What looks good with my sofa: Multimodal Search Engine for Interior Design

Ivona Tautkute, Aleksandra Mozejko, Tomasz Trzcinski, Krzysztof Marasek, Wojciech Stokowiec, Lukasz Brocki

Polish - Japanese Academy of Information Technology,
Warsaw University of Technology,
Tooploox

Presentation plan

1. What is style search?
2. Dataset description
3. Model pipeline
4. Multimodal approaches
5. Results

Problem

Find items that match **not only** visually but also **by style.**

Extend visual query by **text input**.

Visual search → CBIR → Style Search

Dataset challenges

1. Item (product) images
2. Context (room) quality images (e.g designer magazines)
3. One-to-many relationship between items (product) and context (room).
4. Text descriptions for item and context images

Our dataset

- 298 room photos
- 2193 product photos
- 6 product categories

Room images:

Object images:

Description:

You sit comfortably thanks to the armrests.

There's a natural and living feeling of wood, as knots and other marks remain on the surface.

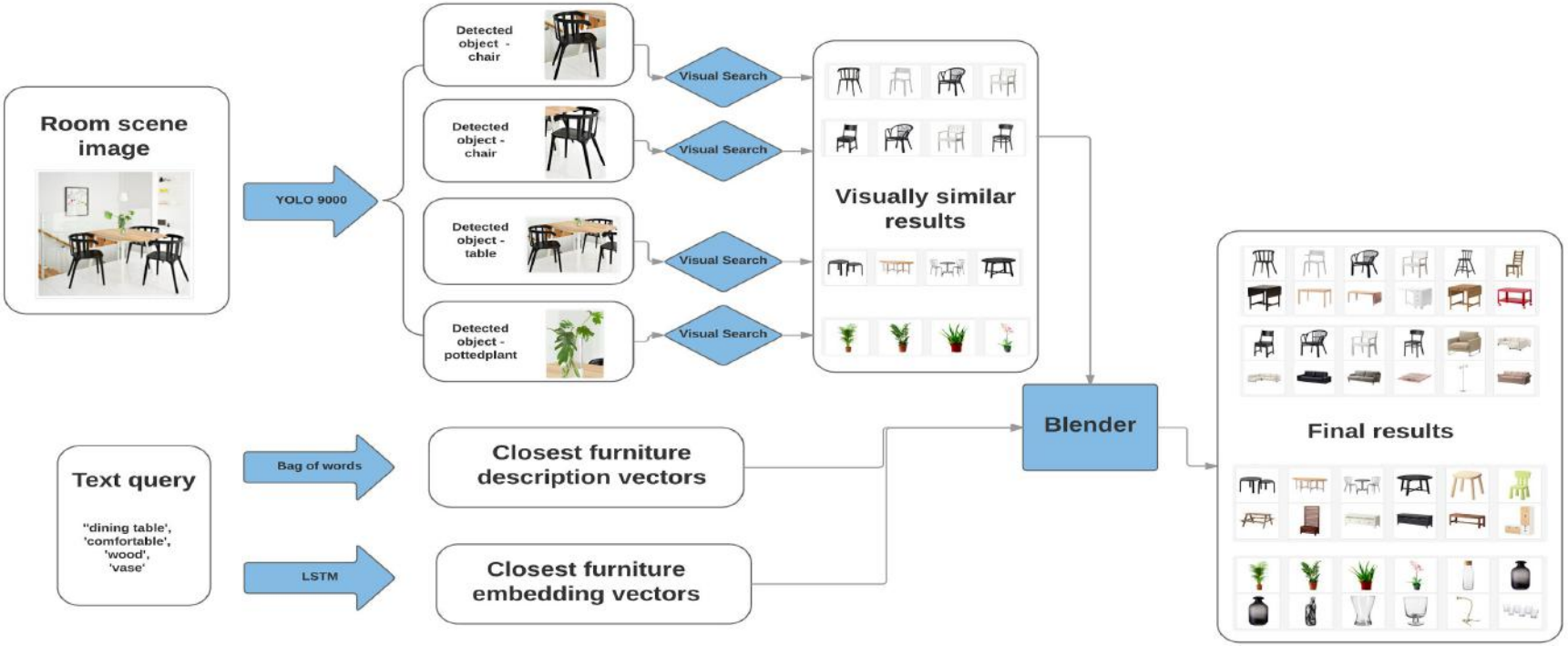
This lamp gives a pleasant light for dining and spreads a good directed light across your dining or bar table.

Extendable, so it can be pulled out as your child grows.

