Understanding the Multimodal World

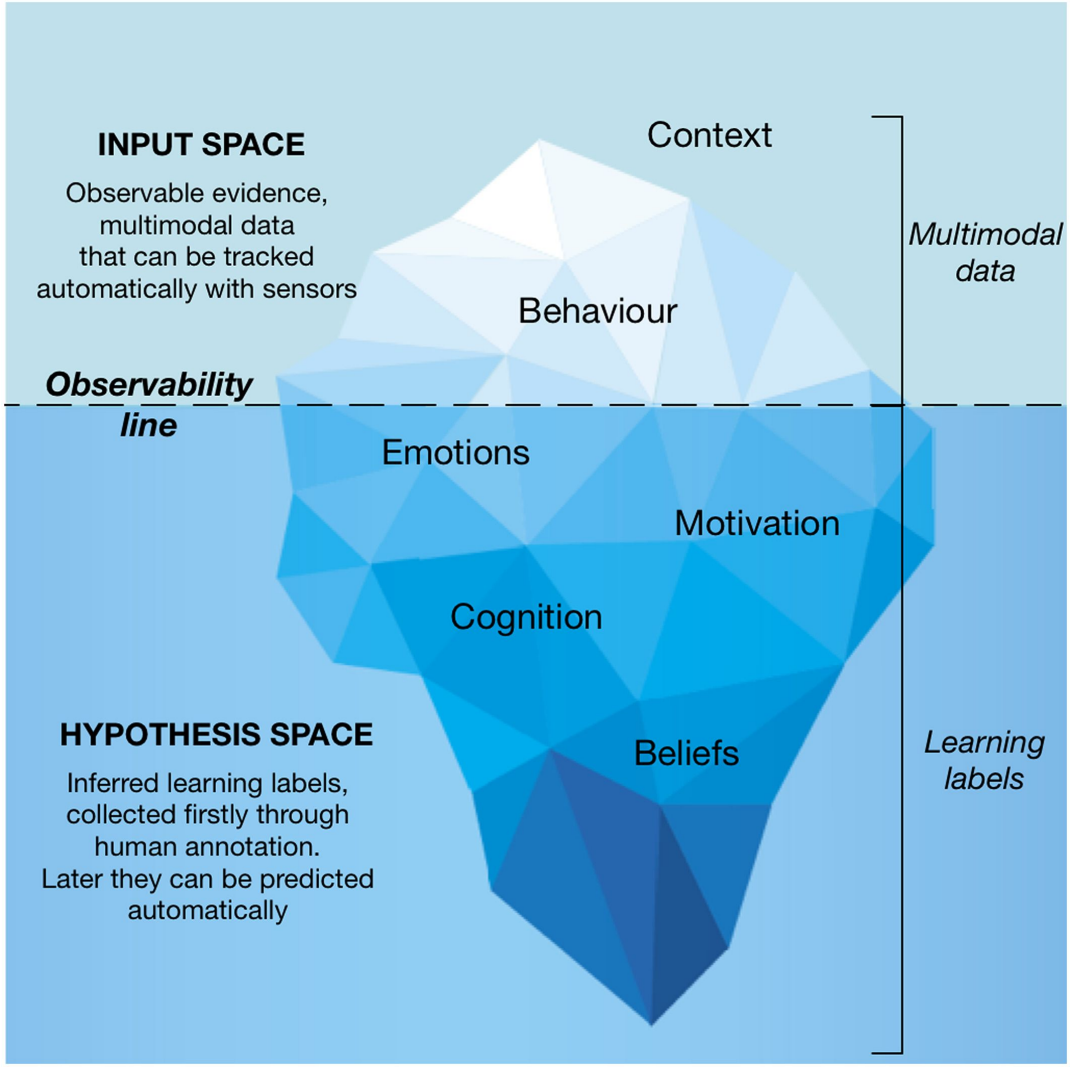
Results and Challenges

Sergey Nikolenko
Artificial Intelligence Lab, PDMI RAS

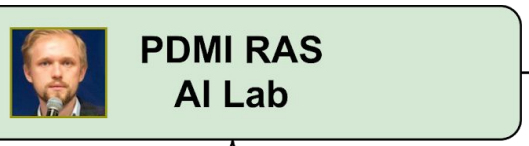
March 23, 2021

Plan

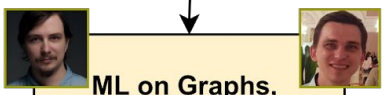
- Embeddings in modern ML
- The world is multimodal: how do we cope?
- Multimodal embeddings, different approaches
- A review of three state of the art 2020 papers on multimodal deep learning
- Our own example: recent work on symbolism prediction



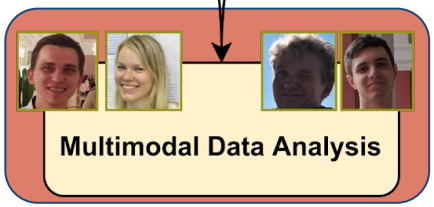
Artificial Intelligence Lab, PDMI RAS



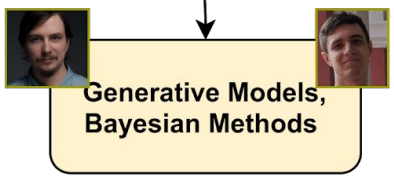
1. New Probabilistic Neural Networks
2. Mobile applications: facial analytics, on-device scene analysis
3. Depth estimation, 3D Model Reconstruction and Mapping
4. Deep learning for rendering
5. Segmentation for satellite images



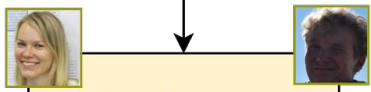
1. Link prediction on graphs with mobile applications
2. ML for imaging mass-spectrometry and cell biology
3. Efficient classification with small samples
4. RL in games and interpretable RL
5. Fault Detection in Tennessee Eastman processes



Symbolism Prediction in Advertising



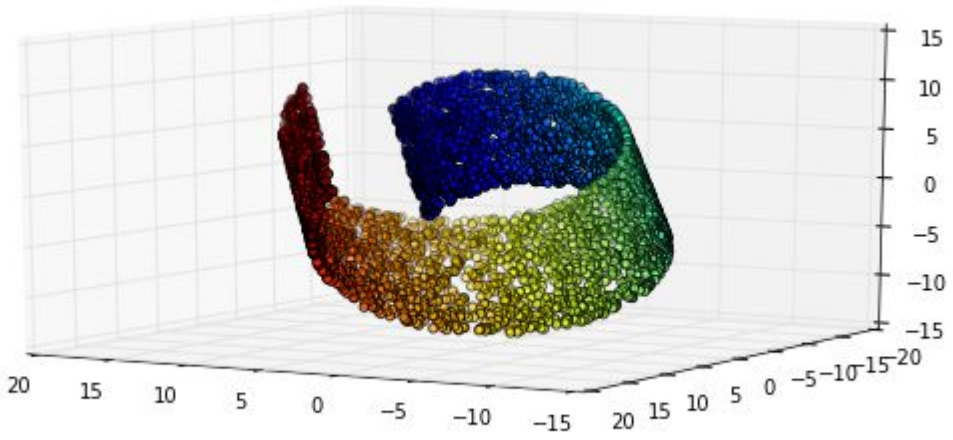
1. VAE for Recommendations
2. GANs for style transfer
3. Generative models for drug discovery
4. Postprocessing of rendered 3D scenes
5. GANs for fashion

