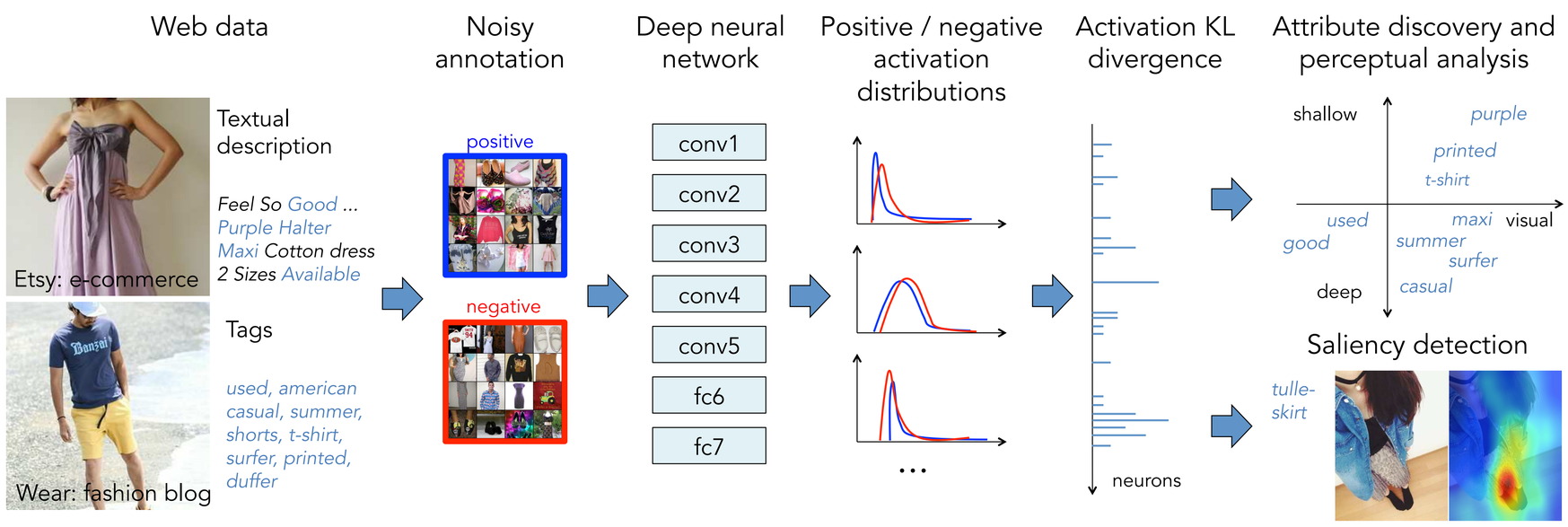
Automatic Attribute Discovery with Neural Activations

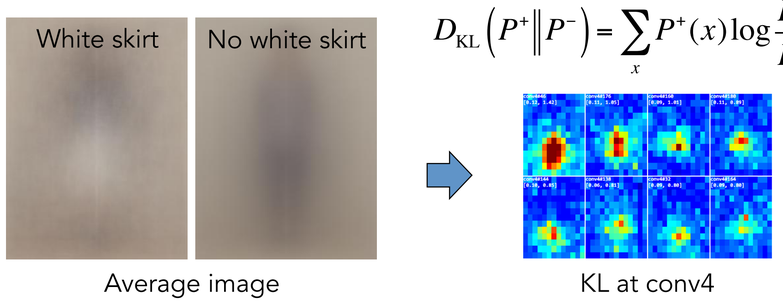
Sirion Vittayakorn[1], Takayuki Umeda[2], Kazuhiko Murasaki[2], Kyoko Sudo[2], Takayuki Okatani[3], Kota Yamaguchi[3]
[1]UNC at Chapel Hill, NC, USA [2]NTT Media Intelligence Laboratories, Japan [3]Tohoku University, Japan
http://cs.unc.edu/~sirionv

Overview

- Can we learn visual attributes without a supervised dataset?
- Our approach: look at neurons
- KL divergence helps identify visual attributes from noisy label association
- Relation to human perception
- Neurons to saliency detection



