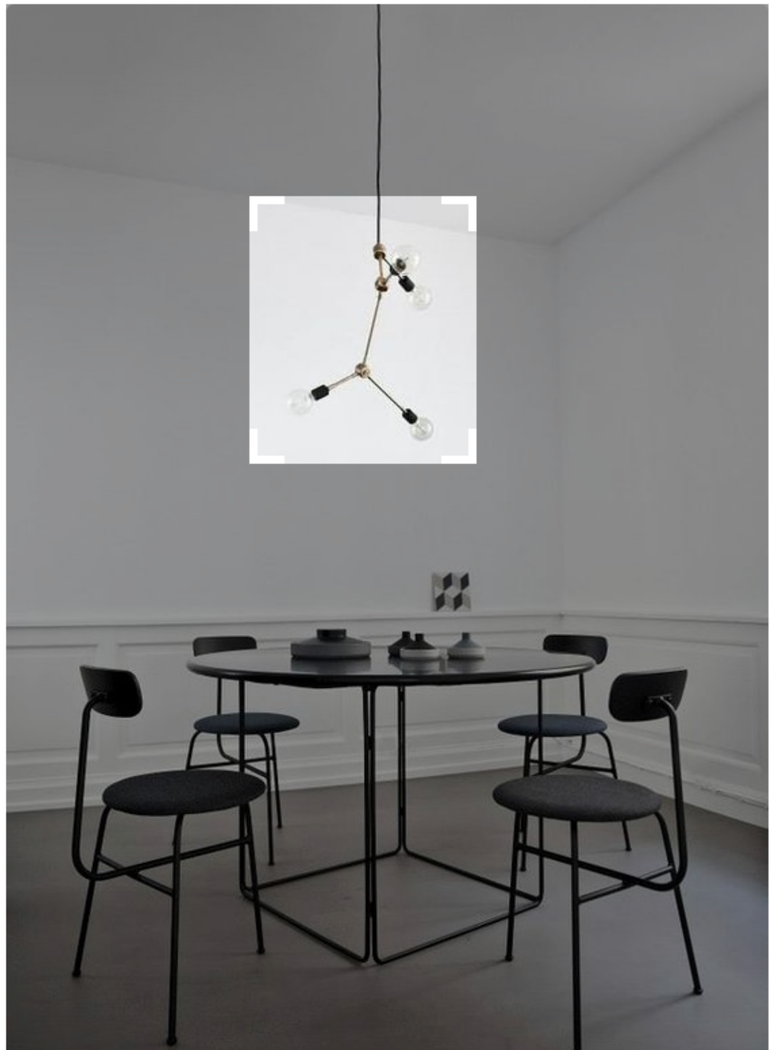
Visual Search in the Deep Learning Era

Dmitry Kislyuk
Visual Search Lead @ Pinterest

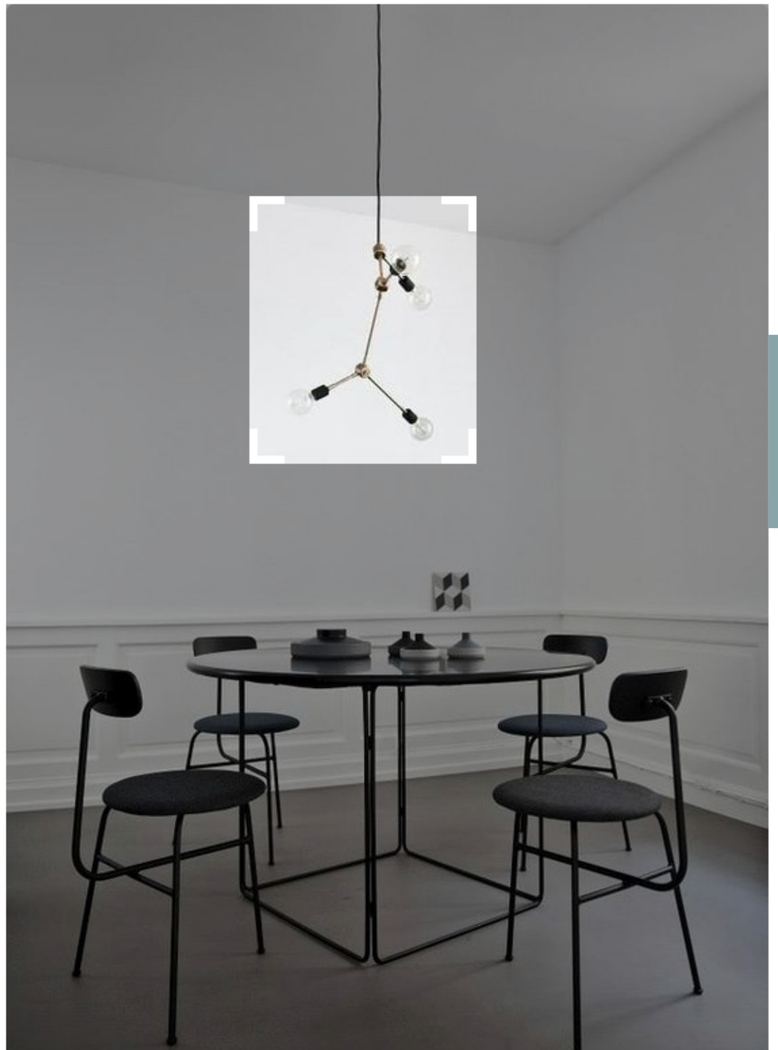
Visual Search

Visually similar results



Visual Search

Visually similar results



franklin chandelier | tribeca franklin | chandeliers | franklin | tribeca

p
New Decor
1

Robin Lane
Lighting

Menu Tribeca Series Franklin Chandelier
Nest.co.uk
Tamsin Paola
Lighting

• from A-TAK DESIGN
A-TAK DESIGN
kasia morstin
lamps

Veronika W
Lighting

WOHNWUNDERBAR
• LAMPEN •
5

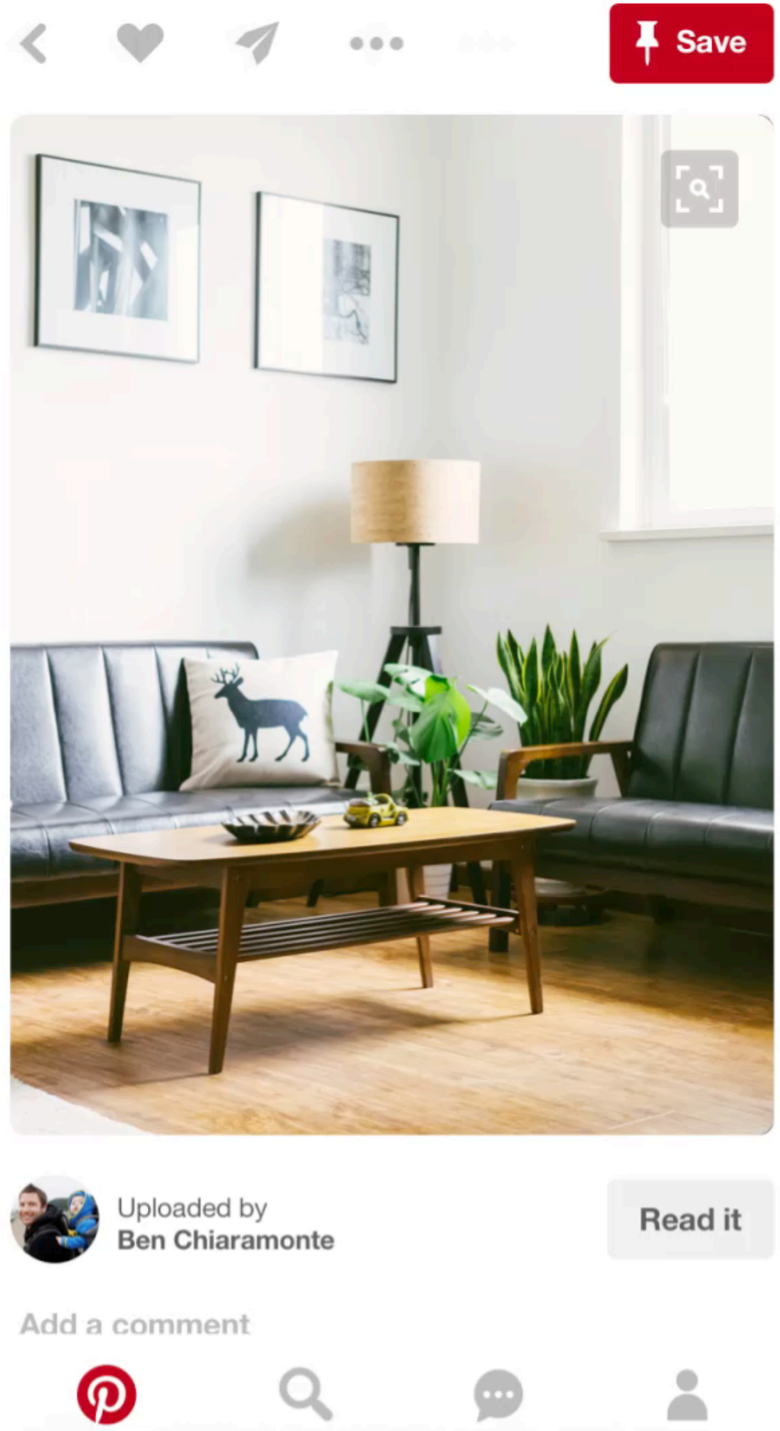
Carla Vent
Home

Karin Daar
Belysning
2

Heidi Risku

Magnificently Modern
Modern & Contemporary ...

Visual Search



Technical Challenge

Define visual similarity between any visual object and images in a dataset, in real time.

Convolutional networks
for image classification

high-dimensional embedding

1 0 1 0 1 1 ⋮ 0 1 1 0 1 0

"table"

