Attention-based Fusion for Outfit Recommendation

Katrien Laenen, Marie-Francine Moens

University: KU Leuven

Motivation

- Explosive growth of e-commerce content on the Web
- Recommendation systems are essential to overcome consumer overchoice
- Limited support for users looking for a full outfit

Problem definition

Outfit compatibility prediction

Outfit completion

Challenges

- **Item understanding**
  - Capture important fine-grained product features in the item representation
  - Effectively fuse the information in the product image and description

  → attention-based fusion
- **Item matching**
  - Compatibility is a complex relationship (e.g., not transitive)

appliqued embellished fit flare dress

pre-owned christian louboutin confusalta T-strap platform peep toe pum

Overview

- Motivation
- Problem definition
- Challenges
- Methodology
  - Baseline model: Common space fusion
  - Our model: Attention-based fusion
- Datasets
- Experiments
- Results
- Conclusions

Methodology: Common space fusion

Baseline model

Common space fusion method of Vasileva et al. (2018)

Input

A triplet of image embeddings $(\boldsymbol{x}_{(u)}, \boldsymbol{x}^{+}_{(v)}, \boldsymbol{x}^{-}_{(v)})$ and a triplet of corresponding description embeddings $(\boldsymbol{t}_{(u)}, \boldsymbol{t}^{+}_{(v)}, \boldsymbol{t}^{-}_{(v)})$

<u>Example:</u> u = Dresses v = Shoes

$(\boldsymbol{x}_{(u)}, \boldsymbol{x}^{+}_{(v)}, \boldsymbol{x}^{-}_{(v)})$ = ( ,  , )

$(\boldsymbol{t}_{(u)}, \boldsymbol{t}^{+}_{(v)}, \boldsymbol{t}^{-}_{(v)})$ = (appliqued embellished fit flare dress , pre-owned christian louboutin confusalta T-strap platform peep toe pum , solid color lace-up high top)

Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David A. Forsyth. 2018. Learning Type-Aware Embeddings for Fashion Compatibility. In *ECCV*.

Methodology: Common space fusion

Multimodal semantic space

visual-semantic loss

applied embellished fit flare dress

solid color lace-up high top

$$\mathcal{L}_{vse} = \frac{\mathcal{L}_{vse,\boldsymbol{x}_{(u)}} + \mathcal{L}_{vse,\boldsymbol{x}^{+}_{(v)}} + \mathcal{L}_{vse,\boldsymbol{x}^{-}_{(v)}}}{3}$$

$$\mathcal{L}_{vse,\boldsymbol{x}_{(u)}} = \frac{\ell(W_i\boldsymbol{x}_{(u)}, W_s\boldsymbol{t}_{(u)}, W_s\boldsymbol{t}^{+}_{(v)}) + \ell(W_i\boldsymbol{x}_{(u)}, W_s\boldsymbol{t}_{(u)}, W_s\boldsymbol{t}^{-}_{(v)})}{2}$$

with $\ell(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \max(0, f(\boldsymbol{x},\boldsymbol{z}) - f(\boldsymbol{x},\boldsymbol{y}) + m)$

and $f(\boldsymbol{x},\boldsymbol{y}) = \dfrac{\boldsymbol{x}^T \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$

visual similarity loss

$$\mathcal{L}_{vsim} = \frac{\ell(W_i\boldsymbol{x}^{+}_{(v)}, W_i\boldsymbol{x}^{-}_{(v)}, W_i\boldsymbol{x}_{(u)}) + \ell(W_i\boldsymbol{x}^{-}_{(v)}, W_i\boldsymbol{x}^{+}_{(v)}, W_i\boldsymbol{x}_{(u)})}{2}$$

textual similarity loss

solid color lace-up high top

pre-owned christian louboutin confusalta T-strap platform peep toe pum

applied embellished fit flare dress

$$\mathcal{L}_{tsim} = \frac{\ell(W_s\boldsymbol{t}^{+}_{(v)}, W_s\boldsymbol{t}^{-}_{(v)}, W_s\boldsymbol{t}_{(u)}) + \ell(W_s\boldsymbol{t}^{-}_{(v)}, W_s\boldsymbol{t}^{+}_{(v)}, W_s\boldsymbol{t}_{(u)})}{2}$$

Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David A. Forsyth. 2018. Learning Type-Aware Embeddings for Fashion Compatibility. In *ECCV*.

