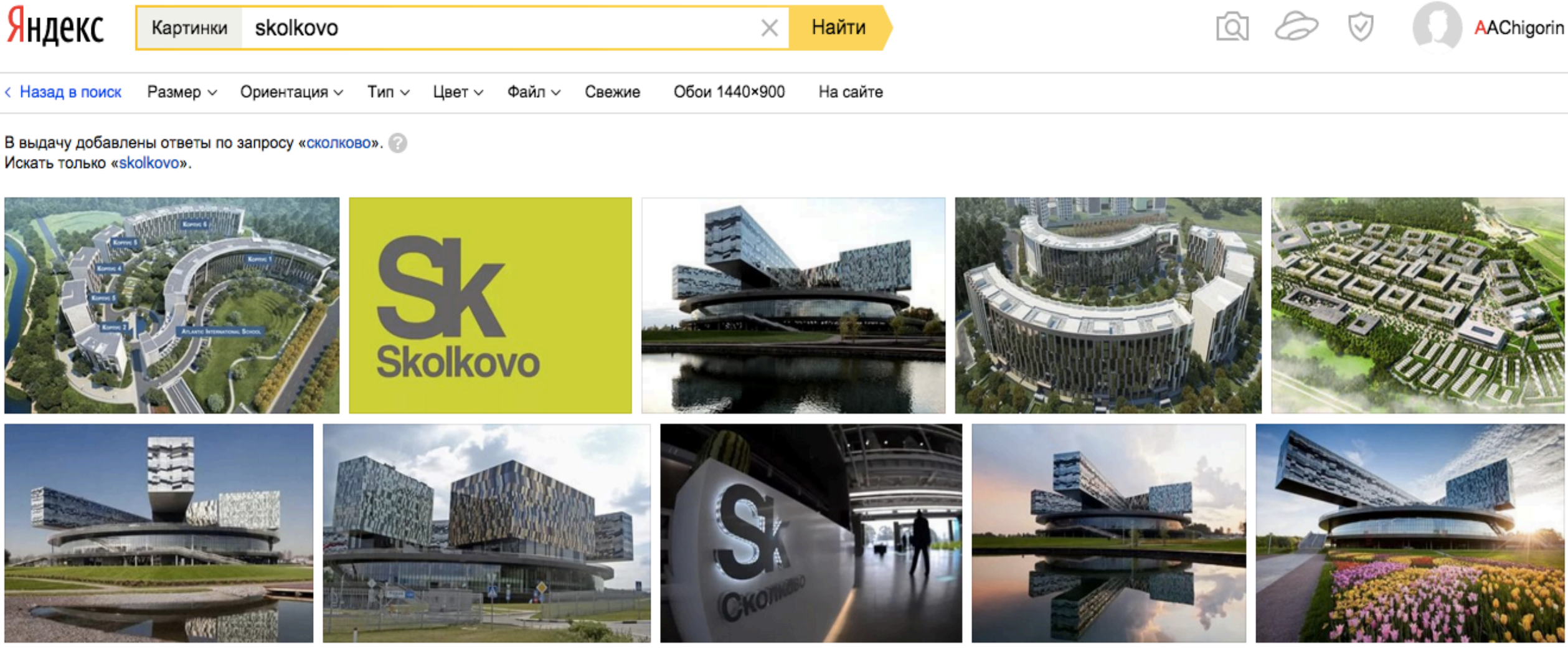
Яндекс

Matching image and query in Yandex.Images search

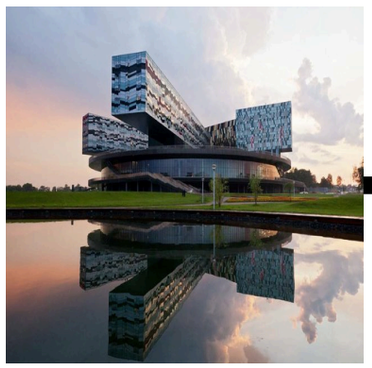
Alexander Chigorin

Yandex images search

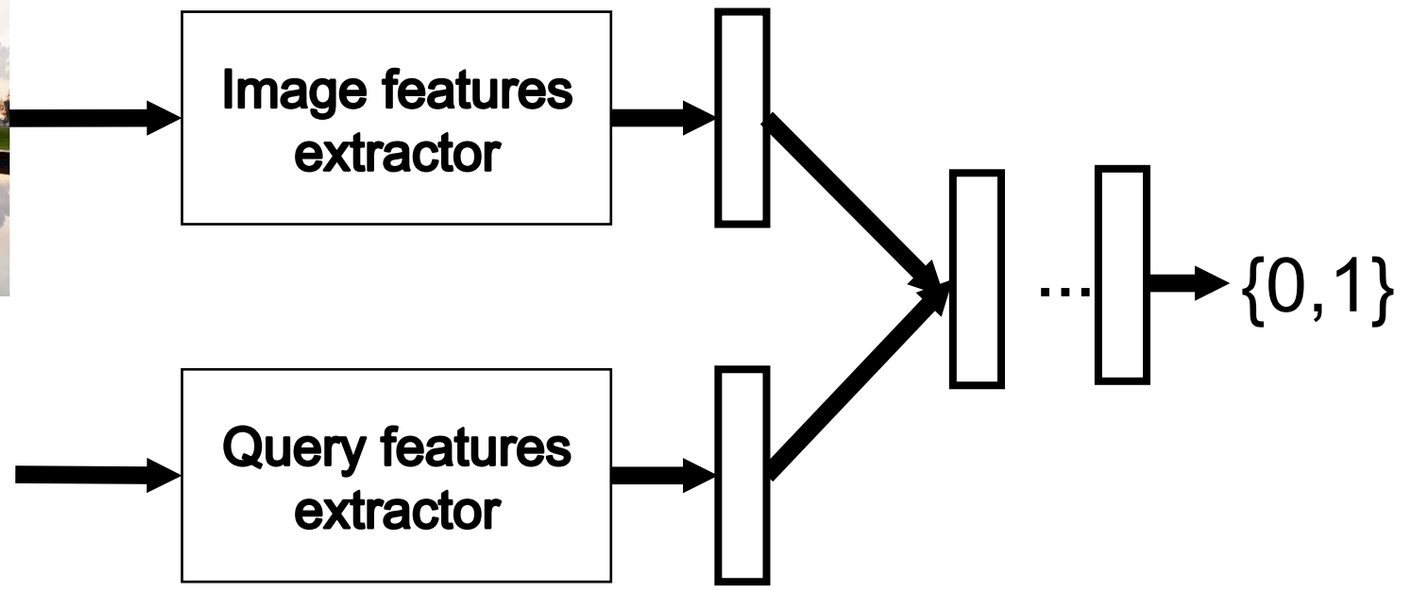


Matching image and query

General pipeline