Extract high level features | Representation | Classify Each Sample

Convolutional networks
for multi-task learning

high-dimensional embedding

1 0 1 0 1 1 ⋮ 0 1 1 0 1 0

Task 1

Task 2

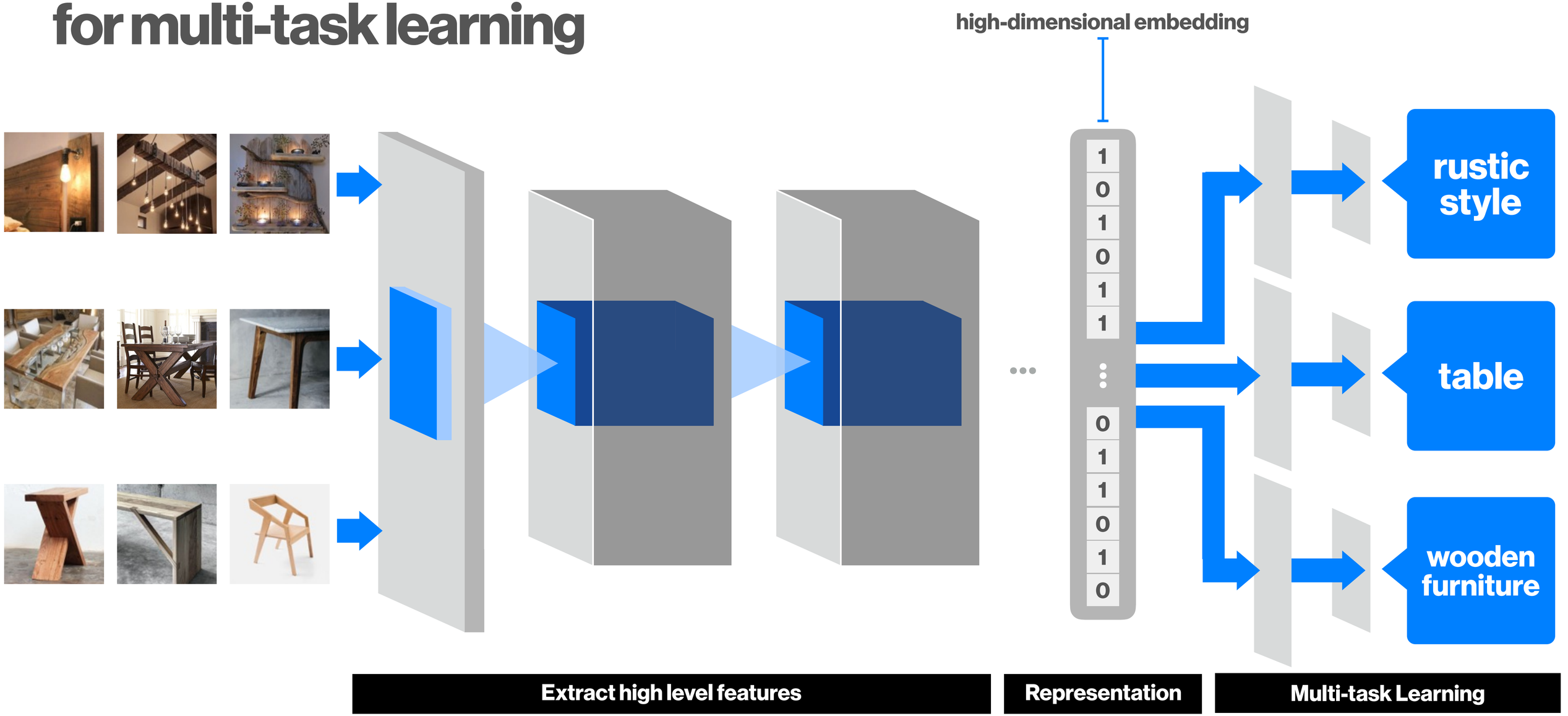
Task 3

Extract high level features

Representation

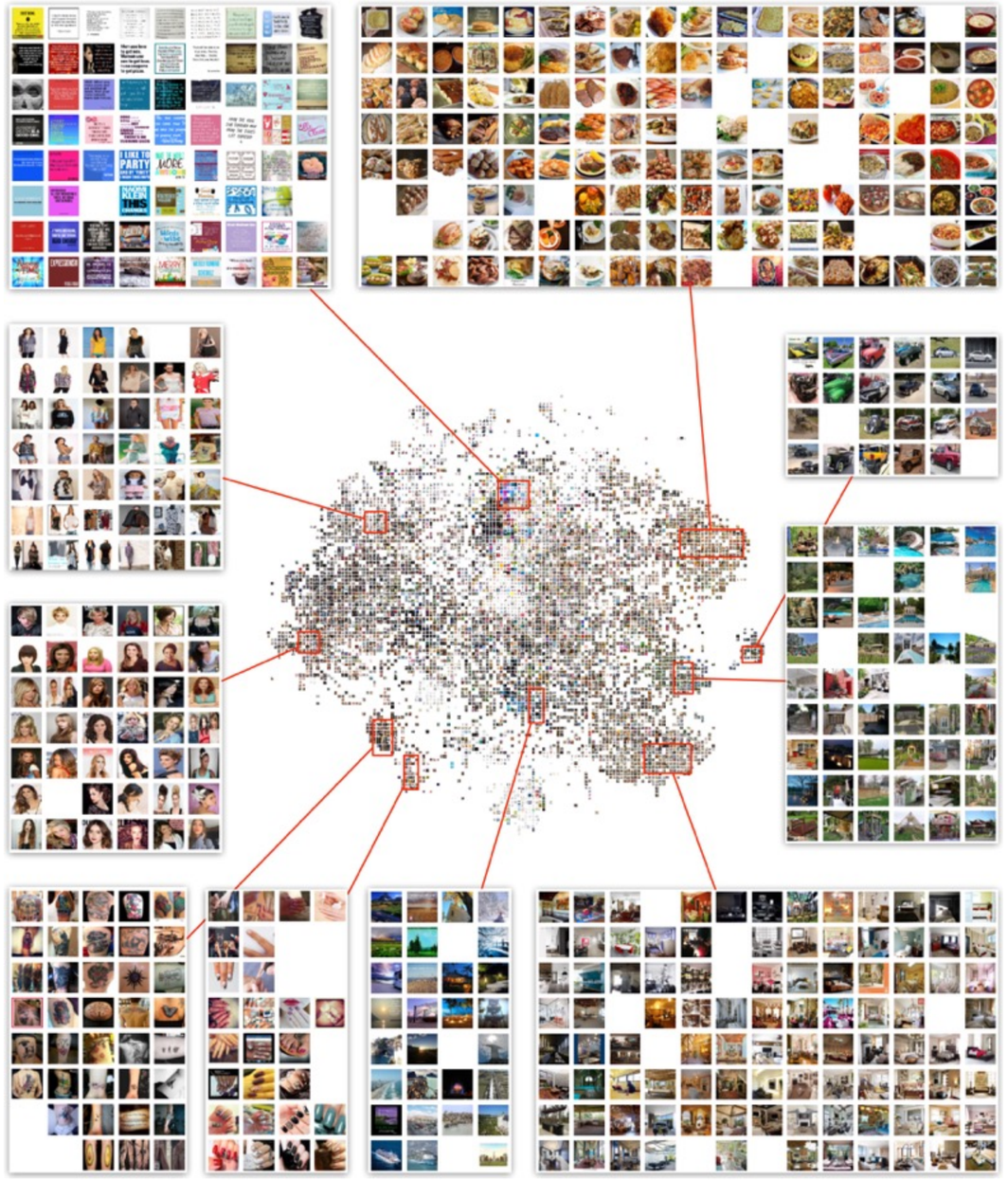
Multi-task Learning

Convolutional networks
for multi-task learning



Deep visual features

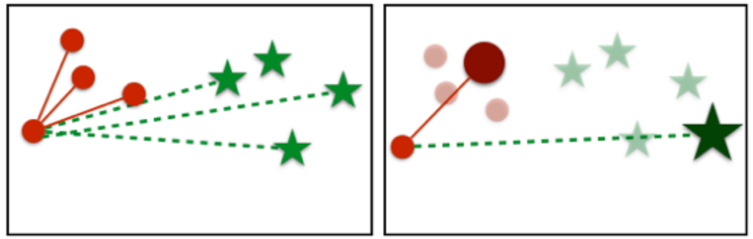
t-SNE of fine-tuned, binarized VGG-16 features



Choice of embedding model

Kislyuk et al., WWW'17

Proxy Ranking Loss (Movshovitz-Attias et al.) + **Multi-Task**



$$L_{\text{triplet}}(x, y, z) = [d(x, y) + M - d(x, z)]_+$$

Other embedding models

ResNet-152 / ResNeXt embeddings lead to best overall p@k performance, but features 4x larger

Binarized **VGG-16 fc6** most compact high-performing embeddings

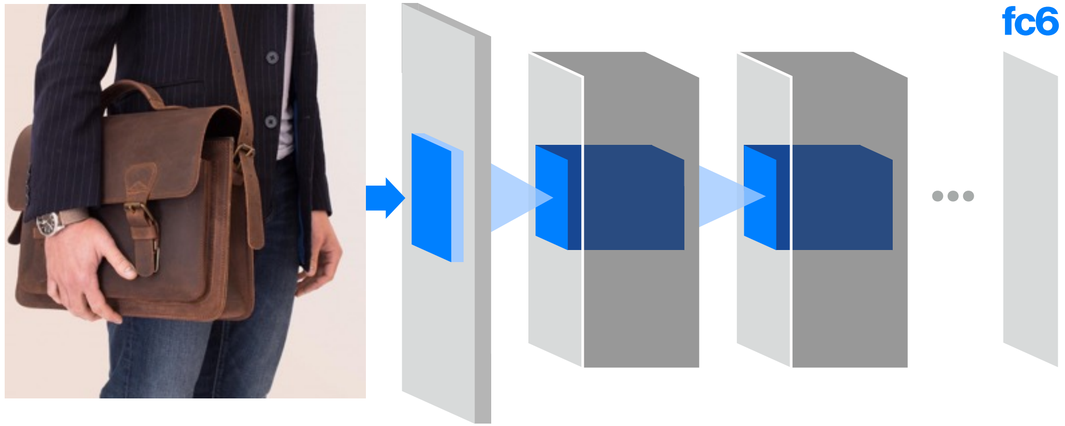
Fine-tuning on Pinterest data helps further