Methodology: Common space fusion

Type-specific compatibility spaces

compatibility loss

$$\mathcal{L}_{comp} = \ell(W_c^{(u,v)} W_i \boldsymbol{x}_{(u)}, W_c^{(u,v)} W_i \boldsymbol{x}^+_{(v)}, W_c^{(u,v)} W_i \boldsymbol{x}^-_{(v)})$$

$$\text{with } \ell(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \max(0, f(\boldsymbol{x}, \boldsymbol{z}) - f(\boldsymbol{x}, \boldsymbol{y}) + m)$$

$$\text{and } f(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x}^T \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$$

Compatibility space for *Dresses* and *Shoes*:

Training

complete loss

$$\mathcal{L} = \mathcal{L}_{comp} + \lambda_1 \mathcal{L}_{vsim} + \lambda_2 \mathcal{L}_{tsim} + \lambda_3 \mathcal{L}_{vse}$$

Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David A. Forsyth. 2018. Learning Type-Aware Embeddings for Fashion Compatibility. In *ECCV*.

Methodology: Attention-based fusion

Input

A triplet of **region-level** image features $(\boldsymbol{x}_{1:N(u)}, \boldsymbol{x}^{+}_{1:N(v)}, \boldsymbol{x}^{-}_{1:N(v)})$ and a triplet of corresponding description-level features $(\boldsymbol{t}_{(u)}, \boldsymbol{t}^{+}_{(v)}, \boldsymbol{t}^{-}_{(v)})$ ***or*** word-level features $(\boldsymbol{t}_{1:M(u)}, \boldsymbol{t}^{+}_{1:M(v)}, \boldsymbol{t}^{-}_{1:M(v)})$ *(depends on attention mechanism)*

Multimodal semantic space

Average region-level and word-level representations, i.e., $\boldsymbol{x}_{(u)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_{i(u)}$, to compute losses

Use **attention** to fuse the visual and textual information to obtain a triplet of multimodal item representations $(\boldsymbol{m}_{(u)}, \boldsymbol{m}^{+}_{(v)}, \boldsymbol{m}^{-}_{(v)})$

Type-specific compatibility spaces

$$\mathcal{L}_{comp} = \ell(W_c^{(u,v)} \overset{\boldsymbol{m}_{(u)}}{\cancel{W_i \boldsymbol{x}_{(u)}}}, W_c^{(u,v)} \overset{\boldsymbol{m}^{+}_{(v)}}{\cancel{W_i \boldsymbol{x}^{+}_{(v)}}}, W_c^{(u,v)} \overset{\boldsymbol{m}^{-}_{(v)}}{\cancel{W_i \boldsymbol{x}^{-}_{(v)}}})$$

Methodology: Visual dot product attention

Input

Region-level image features $X \in \mathbb{R}^{N \times d_g}$

Description-level text feature $\boldsymbol{t} \in \mathbb{R}^{d_g}$

Visual attention weights and context vector

$$a_i = \tanh(\boldsymbol{x_i}) \cdot \tanh(\boldsymbol{t})$$

$$\boldsymbol{c} = \sum_{i=1}^{N} \alpha_i \boldsymbol{x_i}, \text{ with } \alpha_i = \text{softmax}([a_1, a_2, \ldots, a_N])_i$$

Multimodal item representation

$$[\boldsymbol{c}; \boldsymbol{t}]$$

Methodology: Stacked visual attention

Input

Region-level image features $X \in \mathbb{R}^{N \times d_g}$

Description-level text feature $\boldsymbol{t} \in \mathbb{R}^{d_g}$

Visual attention weights and context vector

Computed in R attention hops

$$\boldsymbol{a}^{(r)} = \boldsymbol{w}_p^{(r)} \tanh(W_v^{(r)} X^T \oplus (W_t^{(r)} \boldsymbol{q}^{(r-1)} + \boldsymbol{b}_s^{(r)}))$$

$$\boldsymbol{c}^{(r)} = \boldsymbol{\alpha}^{(r)} X, \text{ with } \boldsymbol{\alpha}^{(r)} = \mathrm{softmax}(\boldsymbol{a}^{(r)})$$

$$\boldsymbol{q}^{(r)} = \boldsymbol{q}^{(r-1)} + \boldsymbol{c}^{(r)}$$

Multimodal item representation

$$[\boldsymbol{q}^{(R)}; \boldsymbol{t}]$$

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *CVPR*. IEEE Computer Society, 21–29.