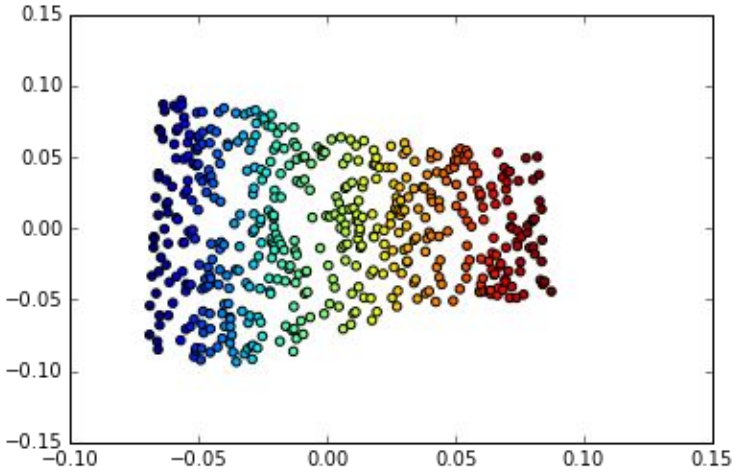


1. Named entity recognition and fact extraction from texts
2. Aspect-based recommendations
3. Transformer-based conversational agents
4. Topic modeling
5. Word forms and semantics in morphology-rich languages

Learning Latent Representations

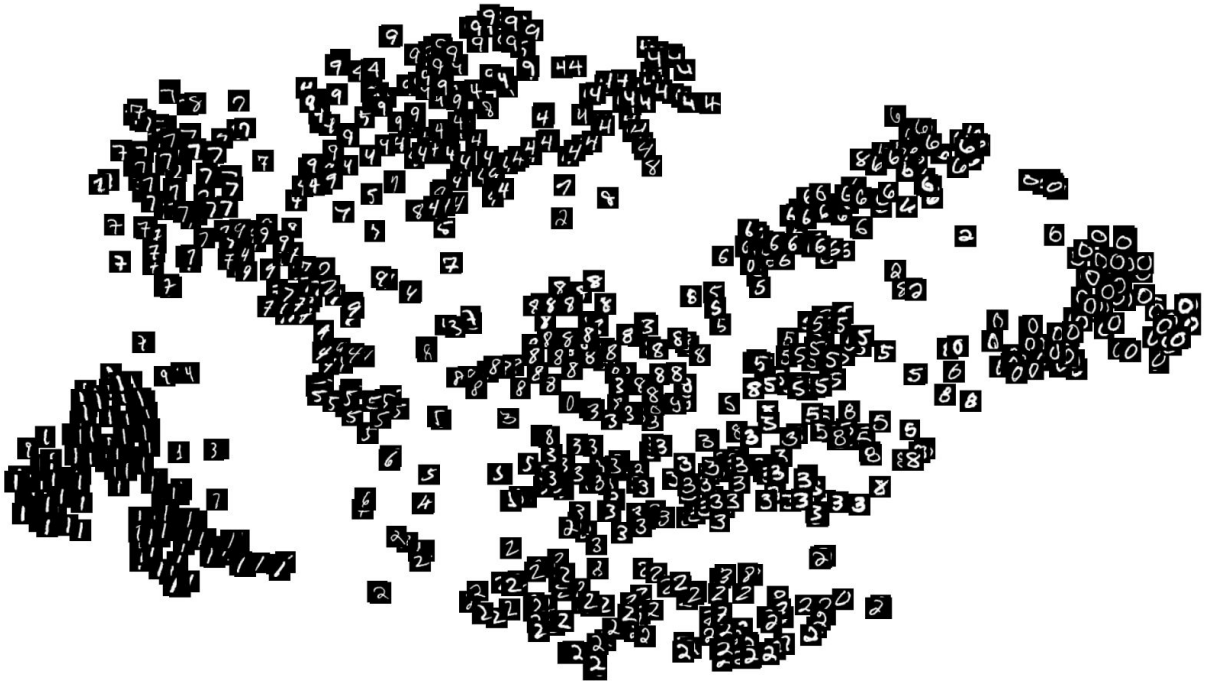
- **Latent representations**, or **embeddings**, is one of the most important concepts in modern machine learning
- Mathematically it's just a parametrization of a manifold: we assume that a set in a high-dimensional space actually has a lower dimension and try to find the parametrization