Model	Layer	Type	Dist.	P@1	P@5	P@10
AlexNet	fc6	raw	L2	0.093	0.045	0.027
AlexNet	fc6	raw	L1	0.099	0.047	0.027
AlexNet	fc7	raw	L2	0.104	0.052	0.031
AlexNet	fc7	raw	L1	0.106	0.052	0.031
AlexNet	fc8	raw	L2	0.088	0.047	0.028
AlexNet	fc8	raw	L1	0.090	0.046	0.028
GoogleNet	loss3/classifier	raw	L2	0.095	0.050	0.032
GoogleNet	loss3/classifier	raw	L1	0.098	0.050	0.032
VGG16	fc6	raw	L2	0.108	0.051	0.030
VGG16	fc6	raw	L1	0.118	0.057	0.035
VGG16	fc7	raw	L2	0.116	0.058	0.036
VGG16	fc7	raw	L1	0.113	0.060	0.038
VGG16	fc8	raw	L2	0.104	0.054	0.034
VGG16	fc8	raw	L1	0.106	0.054	0.034
ResNet101	pool5	raw	L2	0.160	0.080	0.050
ResNet101	pool5	raw	L1	0.149	0.073	0.045
ResNet101	fc1000	raw	L2	0.133	0.068	0.042
ResNet101	fc1000	raw	L1	0.139	0.067	0.041
ResNet152	pool5	raw	L2	**0.170**	**0.083**	**0.050**
ResNet152	pool5	raw	L1	0.152	0.077	0.047
ResNet152	fc1000	raw	L2	0.149	0.073	0.045
ResNet152	fc1000	raw	L1	0.148	0.073	0.044
AlexNet	fc6	binary	H.	0.129	0.065	0.039
AlexNet	fc7	binary	H.	0.110	0.054	0.033
AlexNet	fc8	binary	H.	0.089	0.046	0.027
VGG16	fc6	binary	H.	**0.158**	**0.081**	**0.049**
VGG16	fc7	binary	H.	0.133	0.068	0.044
VGG16	fc8	binary	H.	0.110	0.055	0.035
ResNet101	fc1000	binary	H.	0.125	0.062	0.039
ResNet101	pool5	binary	H.	0.055	0.025	0.014
ResNet152	fc1000	binary	H.	0.133	0.065	0.041
ResNet152	pool5	binary	H.	0.057	0.026	0.015
VGG16 (Pin.)	fc6	binary	H.	**0.169**	**0.089**	**0.056**

Determining visual similarity

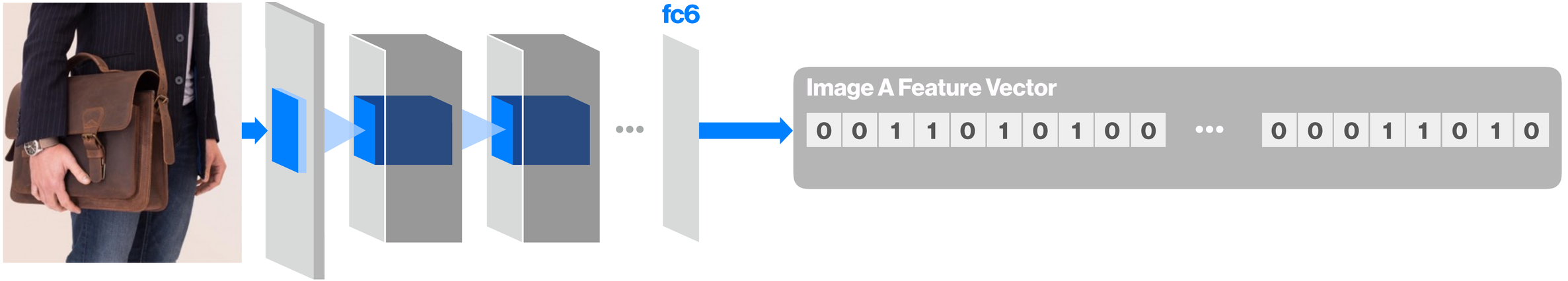
Similarity (A,B)

Determining visual similarity



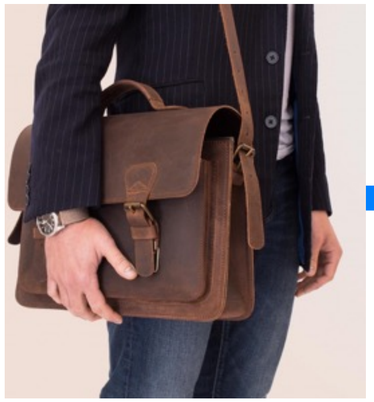
Similarity (A,B)

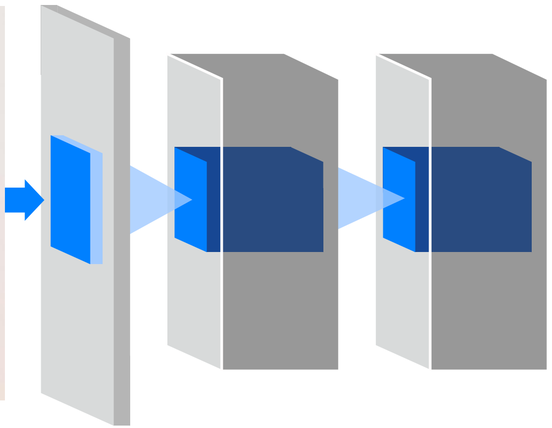
Determining visual similarity



Similarity (A,B)

Determining visual similarity



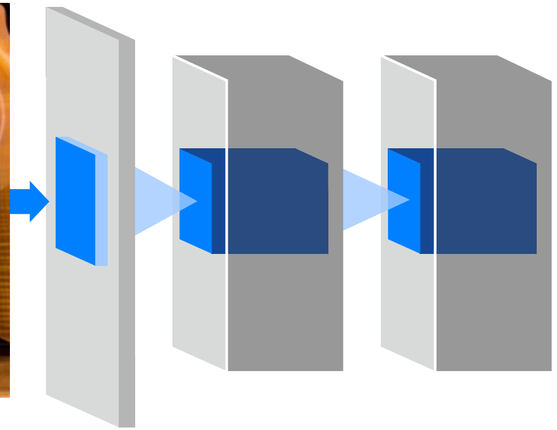


fc6

Image A Feature Vector

0 0 1 1 0 1 0 1 0 0 ••• 0 0 0 1 1 0 1 0

Similarity (A,B)

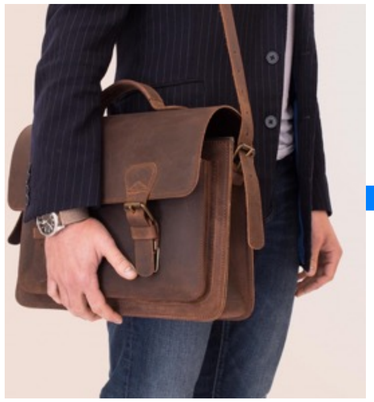


fc6

Image A Feature Vector

1 0 1 0 0 0 0 1 0 1 ••• 1 1 0 0 1 0 0 0

Determining visual similarity



fc6

Image A Feature Vector

0 0 1 1 0 1 0 1 0 0 ••• 0 0 0 1 1 0 1 0

Hamming Distance

Similarity (A,B)

fc6

Image A Feature Vector

1 0 1 0 0 0 0 1 0 1 ••• 1 1 0 0 1 0 0 0

Determining visual similarity

Scaling to billions of images

- Collect training examples from different tasks (predict texture, semantic label, color, etc.)