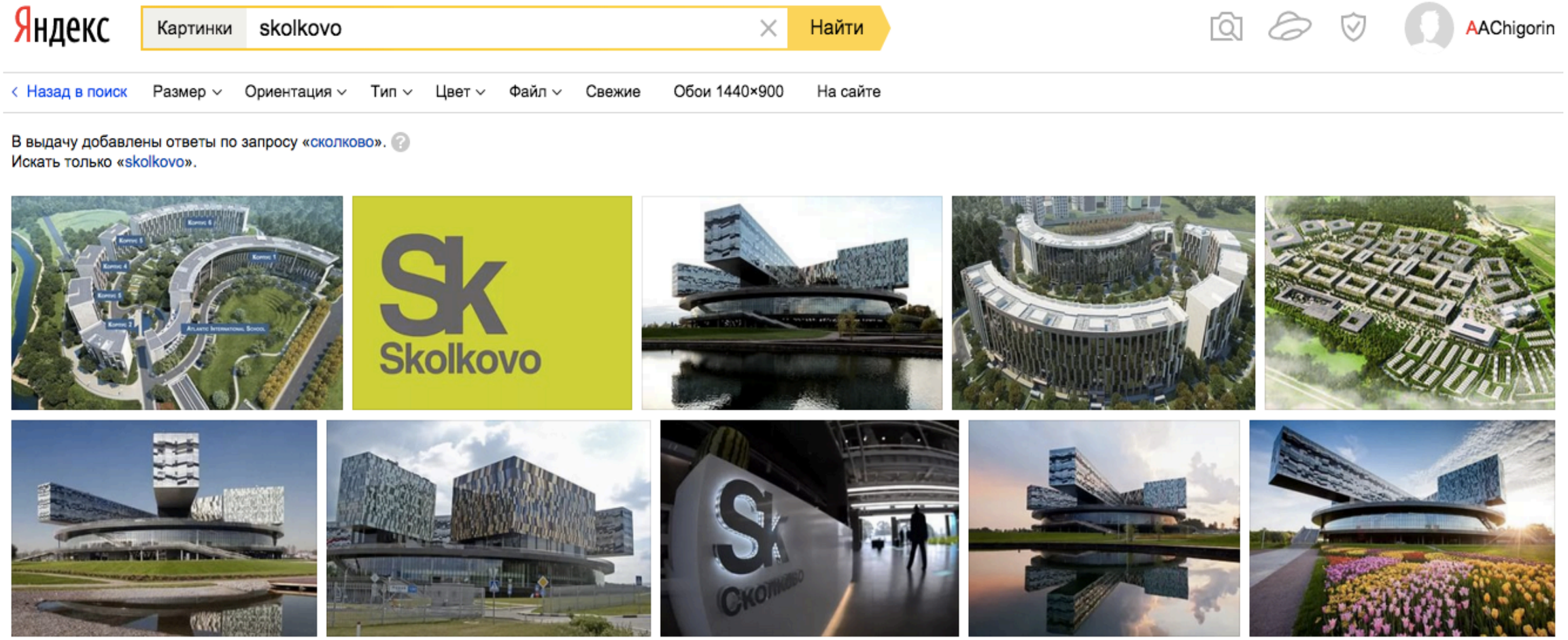


“Moscow School of Management Skolkovo”



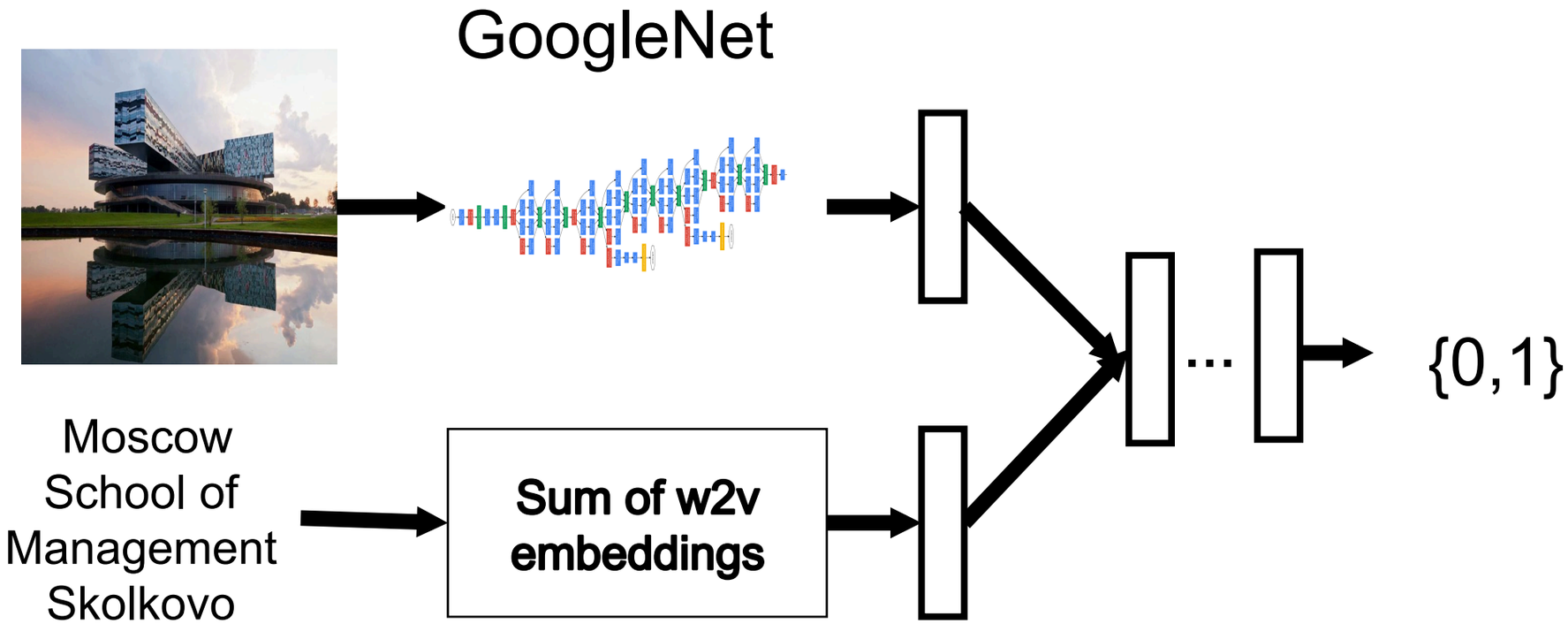
Training data

- Search results of any images search engine
- Top-5 results from 1.1 million queries (positive pairs)
- Random images as negative sample



