Deep Learning the City
Quantifying Urban Perception at a Global Scale

Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, Cesar A Hidalgo

naik@mit.edu

Key Contributions

- **Place Pulse 2.0 Dataset**: Contains 1.17 million pairwise comparisons for 110,988 images from 56 cities, provided by 81,630 online volunteers
- **Six perceptual attributes:** Safe, Lively, Boring, Wealthy, Depressing, and Beautiful
- Deep network that minimizes a **joint classification + ranking loss** to accurately predict perception of urban attributes

The Place Pulse 2.0 Dataset

Figure 1: User Interface for Crowdsourced Online Game

- Goal: **Quantify the perception of urban environments**
- Helps study the relationship between a city's physical appearance and the behavior and health of its residents
- A global dataset of human judgments in the form of pairwise comparisons of urban appearance
- Siamese-like networks, Streetscore-CNN (SS-CNN) and Ranking SS-CNN, to predict pairwise comparisons

Predicting Human Judgments

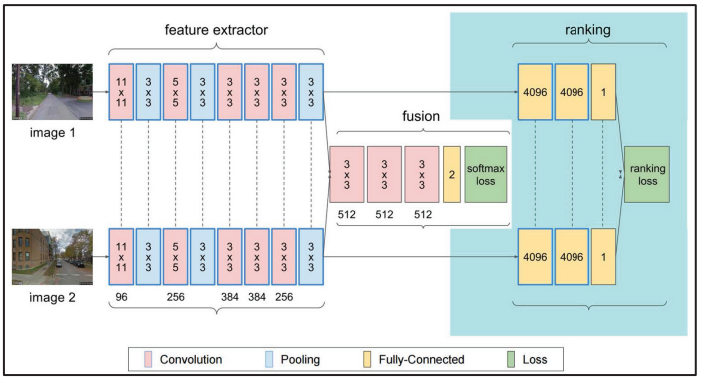

Figure 2: **SS-CNN** and **RSS-CNN** (with additional layers in light blue)

Our dataset contains a set of $\boldsymbol{m}$ images $I = \{x_i\}_{i=1}^{m}$ and set of $\boldsymbol{N}$ comparison triplets $\boldsymbol{P} = \{(i_k, j_k, y_k)\}_{k=1}^{N}$, $i, j \in \{1, \ldots, m\}$, $y \in \{+1, -1\}$. We want to learn a ranking function $f_r(x)$ that satisfies the maximum # of constraints [1, 3]

$$y \cdot (f_r(x_i) - f_r(x_j)) > 0 \quad \forall\ (i, j, y) \in P$$

To solve this problem, we construct networks (1) SS-CNN that minimizes a classification loss L_c and (2) RSS-CNN that minimizes a ranking loss L_r in addition to L_c

$$L_c = \sum_{(i,j,y)\in P} \sum_{k}^{K} -\mathbb{1}[y = k] \log(g_k(x_i, x_j)) \qquad L_r = \sum_{(i,j,y)\in P} \left(\max(0, y \cdot (f_r(x_j) - f_r(x_i))\right)^2$$

k= 2, g → softmax of final layer

RSS-CNN is trained with a joint loss $L = L_c + \lambda L_r$, where λ is set using grid-search to maximize classification accuracy

Performance Analysis

- **SS-CNN:** We calculate the % of pairwise comparisons in test set predicted correctly by
  - **(1) Softmax** of output neurons in final layer
  - **(2)** comparing **TrueSkill** scores [2] obtained from *synthetic* pairwise comparisons from the CNN
  - **(3)** extracting features from penultimate layer of CNN and feeding pairwise feature representations to a **RankSVM** [3]
- **RSS-CNN**: We compare the ranking function outputs for both images in a test pair to decide which image wins, and calculate the binary prediction accuracy.

(a) SS-CNN

Network	Ranking Method		
	Softmax	TrueSkill	RankSVM
AlexNet	53.0%	55.7%	58.4%
SS-CNN (AlexNet)	60.3%	62.6%	65.5%
PlacesNet	56.4%	58.8%	61.6%
SS-CNN (PlacesNet)	62.2%	64.7%	68.1%
VGGNet	60.9%	62.7%	63.5%
SS-CNN (VGGNet)	65.3%	67.8%	**72.4%**

(b) RSS-CNN

Model	Prediction Acc.
AlexNet	64.1%
PlacesNet	68.8%
VGGNet	**73.5%**

Table 1 : Pairwise comparison prediction accuracy for standard networks fine-tuned with the Place Pulse 2.0 dataset. RSS-CNN (VGGNet) obtains the best performance

Prediction Performance Across Attributes

Train \ Test	Safe	Lively	Beautiful	Wealthy	Boring	Depressing
Safe	**73.5%**	67.7%	66.3%	60.3%	47.2%	42.3%
Lively	63.8%	**70.3%**	65.8%	61.3%	58.9%	53.7%
Beautiful	61.2%	67.1%	**70.2%**	53.5%	50.2%	51.4%
Wealthy	60.7%	54.6%	52.7%	**65.7%**	52.8%	55.9%
Boring	48.6%	55.6%	52.3%	53.1%	**66.1%**	59.8%
Depressing	54.5%	54.2%	43.2%	49.7%	57.2%	**62.8%**

Table 2 : Prediction performance of a network trained on one attribute is transferable to similar attributes. Prediction performance is correlated with the number of pairwise comparisons used for training

Visualizations

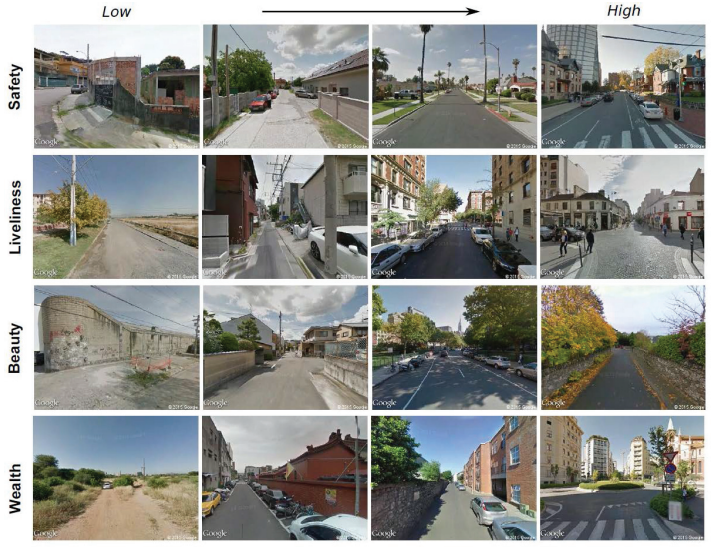

Figure 3: Example results from Place Pulse 2.0 dataset.

References

[1] Parikh, D., Grauman, K.: Relative attributes. ICCV (2011) 503–510
[2]Herbrich, R., Minka, T., Graepel, T.: TrueSkill: A Bayesian skill rating system. Advances in Neural Information Processing Systems (2006) 569–576
[3] Joachims, T.: Optimizing search engines using clickthrough data. KDD (2002) 133–142
[4] Naik, N., Philipoom, J., Raskar, R., Hidalgo, C.: Streetscore–Predicting the perceived safety of one million streetscapes. CVPR Workshops (2014) 793–799