Rustic Style



Table



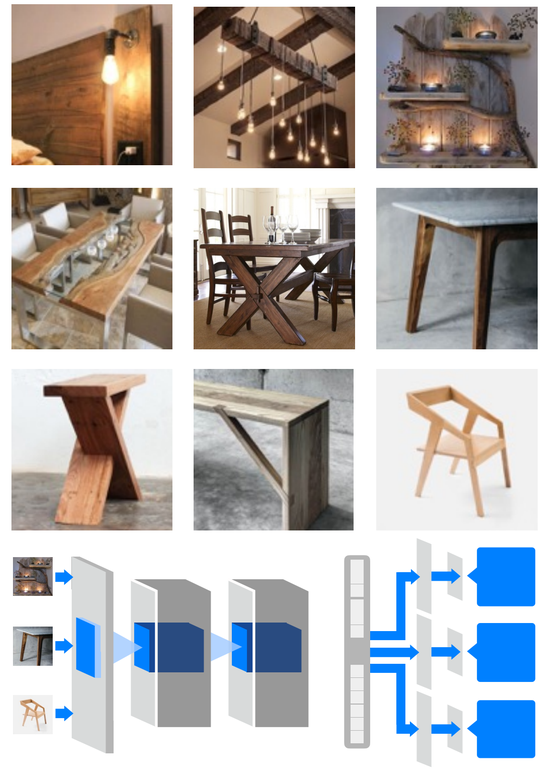
Wooden Furniture



Determining visual similarity

Scaling to billions of images

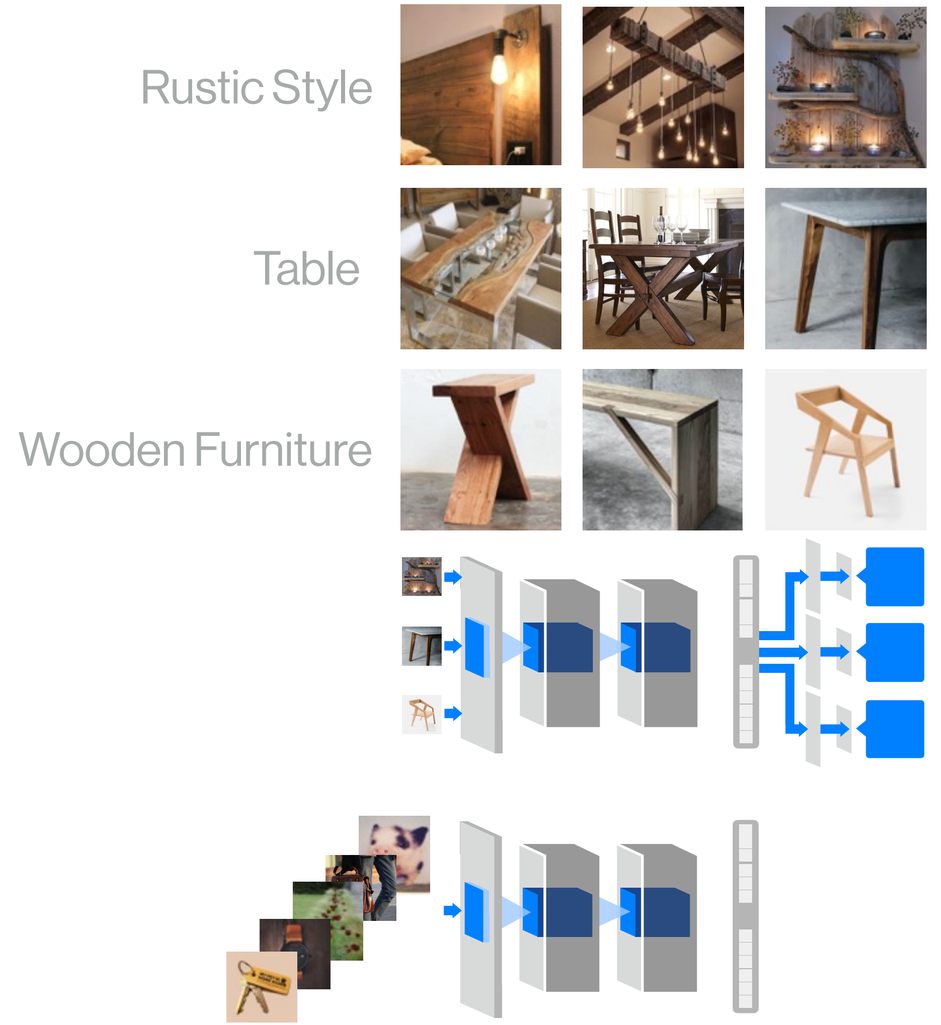
- Collect training examples from different tasks (predict texture, semantic label, color, etc.)
- Train multi-task embedding model on labeled data