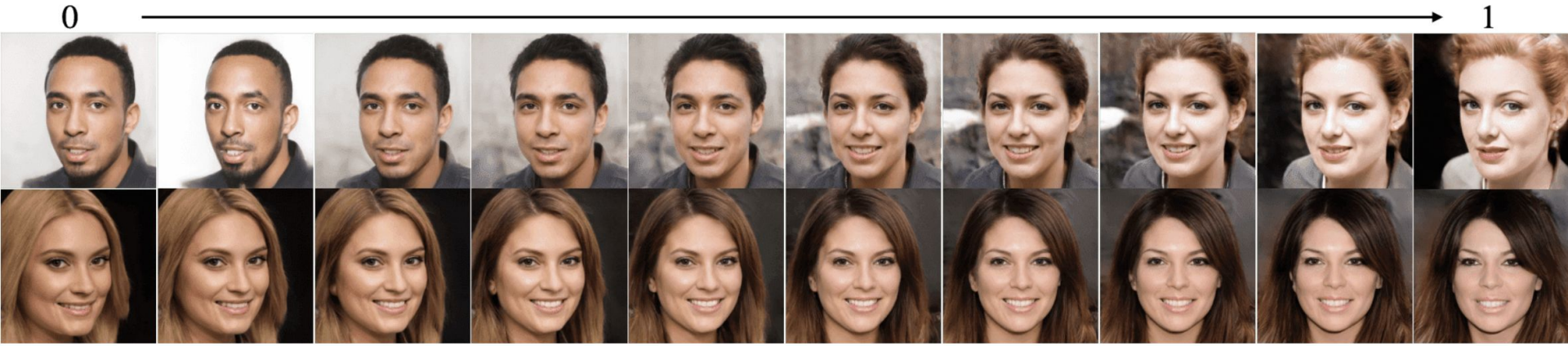




Learning Latent Representations

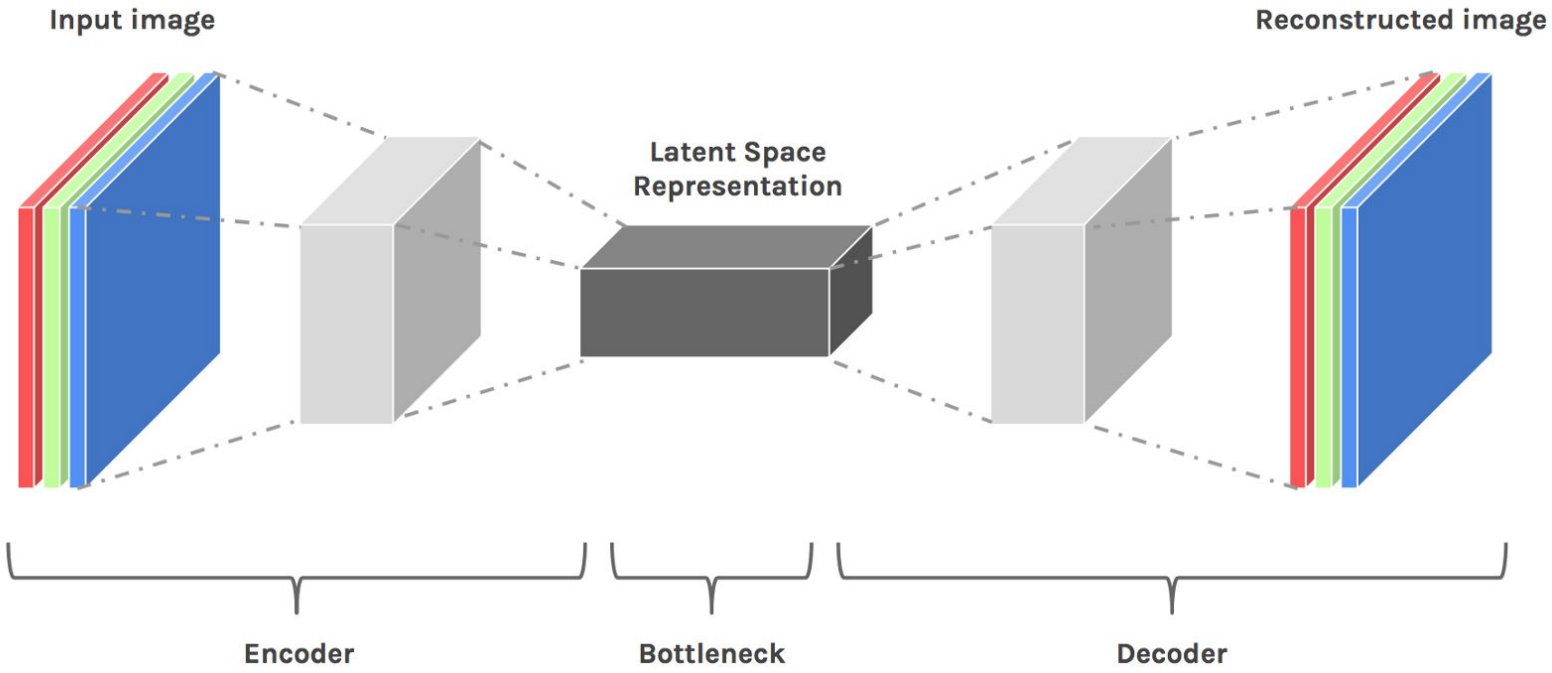
- The high-dimensional points can be images, sentences, sounds...
- We would like the latent space to have good properties; how can we find a good parametrization?





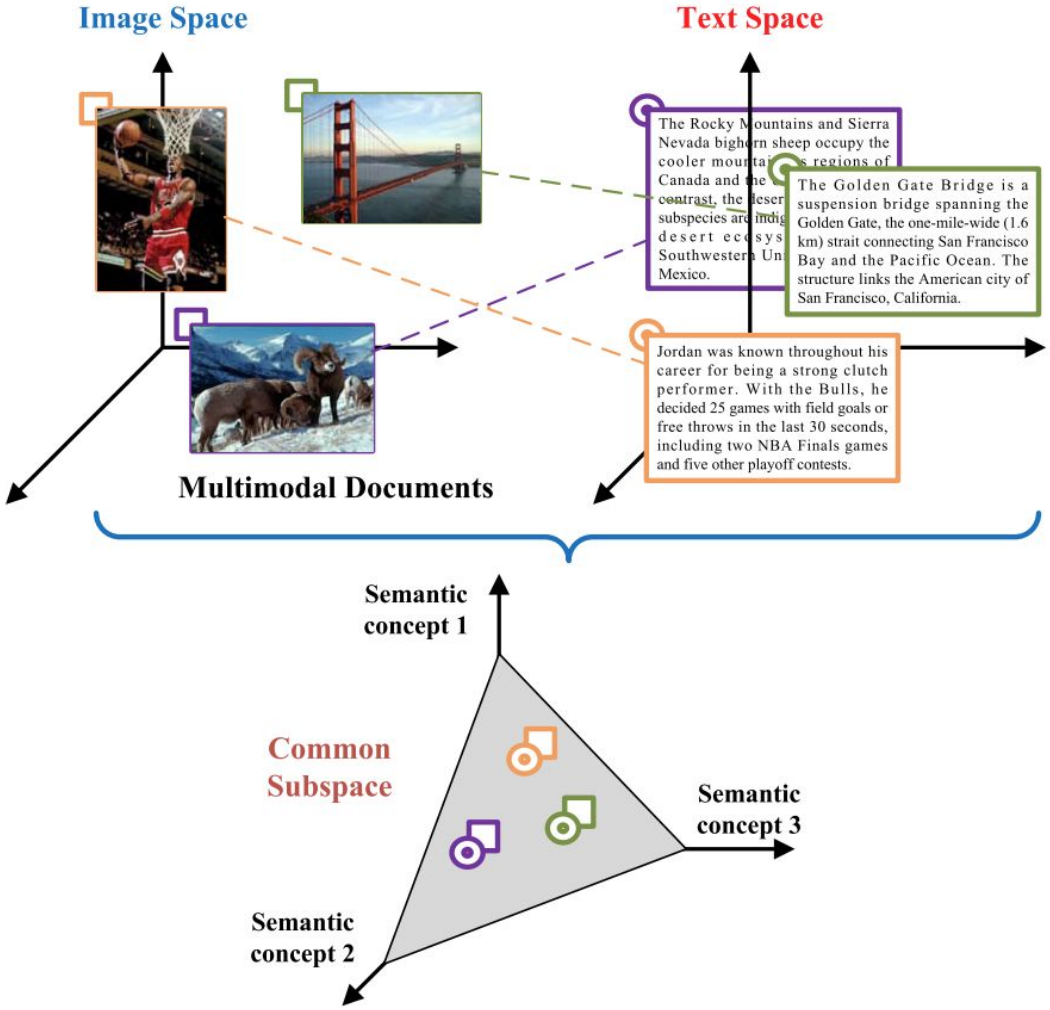
Learning Latent Representations

- **Autoencoders**: let's train a model to reconstruct a point from its latent space representation; datasets are not a problem now
- The important question here is how to ensure a good structure for the distribution of latent representations: VAE, AAE etc.; let's not go there

