evaluation ADDE-M - Manipulation: 2
Manipulation in Fit attribute, - Slim + Regular



Risultati

- Confronto visuale MANN e ADDE-M con 4 manipolazioni

Visual Evaluation ADDE-M - Manipulation: 3
Manipulation in Color attribute, + Charcoal



Risultati

- Confronto visuale MANN e ADDE-M con 4 manipolazioni

Visual Evaluation ADDE-M - Manipulation: 4
Manipulation in Fabric attribute, - Mesh Jersey + Mesh



Conclusioni

- Memory Augmented Neural Network per task di manipolazione su immagini di vestiti
- Semplicità del modello \Rightarrow Performance inferiore su una singola manipolazione
- Memoria esterna indipendente e indirizzabile \Rightarrow Performance maggiore su più manipolazioni in sequenza