Determining visual similarity

Scaling to billions of images

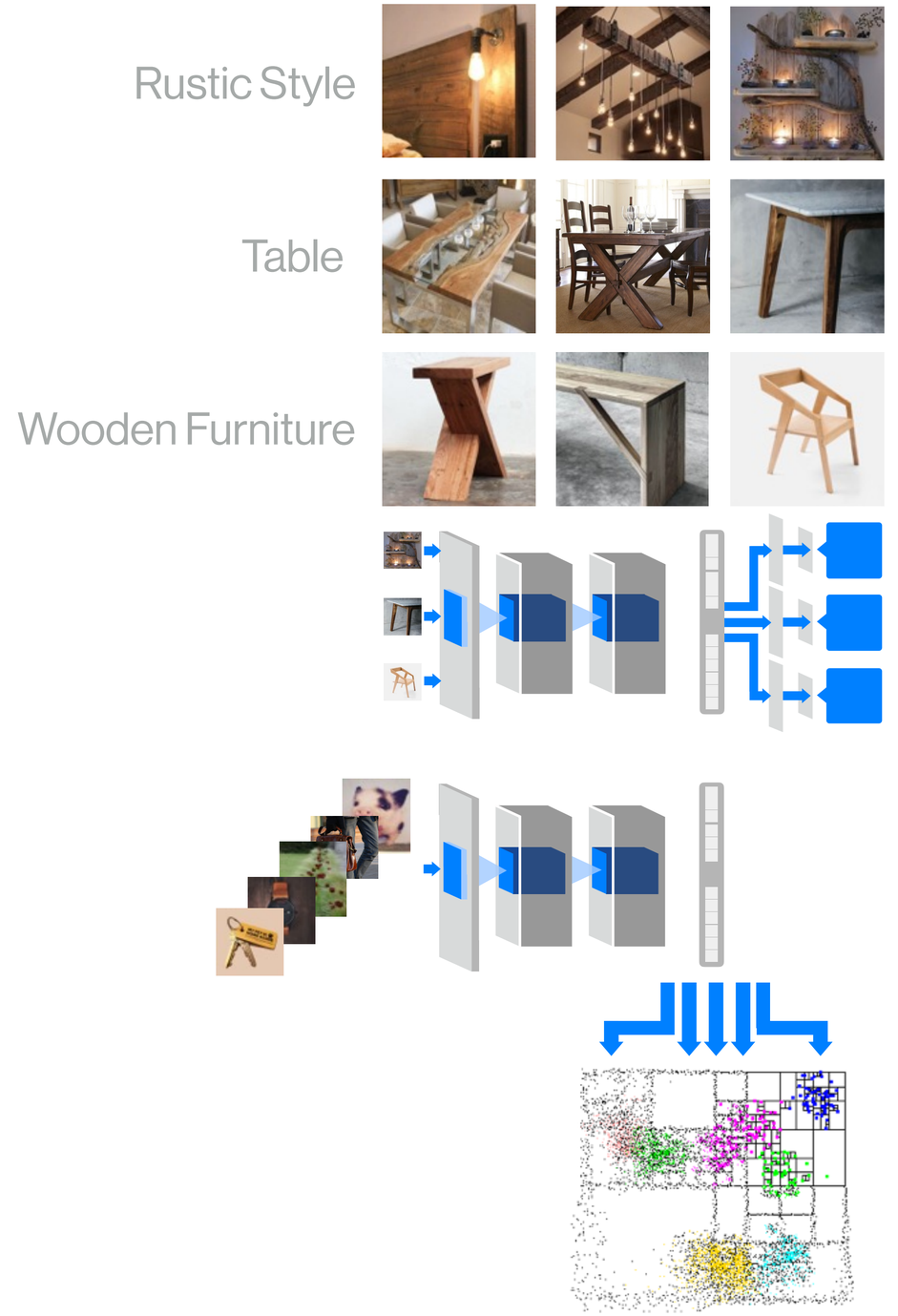
- Collect training examples from different tasks (predict texture, semantic label, color, etc.)
- Train multi-task embedding model on labeled data
- Exact embedding (intermediate convolutional net features) from trained model for the **entire corpus** of images (billions)



Determining visual similarity

Scaling to billions of images

- Collect training examples from different tasks (predict texture, semantic label, color, etc.)
- Train multi-task embedding model on labeled data
- Exact embedding (intermediate convolutional net features) from trained model for the **entire corpus** of images (billions)
- Index image embeddings into Approximate Nearest Neighbor (ANN) data structure (distributed!)