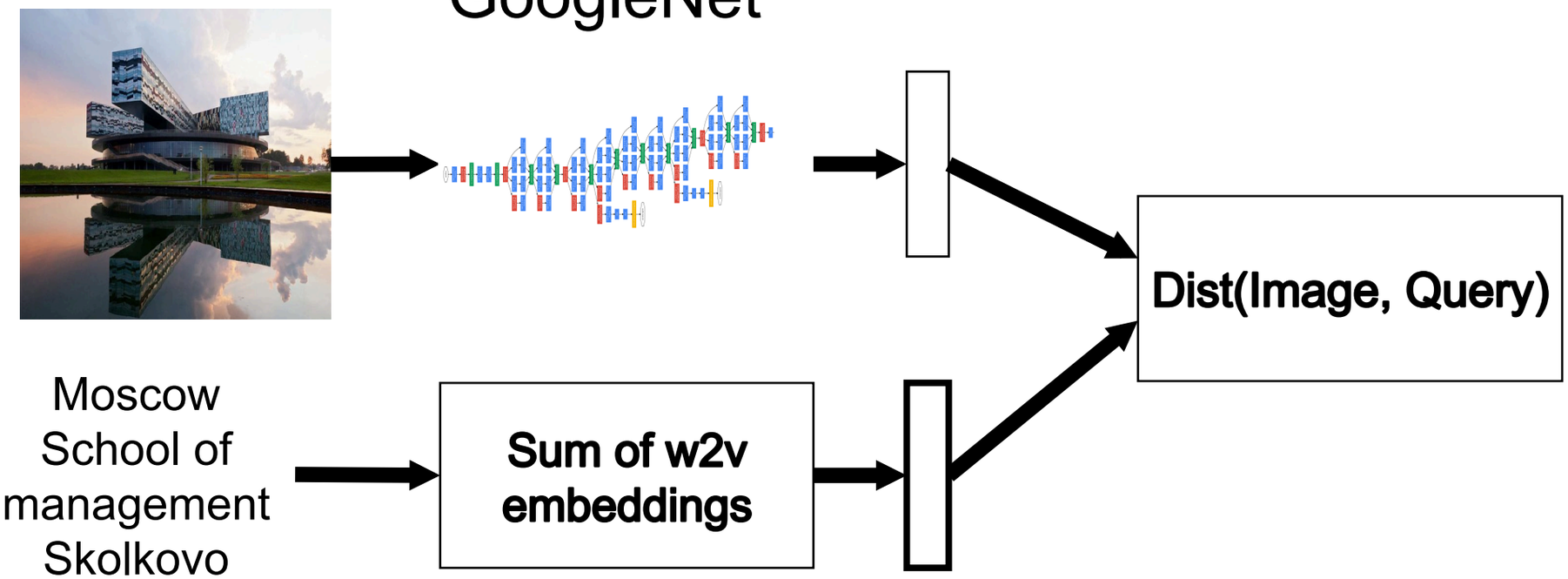
Specifying different parts of the pipeline