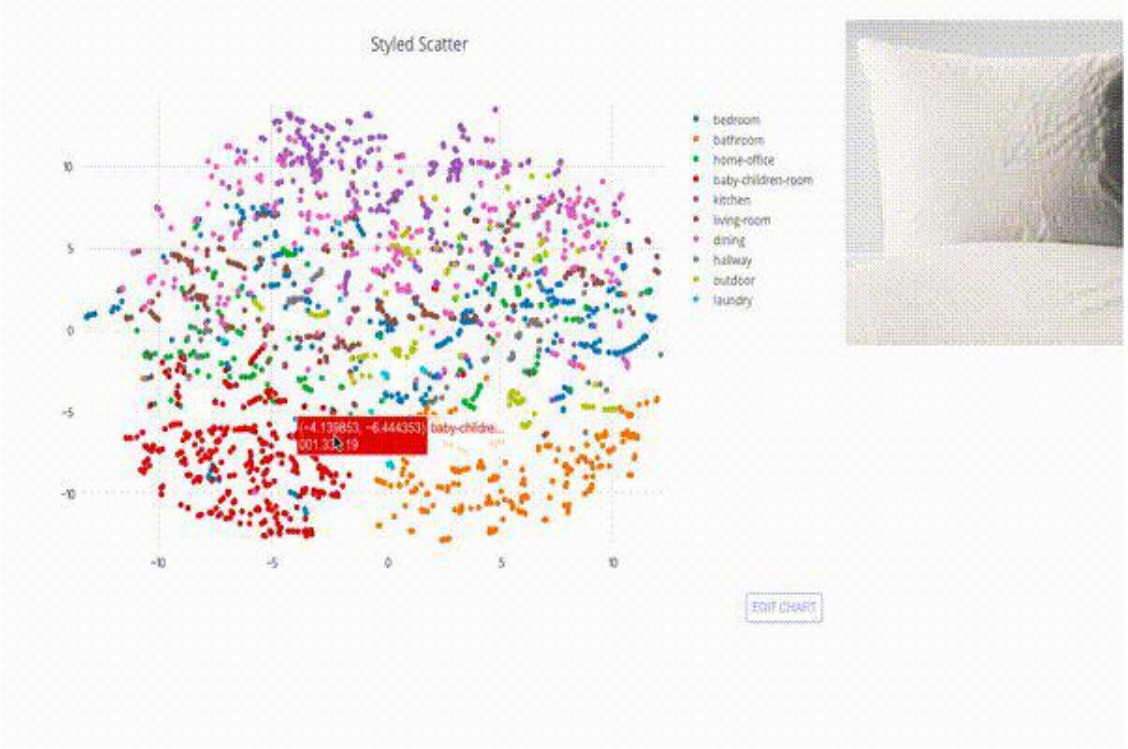
You can adjust the height of the clothes rail and shelves as your child grows.

Made of plastic which makes it easy to carry and move for children.