Methodology: Co-attention

Input

Region-level image features $X \in \mathbb{R}^{N \times d_g}$

Word-level text features $Y \in \mathbb{R}^{M \times d_g}$

Textual attention

$$\boldsymbol{a}^t = \underset{in=d_g, out=1, k=1}{\text{Convolution1D}_{t,2}}(\text{ReLU}(\underset{in=d_g, out=d_g, k=1}{\text{Convolution1D}_{t,1}}(Y)))$$

$$\boldsymbol{c}^t = \boldsymbol{\alpha}^t Y, \text{ with } \boldsymbol{\alpha}^t = \text{softmax}(\boldsymbol{a}^t)$$

$$M = MFB(X, \boldsymbol{c}^t)$$

Visual attention

$$\boldsymbol{a}^{v,(r)} = \underset{in=d_g, out=1, k=1}{\text{Convolution1D}_{v,2}^{(r)}}(\text{ReLU}(\underset{in=2d_g, out=d_g, k=1}{\text{Convolution1D}_{v,1}^{(r)}}(M)))$$

$$\boldsymbol{c}^{v,(r)} = \boldsymbol{\alpha}^{v,(r)} M, \text{ with } \boldsymbol{\alpha}^{v,(r)} = \text{softmax}(\boldsymbol{a}^{v,(r)})$$

$$\boldsymbol{c}^v = W_f[\boldsymbol{c}^{v,(1)}; \boldsymbol{c}^{v,(2)}; \ldots; \boldsymbol{c}^{v,(R)}]$$

Multimodal item representation

$$MFB(\boldsymbol{c}^v, \boldsymbol{c}^t)$$

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal Factorized Bilinear Pooling with Co-Attention Learning for Visual Question Answering. In *ICCV*, 1839–1848.

Datasets

Polyvore68K-ND

- 68,306 outfits
  (78% training, 7% validation, 15% testing)
- 365,054 items

Polyvore68K-D

- 35,140 outfits
  (48% training, 9% validation, 43% testing)
- 175,485 items

Polyvore21K

- 20,925 outfits
  (81% training, 6% validation, 13% testing)

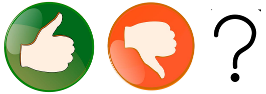
	Item Types
Polyvore68K	Accessories, All body, Bags, Bottoms, Hats, Jewellery, Outerwear, Scarves, Shoes, Sun-glasses, Tops
Polyvore21K	Accessories, Activewear, Baby, Bags and Wallets, Belts, Boys, Cardigans and Vests, Clothing, Cos-tumes, Cover-ups, Dresses, Eyewear, Girls, Gloves, Hats, Hosiery and Socks, Jeans, Jewellery, Jumpsuits, Juniors, Kids, Maternity, Outerwear, Pants, Scarves, Shoes, Shorts, Skirts, Sleepwear, Suits, Sweaters and Hoodies, Swimwear, Ties, Tops, Underwear, Watches, Wedding Dresses

Table 2: Item types kept in the Polyvore68K and Polyvore21K datasets.

Experimental setup

Experiments and evaluation

- *Fashion compatibility (FC) task:* Given a set of items, compute the outfit compatibility score as the average compatibility score across all item pairs in the set



- *Fill-in-the-blank (FITB) task:* Given an incomplete set of items and 4 candidate items, find the most compatible candidate item as the one which has the highest total compatibility score with the items in the set

Training details

- Output of the 7x7x256 res4b_relu layer of ResNet18 to represent images
- Bidirectional LSTM to represent descriptions and words

Results

	Polyvore68K-ND		Polyvore68K-D		Polyvore21K	
	FC	FITB	FC	FITB	FC	FITB
Common space fusion						
baseline [11]	85.62	56.55	85.07	56.91	86.28	58.35
Attention-based fusion						
visual dot product attention	89.43	61.55	86.85	60.12	88.59	**63.11**
stacked visual attention	**89.68**	**61.92**	**87.25**	**60.48**	**88.89**	62.52
co-attention	89.58	61.20	86.25	59.00	85.04	58.20

Table 1: Results on the fashion compatibility and fill-in-the-blank tasks for the Polyvore68K dataset versions and the Polyvore21K dataset.