KL divergence identifies prime units



$$D_{\mathrm{KL}}\left(P^{+}\middle\|P^{-}\right)=\sum_{x}P^{+}(x)\log\frac{P^{+}(x)}{P^{-}(x)}$$

- Some neurons in the pre-trained network respond to visual attributes, even if they are not explicitly supervised
- KL-divergence on activation histogram can identify such neurons (prime units)

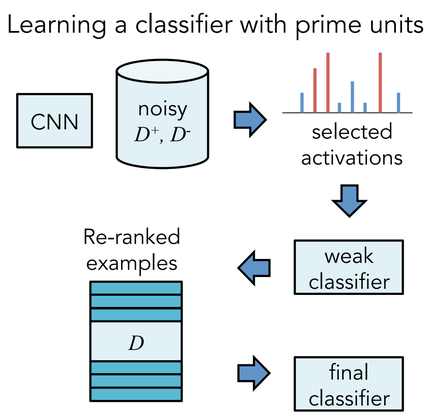
Automatically discovered attributes

Method	Most visual	Least visual
Human	flip pink red floral blue sleeve purple little black yellow	url due last right additional sure free old possible cold
Pre-trained +resample	flip pink red yellow green purple floral blue sexy elegant	big great due much own favorite new free different good
Attribute-tuned	flip sexy green floral yellow pink red purple lace loose	right same own light happy best small different favorite free
Language-prior	top sleeve front matching waist bottom lace dry own right	organic lightweight classic gentle adjustable floral adorable url elastic super

Prime-units perceive attributes like humans do

- Learning a classifier on top of prime units shows close proximity to humans
- Evaluating visualness to human-agreement correlation

$$\mathrm{Visualness}\left(u\,|\,\mathrm{classifier}\right)=\mathrm{accuracy}(\mathrm{classifier},D_u^{+},D_u^{-})$$



Correlation to human perception

Method	Pearson	Spearman
Pre-trained + random (baseline)	0.737	0.637
Pre-trained + resample	**0.799**	0.717
Attribute-tuned	0.662	0.549
Attribute-tuned + random	0.716	0.565
Attribute-tuned + resample	0.782	**0.721**
Category-tuned + random	0.760	0.684
Category-tuned + resample	0.783	0.704
Language prior	0.139	0.032

Layers characterize attribute perception

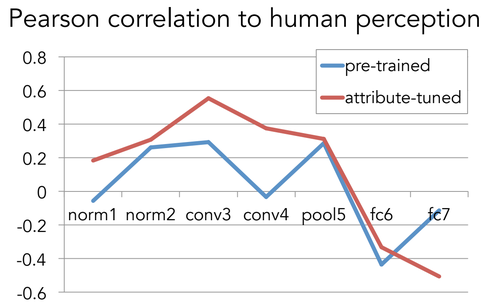
- From early primitive attributes to deeper abstract attributes
- Human agreement highest in the middle?
- Fine-tuning affects maximum KL ratio per layer

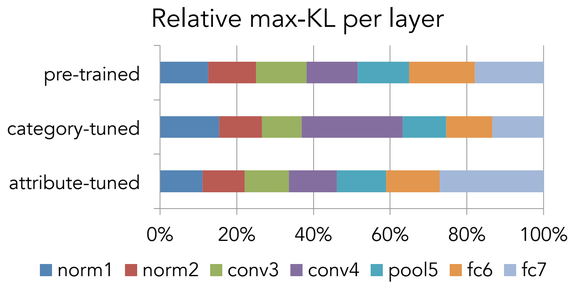
Etsy dataset

norm1	norm2	conv3	conv4	pool5	fc6	fc7
orange colorful vibrant bright blue welcome exact yellow red specific	green red yellow purple colorful blue vibrant ruffle orange only	bright pink red purple green lace yellow sweet french black	flattering lovely vintage romantic deep waist front gentle formal delicate	lovely elegant natural beautiful delicate recycled chic formal decorative romantic	many soft new upper sole genuine friendly sexy stretchy great	sleeve sole acrylic cold flip newborn large floral waist american