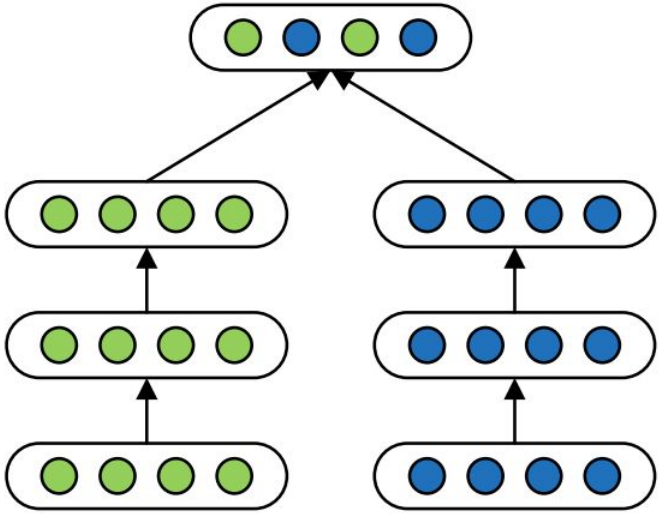


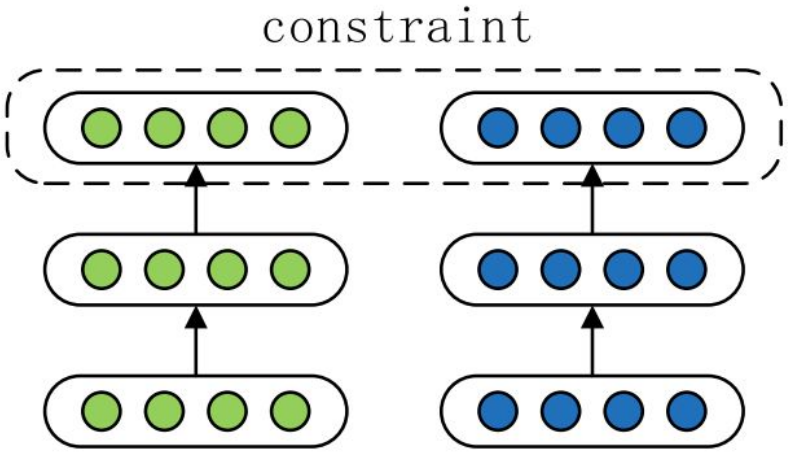
How does multimodal analysis work?

- Most multimodal analysis attempts to learn joint embeddings in some unified latent space
- The objective is to construct a vector space that brings semantically similar elements from different modalities close together

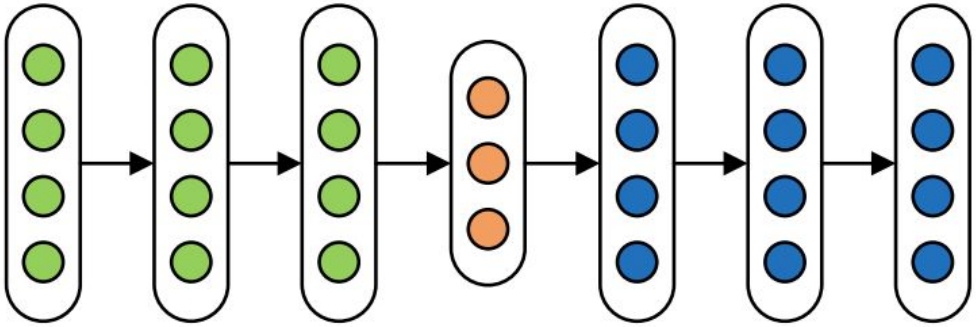


How can we learn joint embeddings?



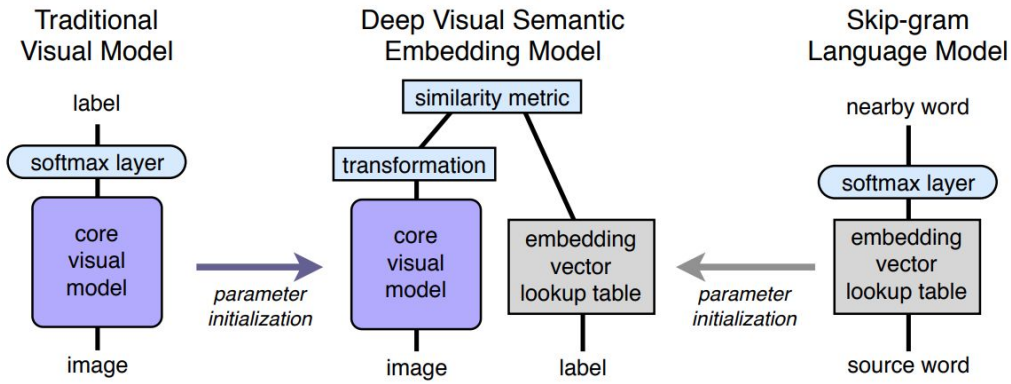


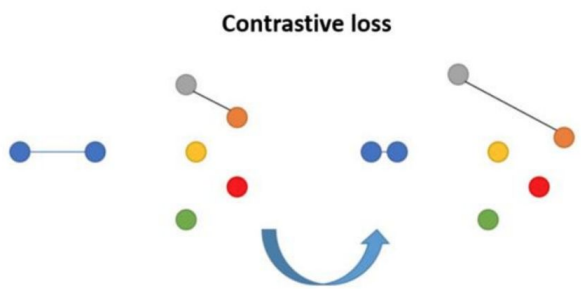
- Learn a **joint** representation
- Learn **coordinated** representations
- Learn an **encoder-decoder** transformation from one modality to another

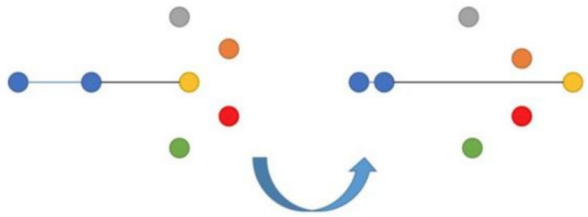


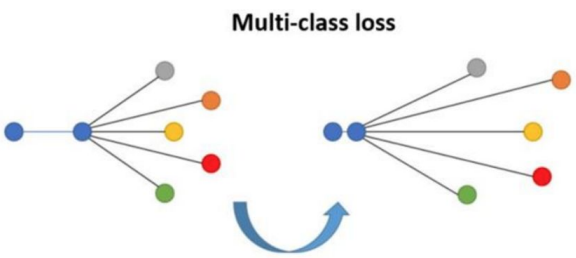
Constraints via ranking losses

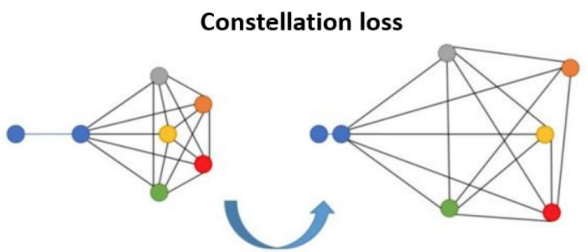
- The “constraint”, bringing together vectors from different modalities, is usually implemented via some kind of ranking loss: contrastive, triplet loss etc.
- DeViSE (Frome et al., 2013); VSE++ (Faghri et al., 2018)