Visual Search on Pinterest

with distributed nearest neighbors

Query

Query

Visual Search Ecosystem

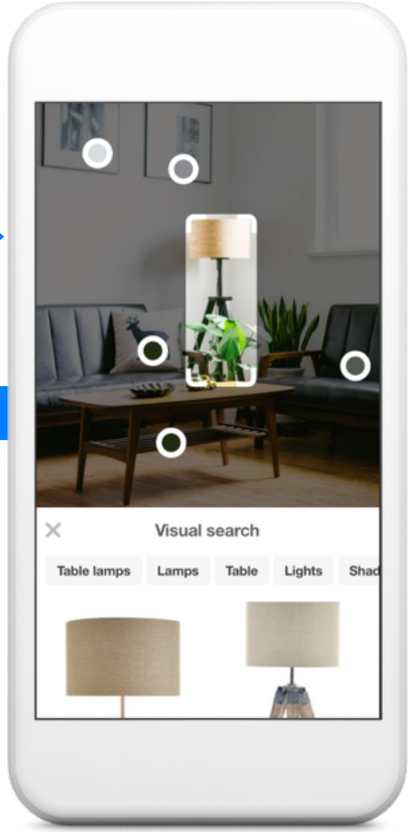
leveraging object detection

Visual search

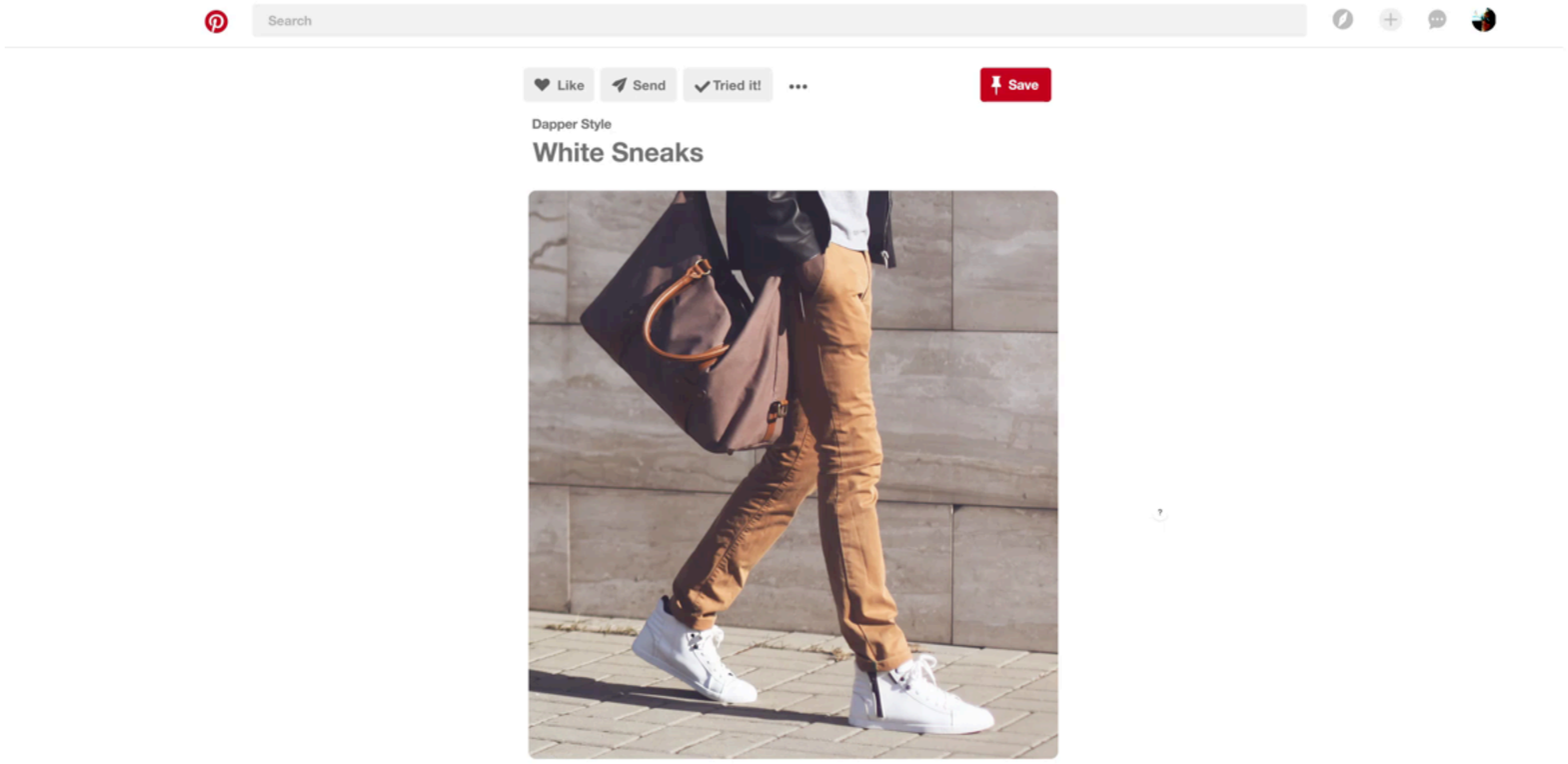
Training data

Improved experience

Object detection

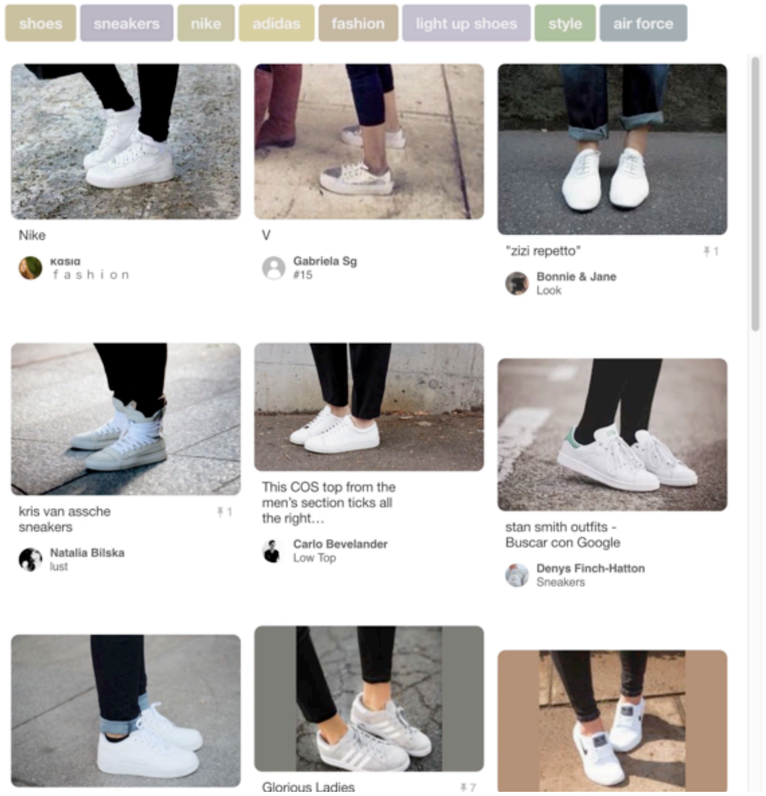