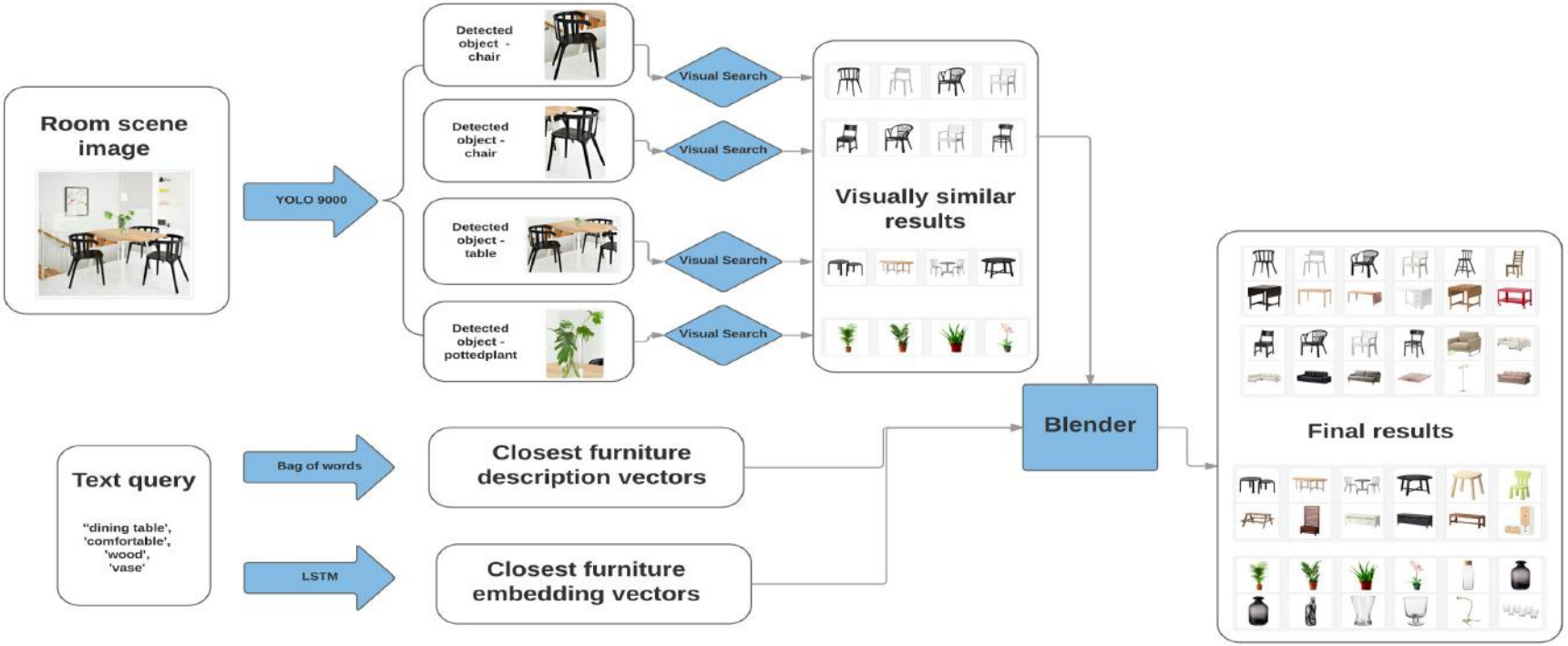
Model pipeline



Furniture products embedding



Model pipeline



Methods

- YOLO 9000 (Darknet) [1]
- Convolutional Neural Networks (VGG, Resnet)
- CBOW (word2vec) [2]

Type	Filters	Size/Stride	Output
Convolutional	32	3 × 3	224 × 224
Maxpool		2 × 2/2	112 × 112
Convolutional	64	3 × 3	112 × 112
Maxpool		2 × 2/2	56 × 56
Convolutional	128	3 × 3	56 × 56
Convolutional	64	1 × 1	56 × 56
Convolutional	128	3 × 3	56 × 56
Maxpool		2 × 2/2	28 × 28
Convolutional	256	3 × 3	28 × 28
Convolutional	128	1 × 1	28 × 28
Convolutional	256	3 × 3	28 × 28
Maxpool		2 × 2/2	14 × 14
Convolutional	512	3 × 3	14 × 14
Convolutional	256	1 × 1	14 × 14
Convolutional	512	3 × 3	14 × 14
Convolutional	256	1 × 1	14 × 14
Convolutional	512	3 × 3	14 × 14
Maxpool		2 × 2/2	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	512	1 × 1	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	512	1 × 1	7 × 7
Convolutional	1024	3 × 3	7 × 7
Convolutional	1000	1 × 1	7 × 7
Avgpool		Global	1000
Softmax			

Table 6: Darknet-19.

1. J. Redmon and A. Farhadi, “YOLO 9000: better, faster, stronger,” [25] CoRR, vol. abs/1612.08242, 2016.
2. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” CoRR abs/1301.3781, 2013.

Baselines

Visual Search

- SIFT [1]
- Bag-of-visual-words [2]
- DL architectures (VGG, Resnet)

Results blending

- Simple blending (*k* best results)
- Vanilla text search
- Vanilla visual search

1. D. G. Lowe, “Distinctive image features from scale-invariant key - points,” International Journal of Computer Vision, vol. 60, no. 2, p. 91110, 2004.
2. J. Sivic and A. Zisserman, “Video google: Efficient visual search of videos,” Toward Category-Level Object Recognition Lecture Notes in Computer Science, p. 127144, 2006.