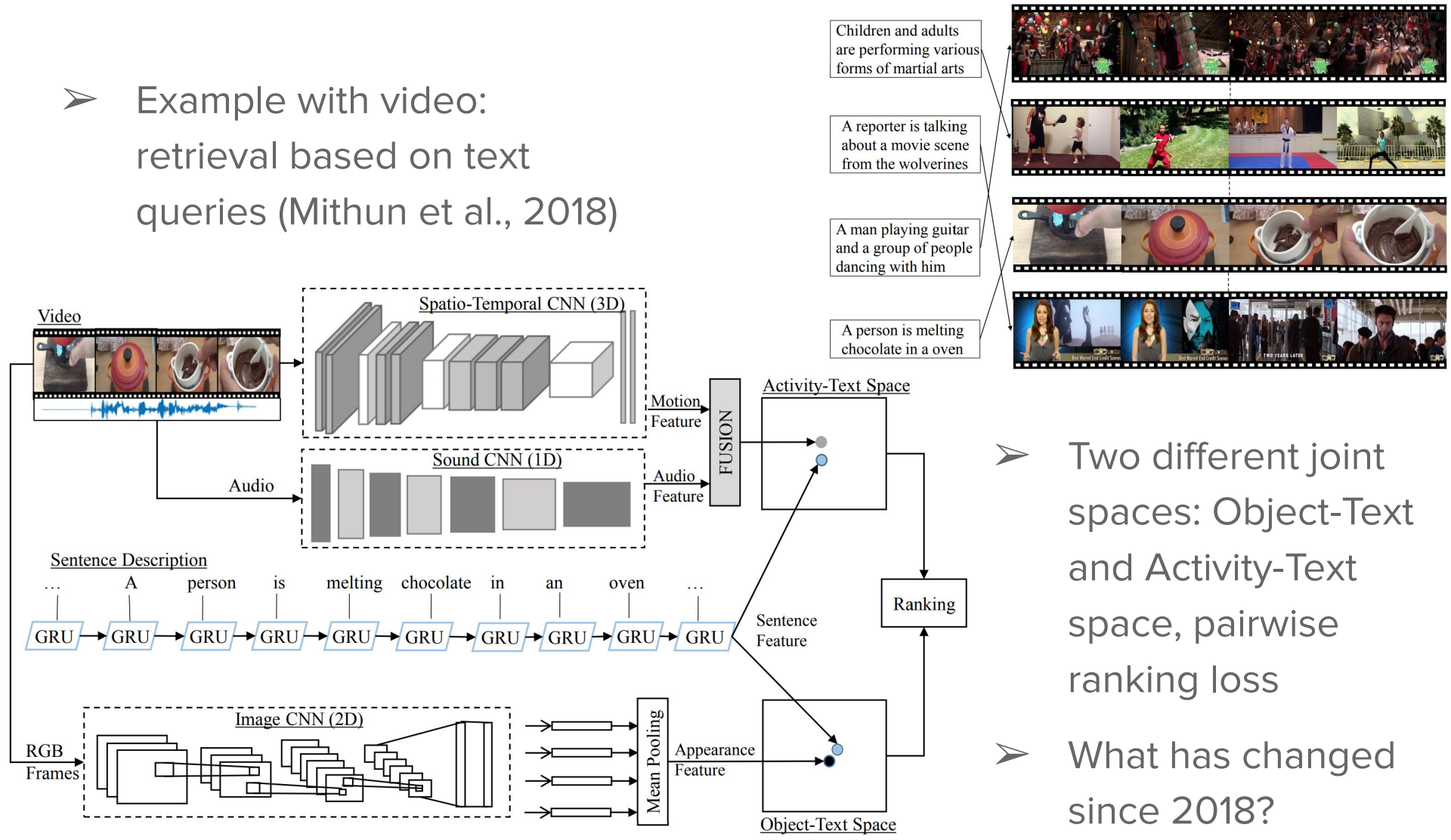






A classical example

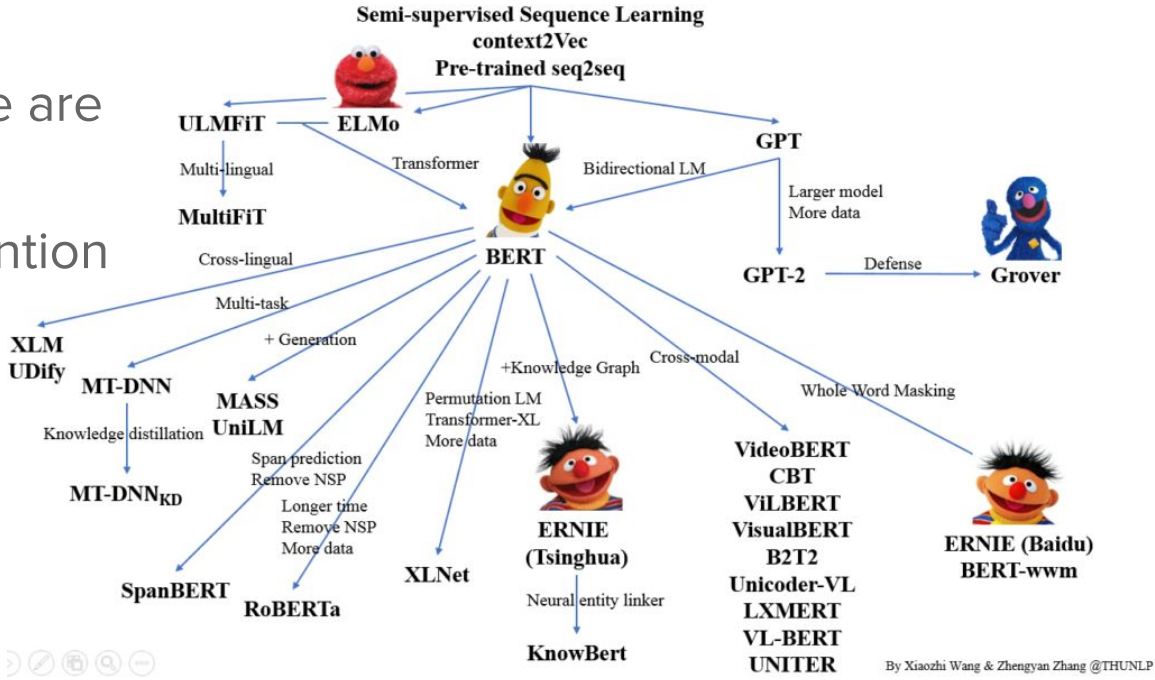
- Example with video: retrieval based on text queries (Mithun et al., 2018)