Improving detection with search



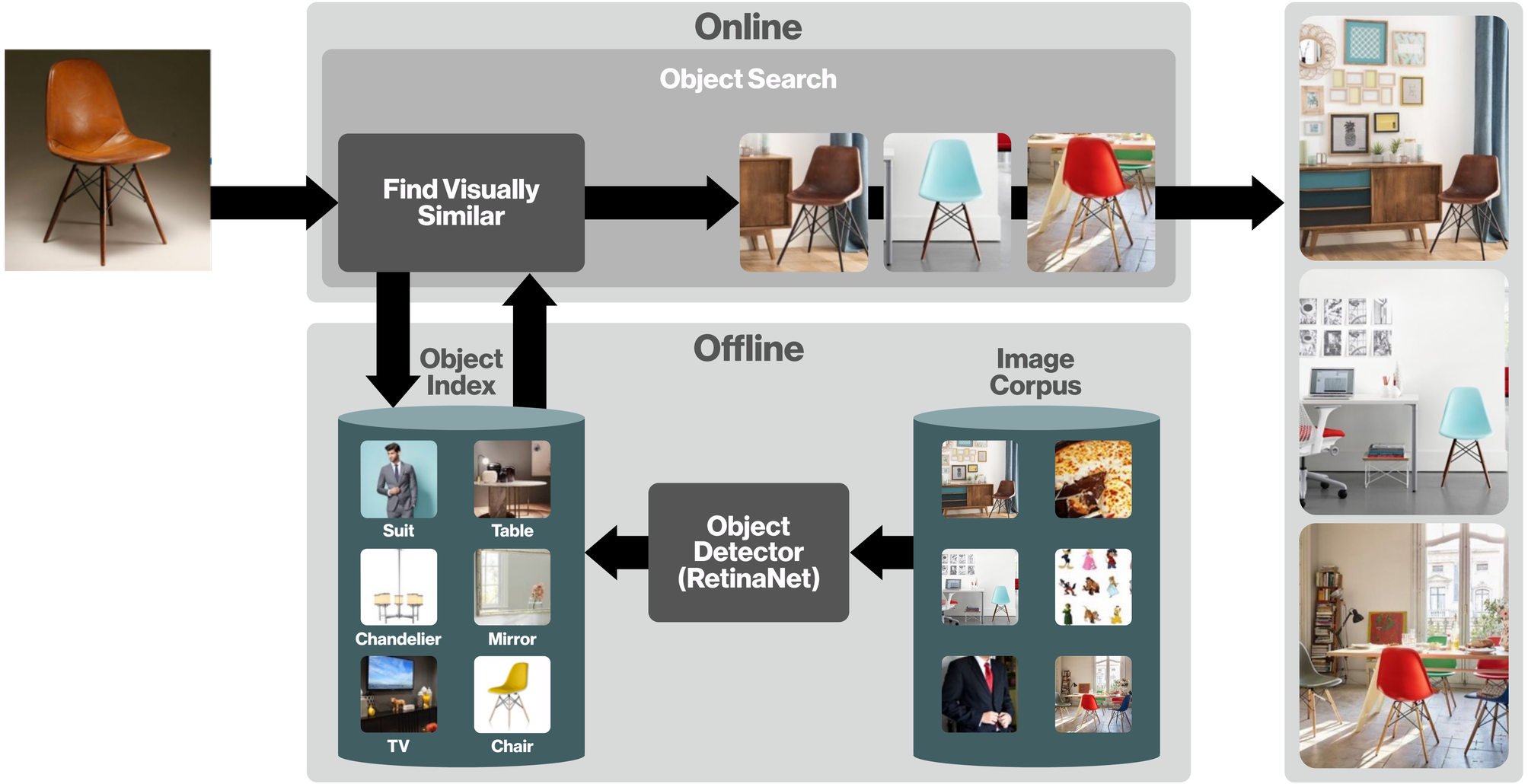
Improving detection with search

Visually similar results

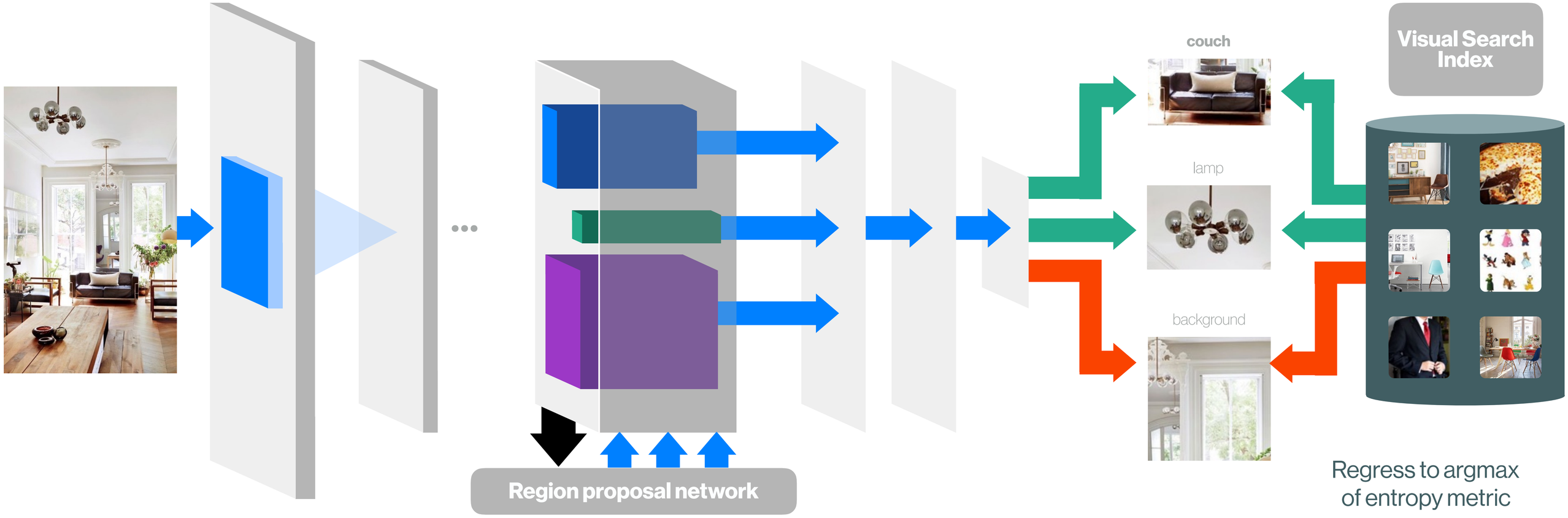


Building **object-to-scene** visual search

leveraging large-scale product detection



Improving **semantic object detectors**

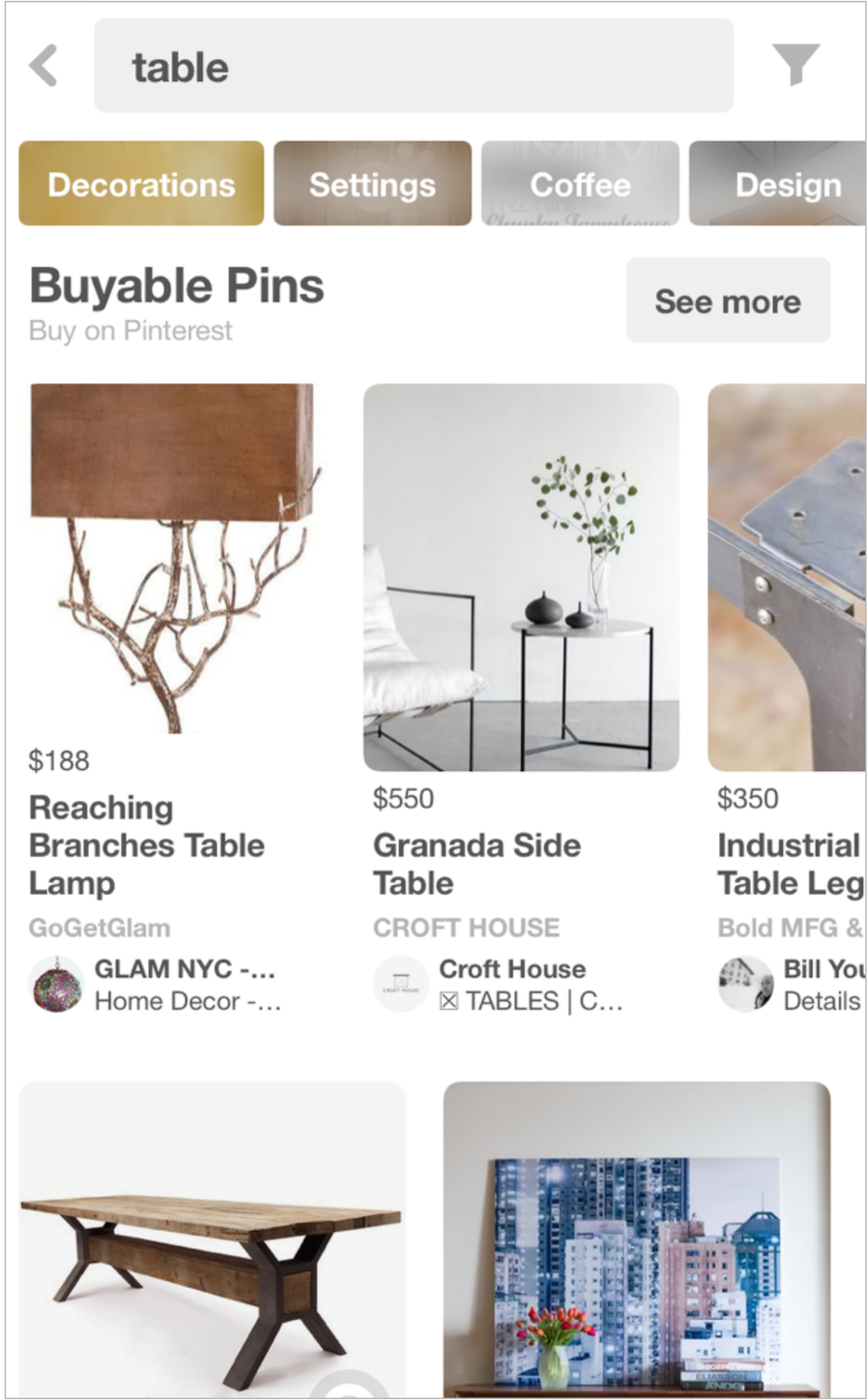


Multi-modal search

with no intent information from user, we must infer a **use-case**

Multi-modal search

with no intent information from user, we must infer a **use-case**



Conclusions

- Deep learning = powerful **featurizer**
- Object detectors = product co-occurrence signals
- Search products moving towards **multi-modal search**