Wear dataset

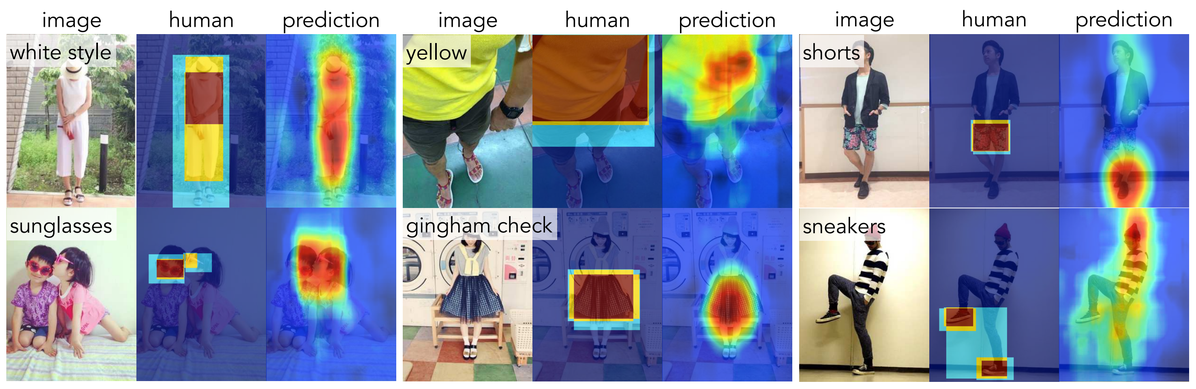
norm1	norm2	conv3	conv4	pool5	fc6	fc7
blue green red-black red denim-on-denim denim-shirt pink denim yellow leopard	denim-jacket pink red red-socks red-black champion blue white shirt i-am-clumsy yellow	border-striped-tops border-stripes dark-style stripes backpack red dark-n-dark denim-shirt navy outdoor-style	kids bucket-hat hat-n-glasses black sleeveless american-casual long-cardigan white-n-white stole mom-style	shorts half-length pants denim dotted border-stripes white-pants border-striped-tops gingham-check sandals chester-coat	white-skirt flared-skirt spring upper beret shirt-dress overalls hair-band loincloth style matched-pair	long-skirt suit-style midi-skirt gaucho-pants handmade straw-hat white-n-white white-coordinate white-pants white

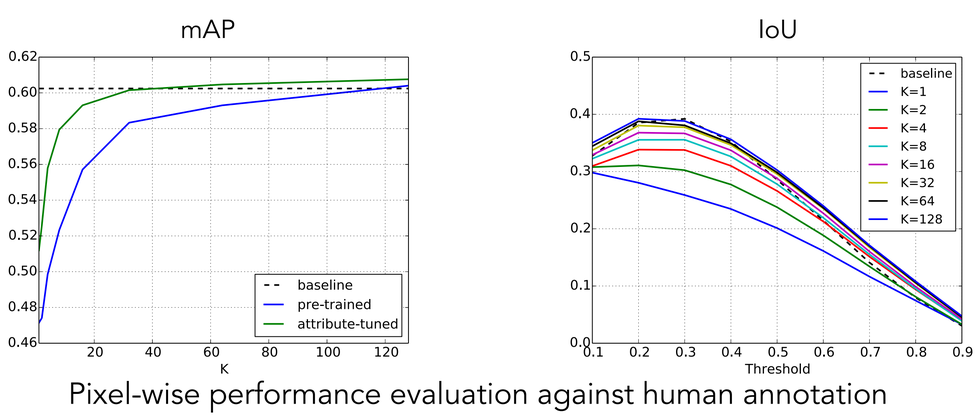




Prime units can identify salient region

- Accumulating sliding mask [Zhou 14] by largest KL units
- Baseline: smoothed gradients [Simonyan 2014]





Pixel-wise performance evaluation against human annotation