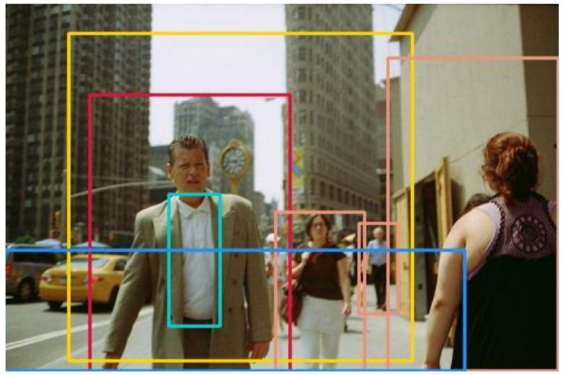


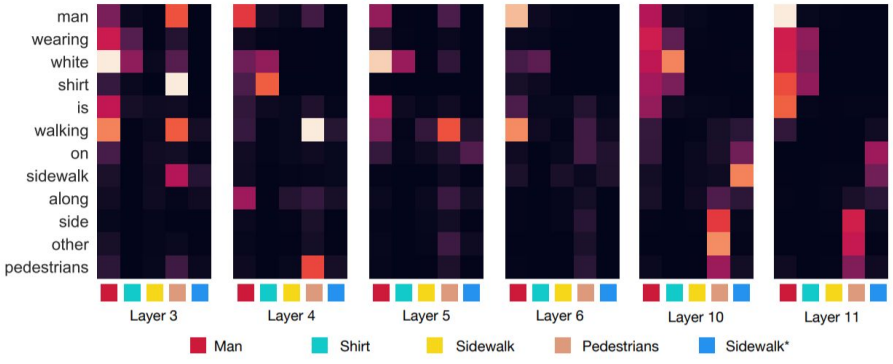
- Two different joint spaces: Object-Text and Activity-Text space, pairwise ranking loss
- What has changed since 2018?

Transformers!

- By now, many state of the are models are based on Transformer-like self-attention
- Multimodal too: ViLBERT, LXMERT, VL-BERT, VisualBERT, UNITER, InterBERT, PixelBERT, ERNIE-VIL, DeVLBert...



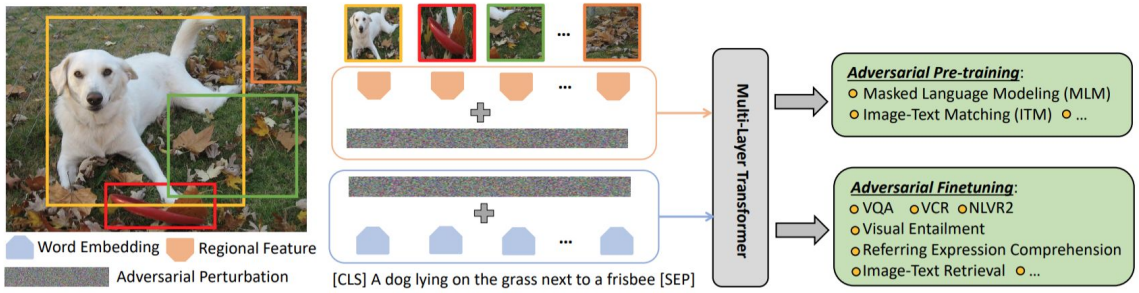


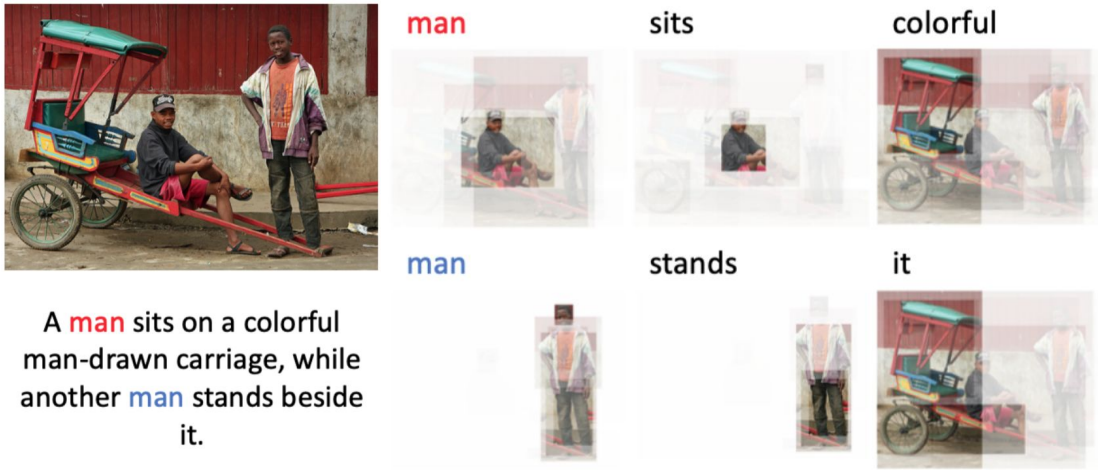


- Let's check out 2020...

Adversarial Training for Vision-and-Language

- Gan et al. (NeurIPS 2020) develop VILLA (Vision-and-Language Large-scale Adversarial training): task-agnostic adversarial pretraining (representation learning), then task-specific adversarial fine-tuning
- V+L Transformers