KU LEUVEN

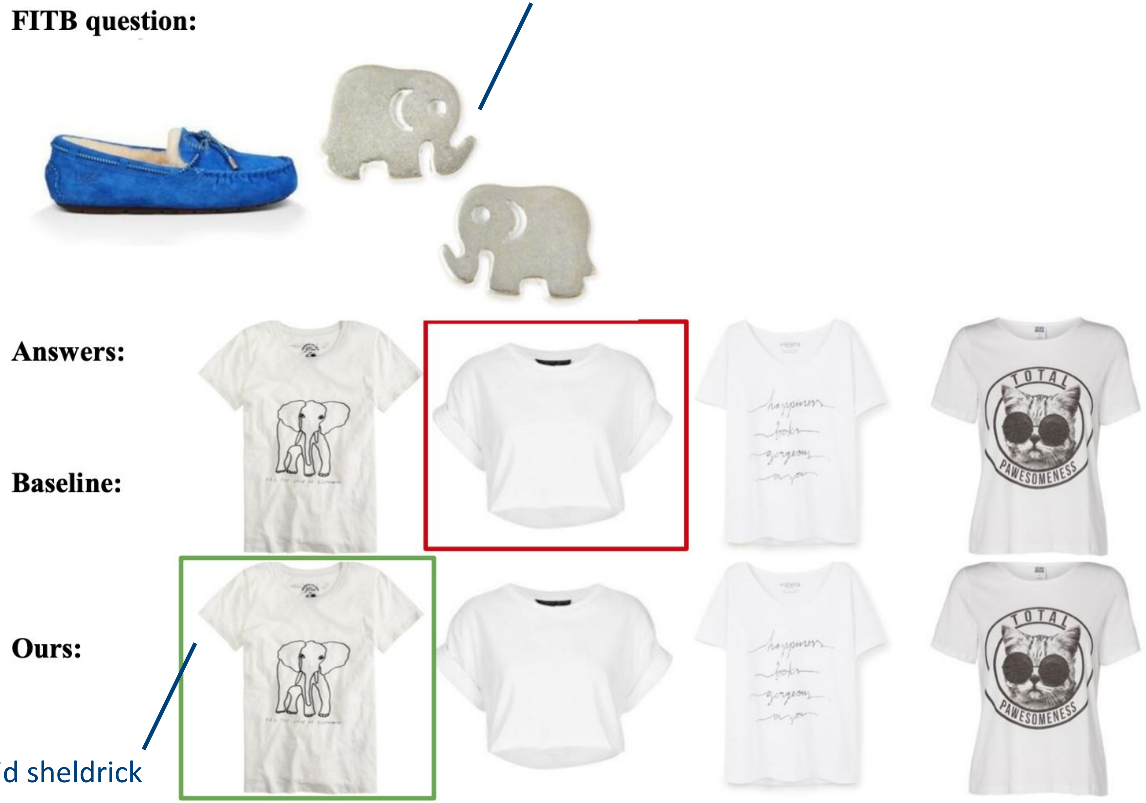
Results

Results

Results

Results

dogeared sterling silver **elephant** stud



women's j.crew for david sheldrick wildlife trust **elephant** t-shirt

KU LEUVEN

Conclusions and future work

- Attention on region-level image features and word-level text features allows to bring certain product features to the forefront in the multimodal item representations, which benefits the outfit recommendation task
- Improve state-of-the-art results on an outfit compatibility prediction task and an outfit completion task on three datasets
- Investigate neural architectures that still better recognise fine-grained fashion attributes in images
- Design novel co-attention mechanisms


1425