Results

Object detection pre-processing

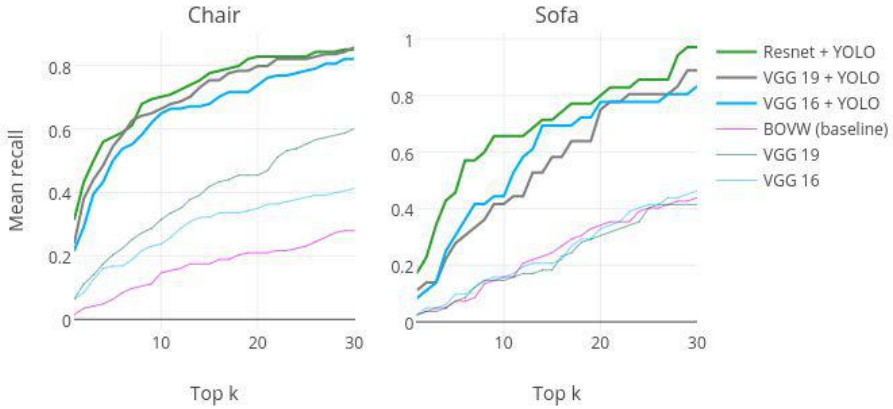


TABLE I

RESULTS FOR CONTENT BASED IMAGE RETRIEVAL EXPERIMENT FOR DIFFERENT MODELS AND ALL OBJECT CLASSES. CONFIGURATION OF RESNET NEURAL NETWORK WITH YOLO 9000 OBJECT DETECTION AS A PRE-PROCESSING STEP SIGNIFICANTLY OUTPERFORMS BOTH THE BASELINE BOVW MODEL AND OTHER DEEP NEURAL NETWORK ARCHITECTURES.

Model	Layer	Hit@6	
		whole image	*with object detection*
BoVW	N/A	0.066	0.26
VGG-16	fc6	0.126	0.392
	fc7	0.153	0.314
VGG-19	fc6	0.141	0.43
	fc7	0.136	0.445
ResNet	avg pool	0.167	**0.458**

$$\frac{1}{|\mathfrak{R}|}\sum_{\mathcal{R}\in|\mathfrak{R}|}\vee_{f\in G_{\mathcal{R}}}\mathbb{I}(rank_{f,\mathcal{R}}\leq k),$$

Hit@k metric S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici,B. Varadarajan, and S. Vijayanarasimhan, “Youtube-8m: A large-scale video classification benchmark,” CoRR, vol. abs/1609.08675, 2016.

Results

Increase in average style similarity score

$$C(f_1, f_2) = |\{\mathcal{R} : f_1 \in \mathcal{R} \wedge f_2 \in \mathcal{R}\}|.$$

$$s(f_1, f_2) = \frac{C(f_1, f_2)}{\max_{f_i, f_j \in \mathcal{F}} C(f_i, f_j)}.$$

11 %

TABLE II

MEAN SIMILARITY RESULTS AVERAGED FOR ALL ROOM PICTURES IN IKEA DATASET AND SAMPLE TEXT QUERIES.

Text query	Visual search	Text search	Simple blending	Feature similarity blending	Joint embedding blending
-	0.2295	-	-	-	-
object class name	-	-	0.2486	0.2374	-
decorative	-	0.1358	0.2316	0.2517	0.2441
black	-	0.1538	0.2493	0.2244	0.1308
white	-	0.2036	0.2958	0.2793	0.2022
smooth	-	0.3520	0.2415	0.3052	0.3142
cosy	-	0.2419	0.2126	0.2334	0.2434
fabric	-	0.0371	0.1269	0.1344	0.2440
colourful	-	0.4461	0.3032	0.3215	0.2423
Average	0.2295	0.2243	0.2387	**0.2484**	0.2315

Results

Query image:

Text query: child
Object class: chair
Object image:

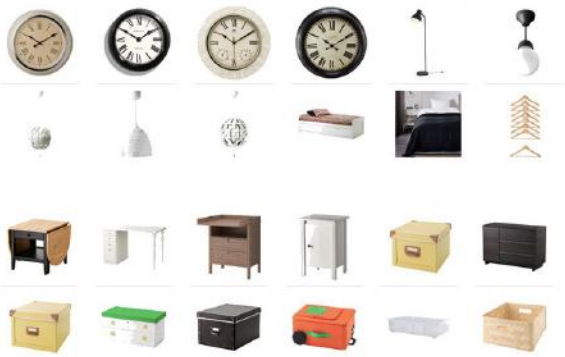
Blended results:

Query image:

Text query: light
Object class: clock
Object image:

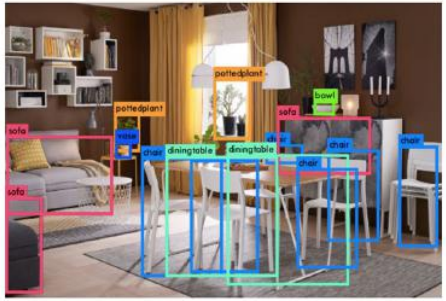
Text query: box
Object class: table
Object image:

Blended results:



Results

Query image:



Text query: white table
Object class: chair
Object image:

Text query: comfortable
Object class: sofa
Object image:

Blended results:

Query image:

Text query: garden
Object class: chair
Object image:

Text query: vase
Object class: diningtable
Object image:

Blended results:

Conclusions and Future work

- **Object detection step** improved content based image retrieval by over 200%.
- By using feature blending approach we increased **overall similarity**.
- We proposed a novel pipeline that tries to tackle difficult topic of **style based retrieval engine.**
- Further joint embedding methods need to be tested.

Thank you!