- GoogleNet trained on ImageNet
- Word2Vec embeddings **trained on queries**

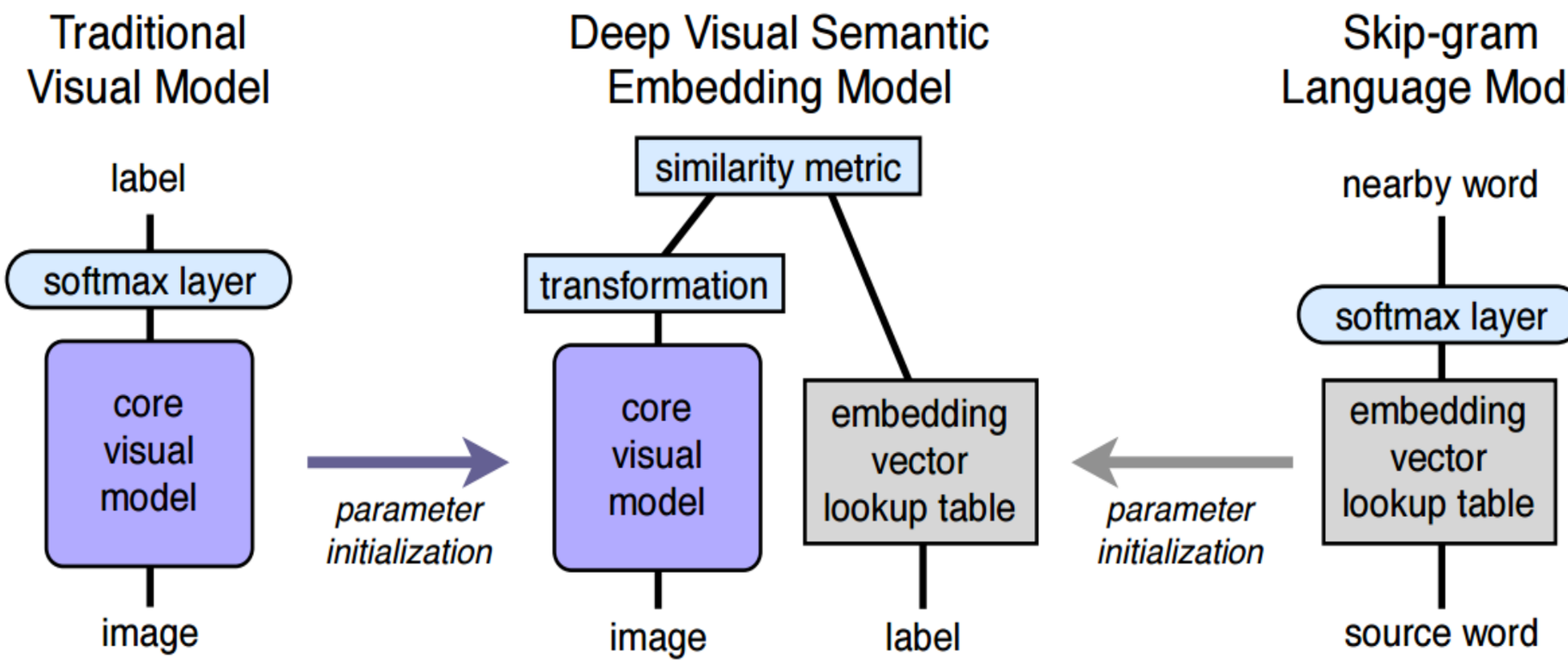


Better solution for production

- Faster at runtime. No need to perform matrix multiplications



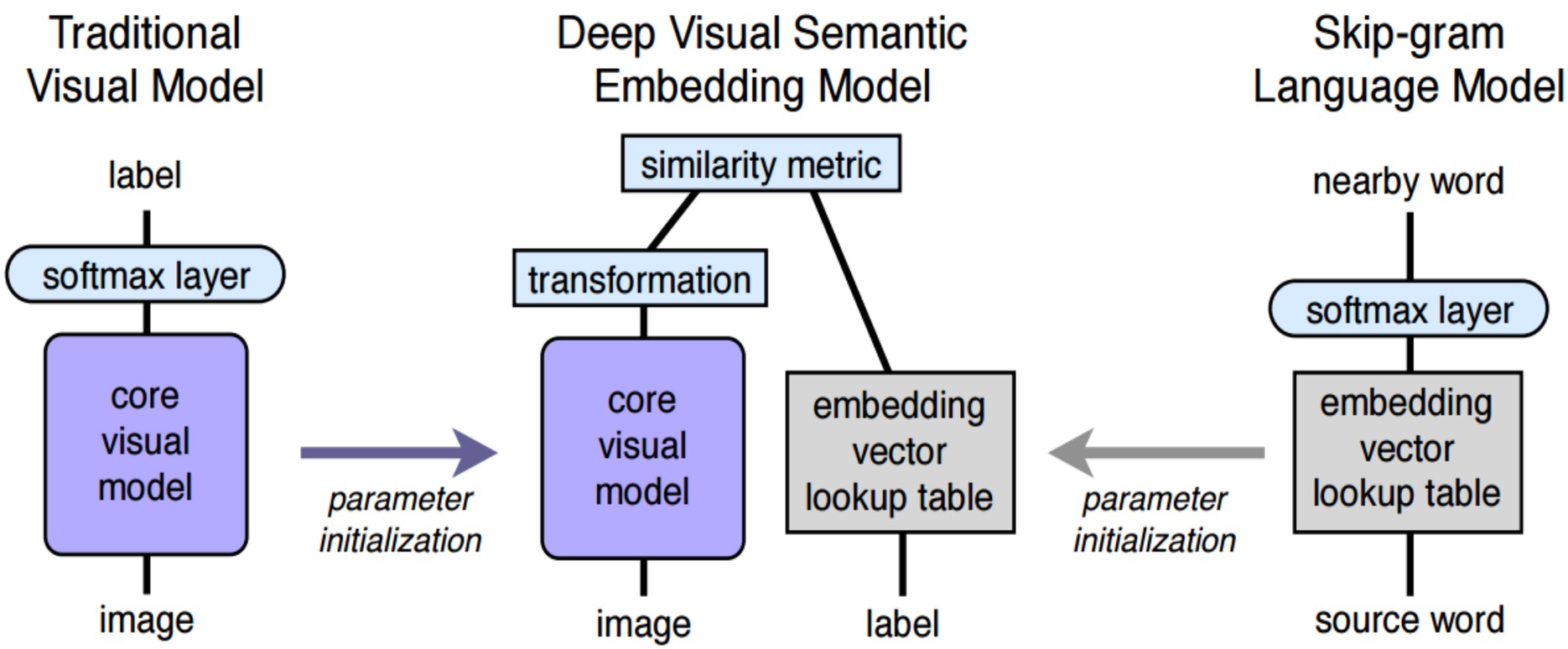
Common semantic space



DeViSE: A Deep Visual-Semantic Embedding Model, Frome A., Corrado G.S. et al.

DSSM: Deep Structured Semantic Model, Huang, He, Gao, Deng et al.

Common semantic space



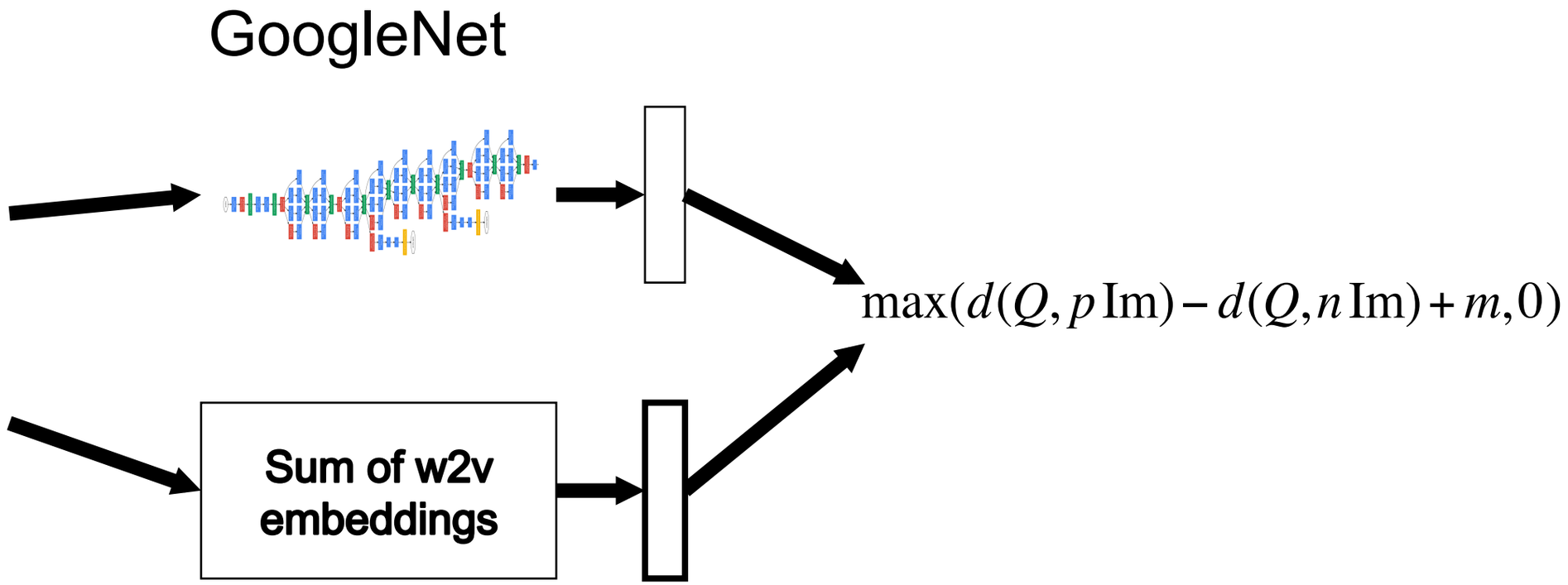
$$loss(image, label) = \sum_{j \neq label} \max[0, margin - \vec{t}_{label} M\vec{v}(image) + \vec{t}_j M\vec{v}(image)]$$

DeViSE: A Deep Visual-Semantic Embedding Model, Frome A., Corrado G.S. et al.