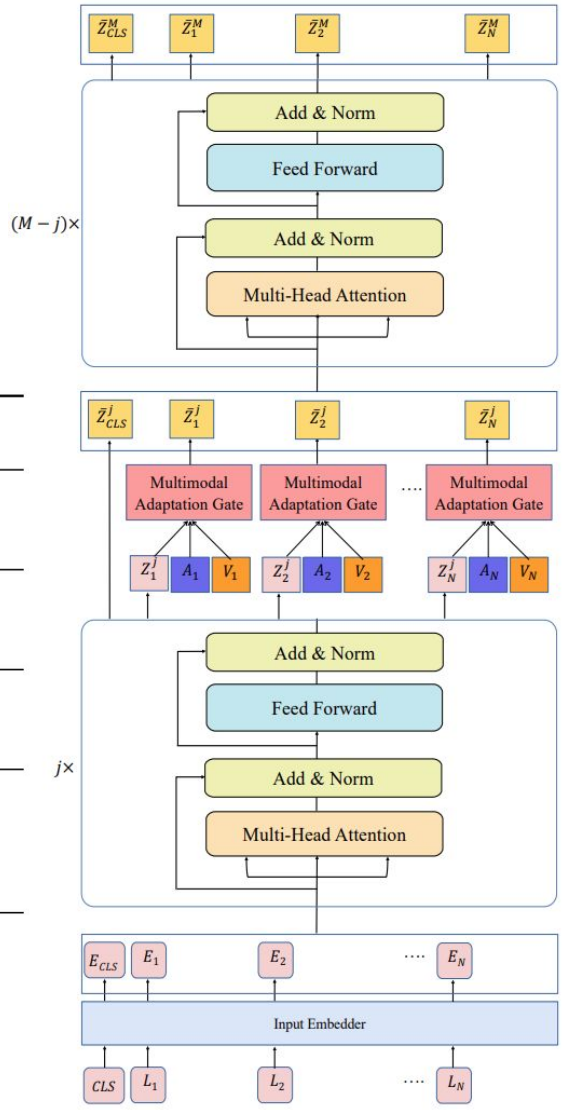


Russian Academy of Sciences

Multimodal Information in Large Transformers

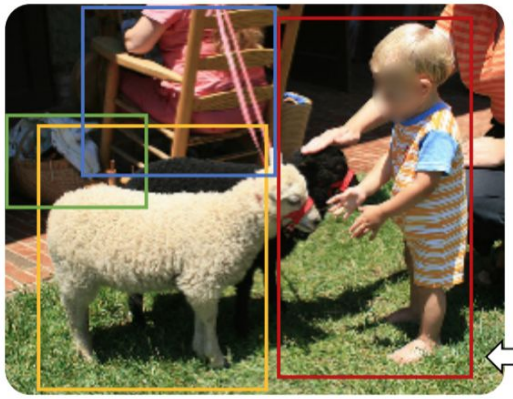
- How do we model face-to-face communication?
- Rahman et al. (ACL 2020) propose a new gate for Transformers to accept multimodal nonverbal data during fine-tuning

#	Spoken words + acoustic and visual behaviors	Ground Truth	MAG-XLNet	XLNet
1	"And it really just lacked what made the other movies more enjoyable." + Frustrated and disappointed tone	-1.4	-1.41	-0.9
2	"But umm I liked it." + Emphasis on tone + positive shock through sudden eyebrow raise	1.8	1.9	1.2
3	"Except their eyes are kind of like this welcome to the polar express." + tense voice + frown expression	-0.6	-0.6	0.8
4	"Straight away miley cyrus acting miley cyrus, or lack of, she had this same expression throughout the entire film" + sarcastic voice + frustrated facial expression	-1.0	-1.2	0.2



12-in-1: Multi-Task Vision and Language

- Lu et al. (CVPR 2020): representation learning by solving 12 different problems with a ViLBERT-derived model

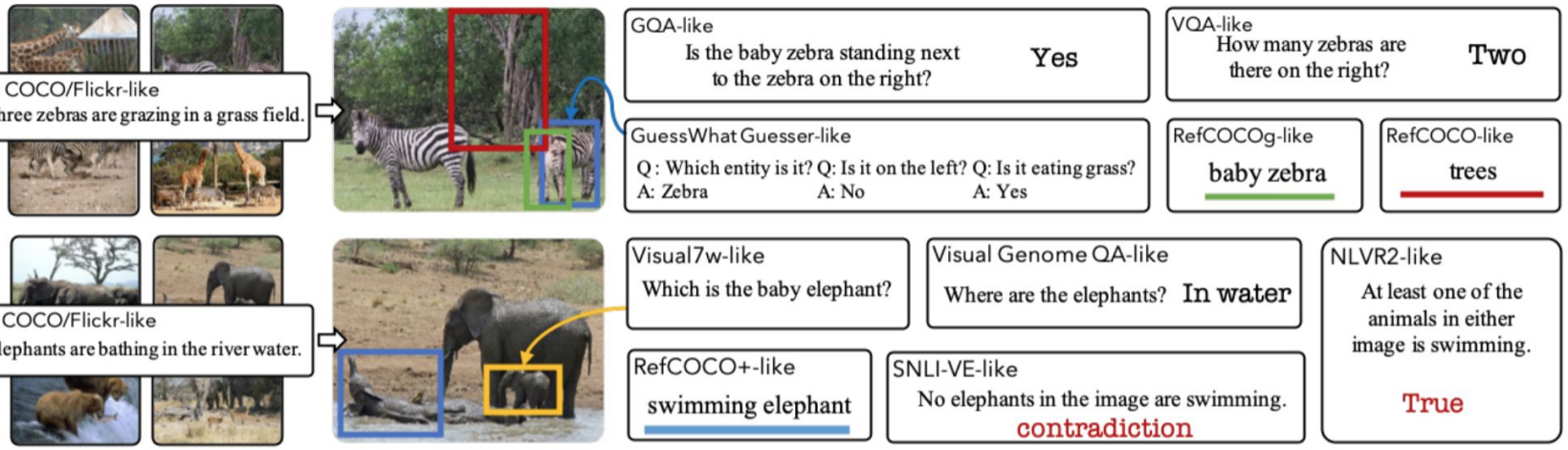


Visual Question Answering
What color is the child's outfit? Orange

Referring Expressions
child sheep basket people sitting on chair

Multi-modal Verification
The child is petting a dog. false

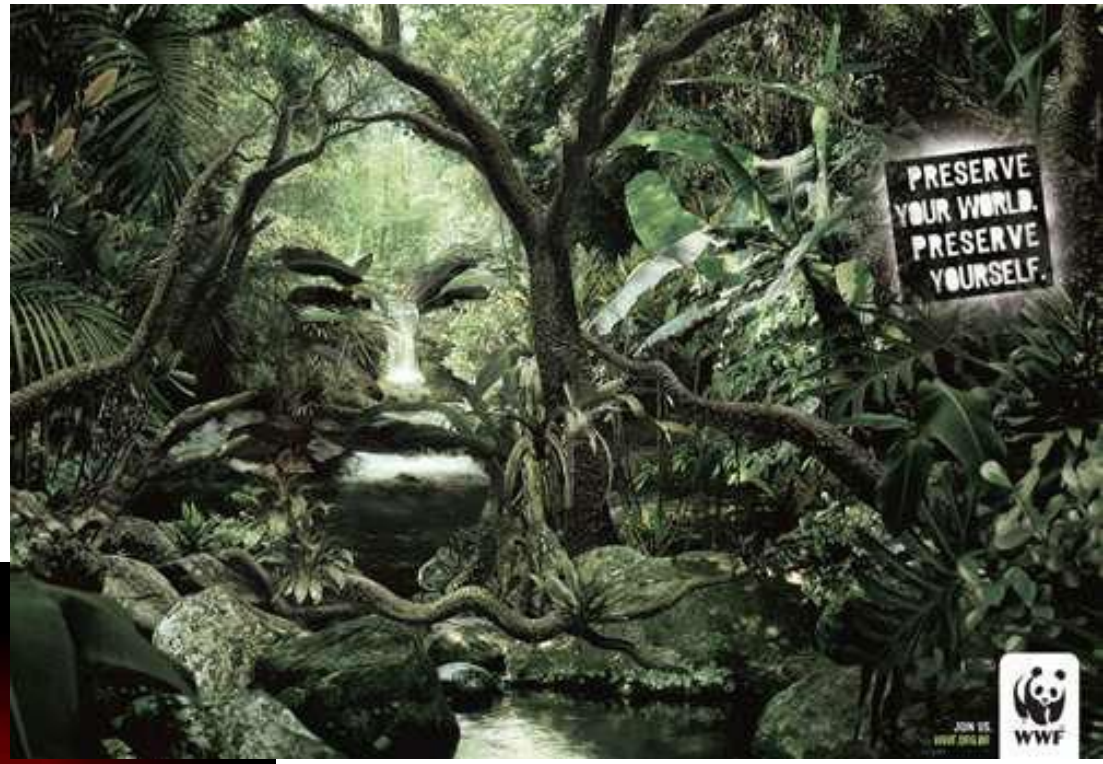
Caption-based Image Retrieval
A child in orange clothes plays with sheep.