DSSM: Deep Structured Semantic Model, Huang, He, Gao, Deng et al.

The final model

Sample
(*Query*,
Positive image,
Negative image)



First results

Straightforward training of this model already gives profit in production:

- Triplets classification error: 4.01%
- Internal relevance metrics: +2000

Improving basic model. Hard negatives

Possible strategies to get hard negatives:

- Iterative addition of hard negatives during training:
  - Train the network for some time
  - Mine hard negatives for the trained model
  - Add hard negatives to the dataset
  - Iterate
- Dynamic hard negatives
  - Mine hard negatives dynamically during training from the mini-batch
  - No need to recompute entire network!

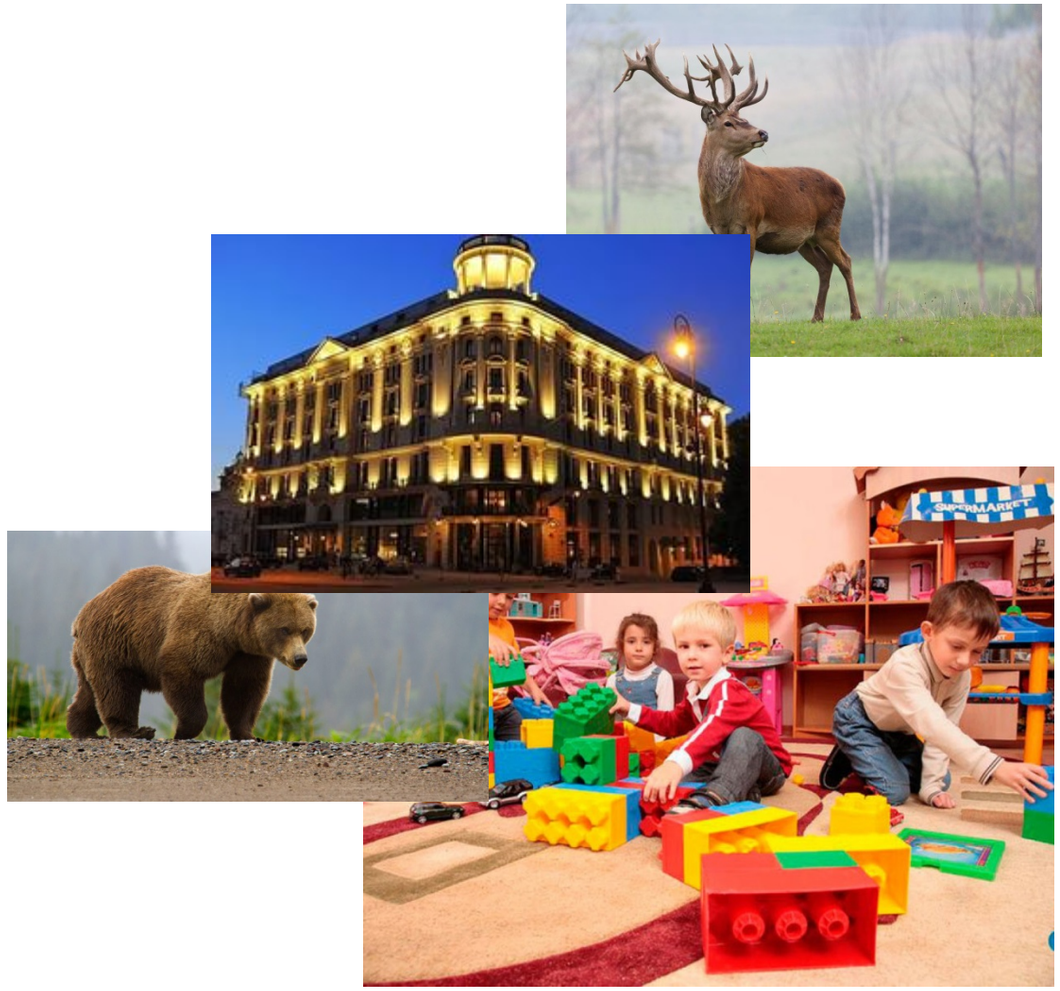
Dynamic hard negatives from mini-batch

Mini-batch

	**racehorse**
	**hotel**
	**kindergarten**
	**deer**

Dynamic hard negatives from mini-batch

racehorse



Dynamic hard negatives from mini-batch

- Take N random images from the mini-batch and select the hardest example among them
- If N = 1 – we get random negatives

racehorse

Similarity

0.65

0.3

Dynamic hard negatives from mini-batch

- Take N random images from the mini-batch and select the hardest example among them
- If N = 1 – we get random negatives

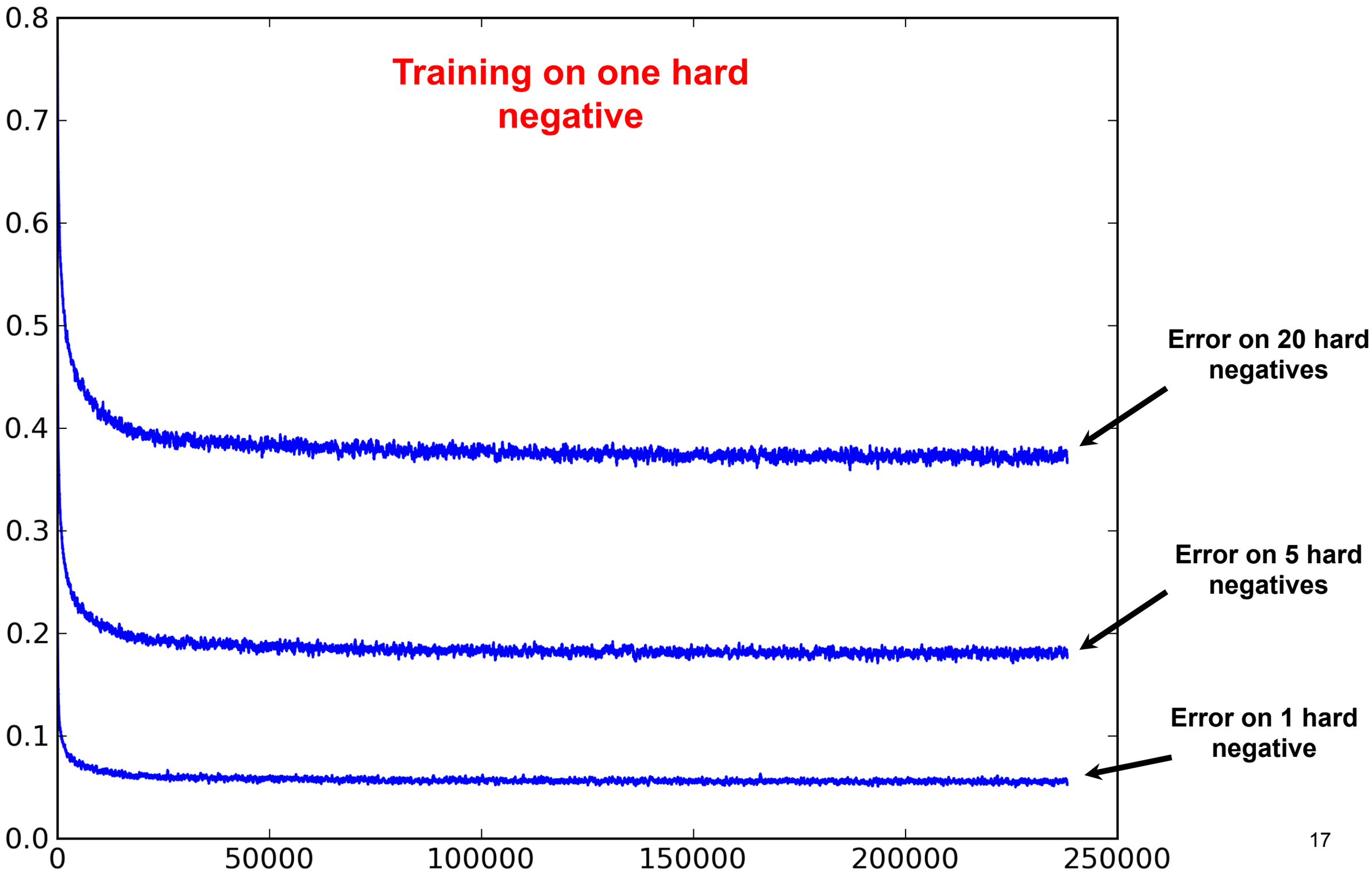
racehorse

Similarity

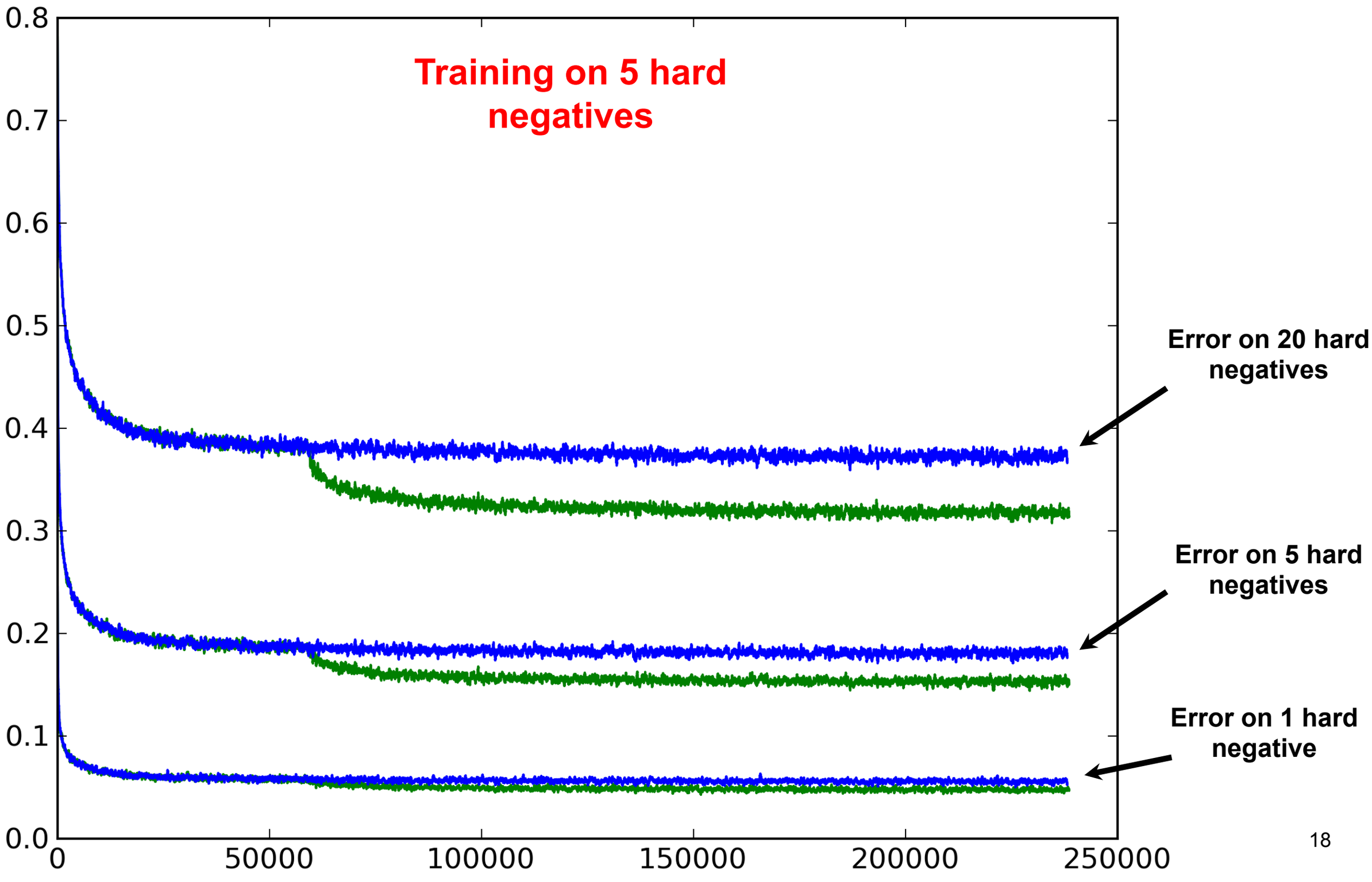
0.65

0.3

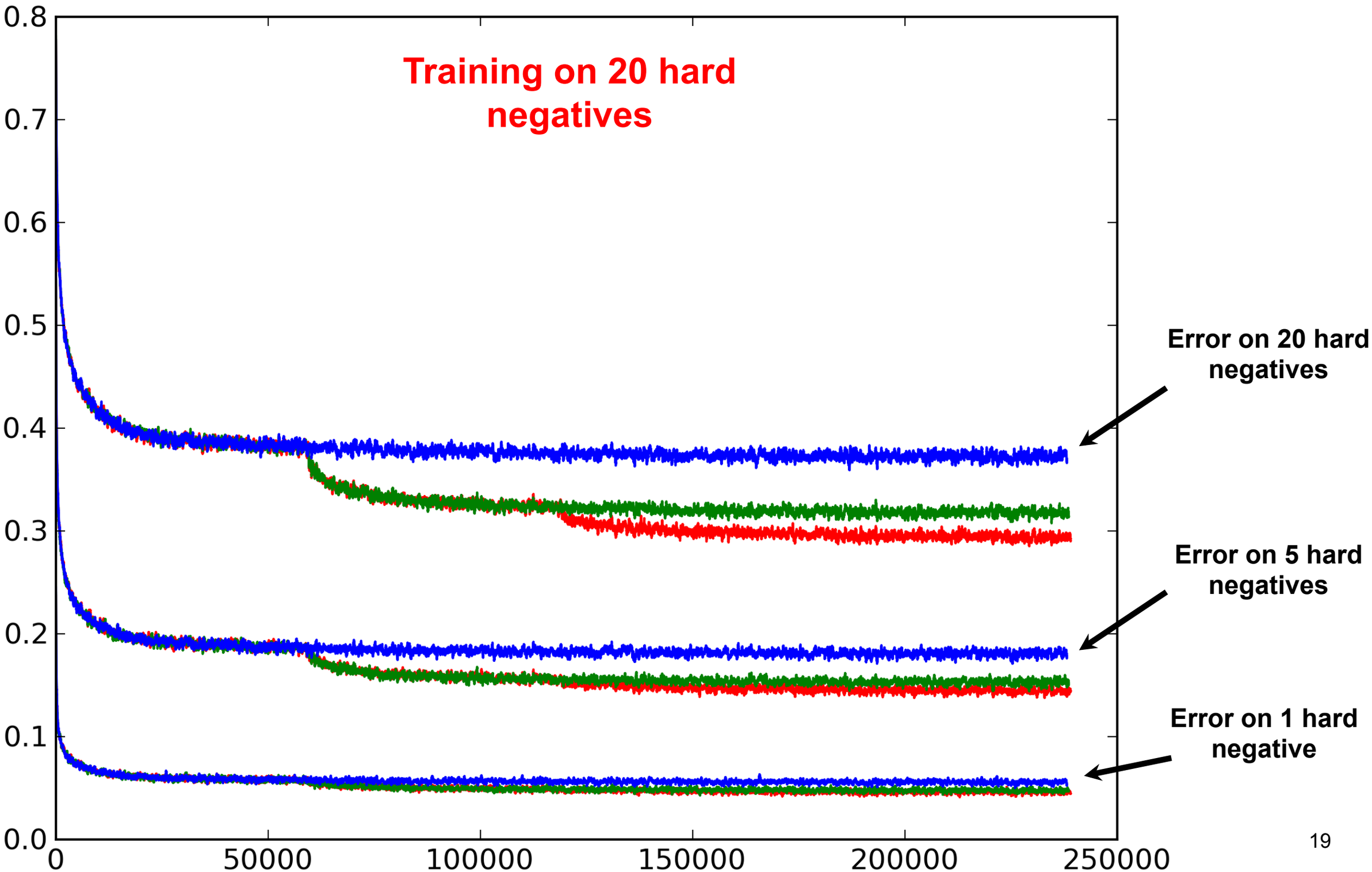
Changing the number of hard negatives



Changing the number of hard negatives



Changing the number of hard negatives



Curriculum learning

- Sometimes it is beneficial to gradually increase the number of hard negatives, otherwise network does not start to learn
- But in our experiments curriculum learning have not increased the final accuracy of the classifier

Results

One hard negative during training

Triplets classification error:

- 1 image: 4.01%
- 10 images: 20.23%
- 20 images: 28.72%
- 40 images: 38.81%

20 hard negatives during training

Triplets classification error:

- 1 image: 3.72%
- 10 images: 16.74%
- 20 images: 23.44%
- 40 images: 31.10%

Results

One hard negative during training

Triplets classification error:

- 1 image: 4.01%
- 10 images: 20.23%
- 20 images: 28.72%
- 40 images: 38.81%

Relevance metrics: +2000

20 hard negatives during training

Triplets classification error:

- 1 image: 3.72%
- 10 images: 16.74%
- 20 images: 23.44%
- 40 images: 31.10%

Relevance metrics: +1900

That is strange. We should get more

Adding reverse ranking to the model

Sim

0.65

0.5

0.3

-0.1

racehorse

Sim: 0.8

Adding reverse ranking to the model

Sim

0.65

0.5

0.3

-0.1

racehorse

Sim: 0.8

Sim

deer 0.58

bear 0.53

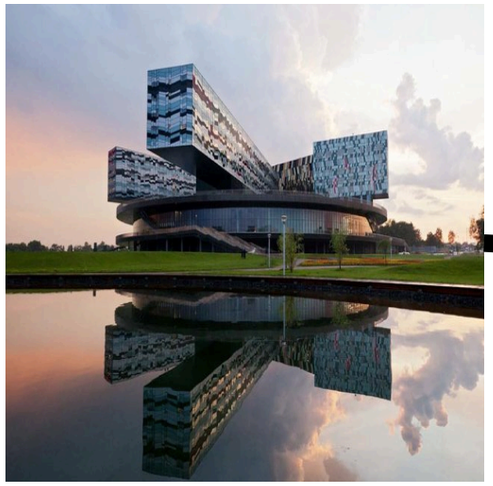
kindergarten 0.2

hotel -0.18

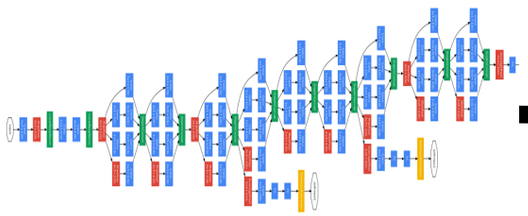
Result with additional reverse ranking

Test \ Train	**Direct ranking** 1 negative	**Direct ranking** 20 negatives	**Direct+reverse ranking** 20 negatives
Direct ranking 1 negative 40 negatives	**4.01%** **38.81%**	**3.72%** **31.10%**	**3.89%** **32.22%**
Reverse ranking 1 negative 40 negatives	**5.69%** **47.55%**	**9.16%** **50.69%**	**4.39%** **34.55%**
Relevance metrics	**~2000**	**~1900**	**~2500**

Going deep. Finetuning the text part



GoogleNet

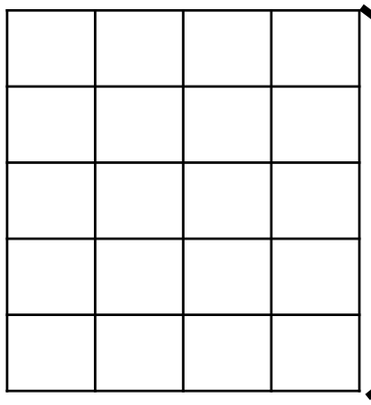


Cos(Image, Text)

w2v trained on queries

Moscow School of Management Skolkovo

Moscow
School
of
Management
Skolkovo



sum

Results of dictionary finetuning

Train / Test	Without dictionary finetuning	With dictionary finetuning
Direct ranking 1 image 40 images	**3.89%** **32.22%**	**3.17%** **27.12%**
Reverse ranking 1 image 40 images	**4.39%** **34.55%**	**3.21%** **28.21%**
Relevance metrics	**~2500**	**~4000**

The largest change in word embeddings

Largest change of distance: dist(old_emb,new_emb)

L2		**cosine**	
صور	8.00618	co	1.89609
nasıl	7.1368	c	1.87288
фриске	6.65218	по	1.85821
minecraft	6.46264	в	1.82775
lisesi	6.44262	о	1.81785
кроссворд	6.21367	фото	1.81259
демотиваторы	6.21074	nasıl	1.79939
ilgili	6.19246	of	1.79792
wolfteam	6.11097	как	1.79545
hentai	5.9108	на	1.79018

The final recipe

- Map different modalities in the common space (fast in production)
- Mine dynamic hard negatives from the mini batch
- Gradually increase the number of hard negatives (if network does not start to train)
- Mine hard negatives from all modalities
- Finetune all parts of the model

Generating query by image

Generating query by image

меню в макдональдсе

макдональдс меню и цены

бургер кинг меню и цены

рыбалка на алтае

турпоход

рыбалка на камчатке

Generating query by image

животные медведь

самый большой медведь

гризли фото

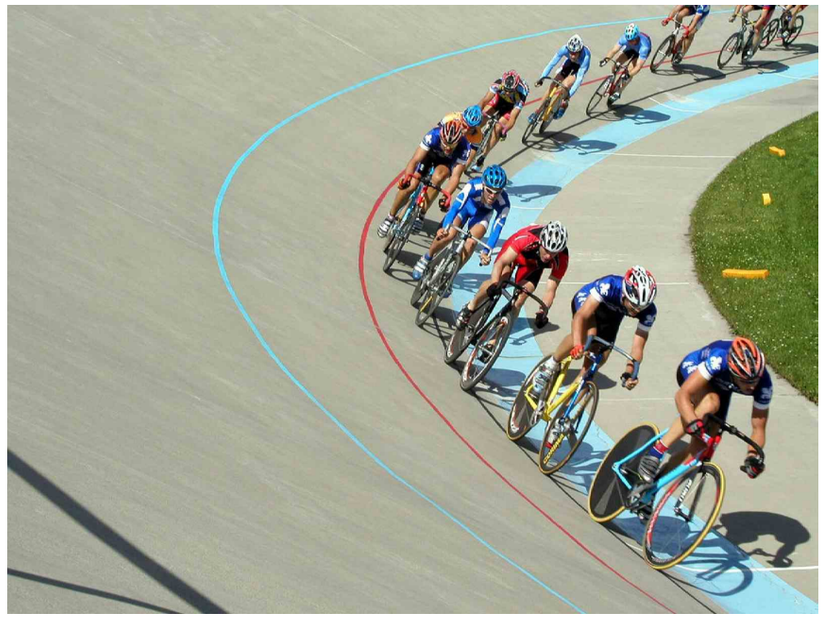
медведь с медвежатами

киев фото майдан

митинг на манежной площади 1991

киев беспорядки

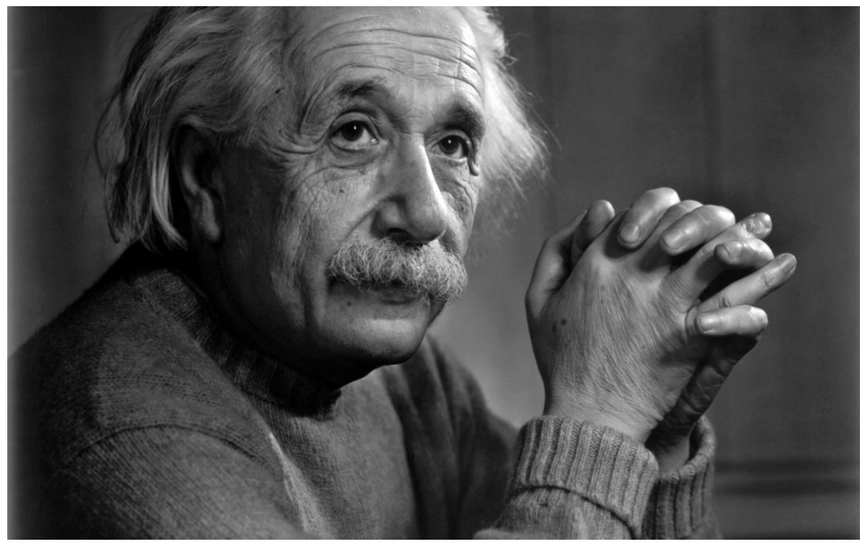
Sometimes it fails…



скоростной бег на коньках фото

парасноуборд

паралимпийский биатлон



смоктуновский фото

гельман

старый человек

Other applications

The same method is also verified in other task:

- Video + Query
- Web page + Query
- User + Ads
- User + Web page