- And now for a detailed example...

Example: symbolism in advertising

- Ads are all about symbols
- Often they are not easy to parse even for a human
- Sometimes an object resembles another



- Sometimes it's not even an object...

Example: symbolism in advertising

- But **text** really helps! What is this ad about?

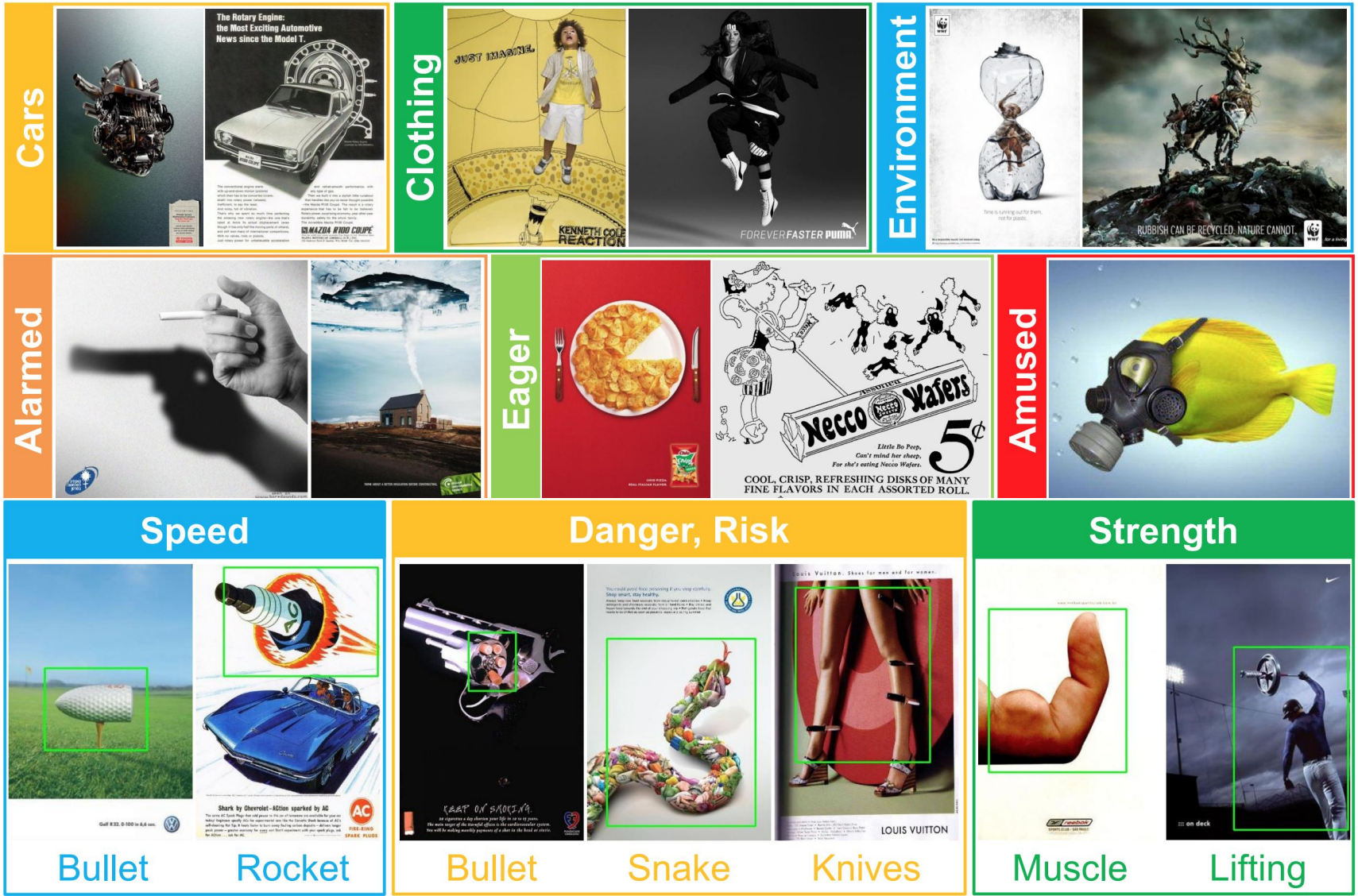
Example: symbolism in advertising

➢ But **text** really helps! What is this ad about? The text explains it all.

Example: symbolism in advertising

- In 2017, researchers from the University of Pittsburgh published a dataset on understanding ads; thousands of labeled images

Labeling: topics, sentiment, and symbols



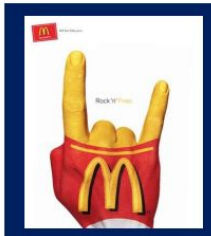
Labeling: visual understanding strategy

Straightforward/literal ads (+OCR/NLP required)

Understanding physical processes

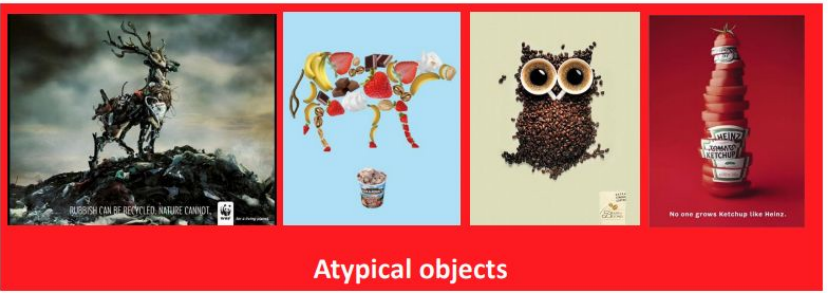
Transfer of qualities

Symbolism



Culture/memes

Humans experience product

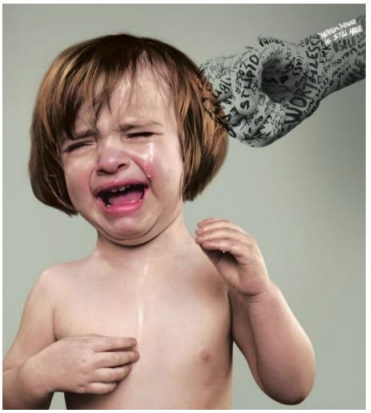


Atypical objects

Surprise/shock or humor/pun

Contrast

Labeling: Q&A



What should I do? I should be careful what words I use on my kid.

Why? Because words can hurt as much as fists.

Q: *Why should I be careful what words I use on my kid?*

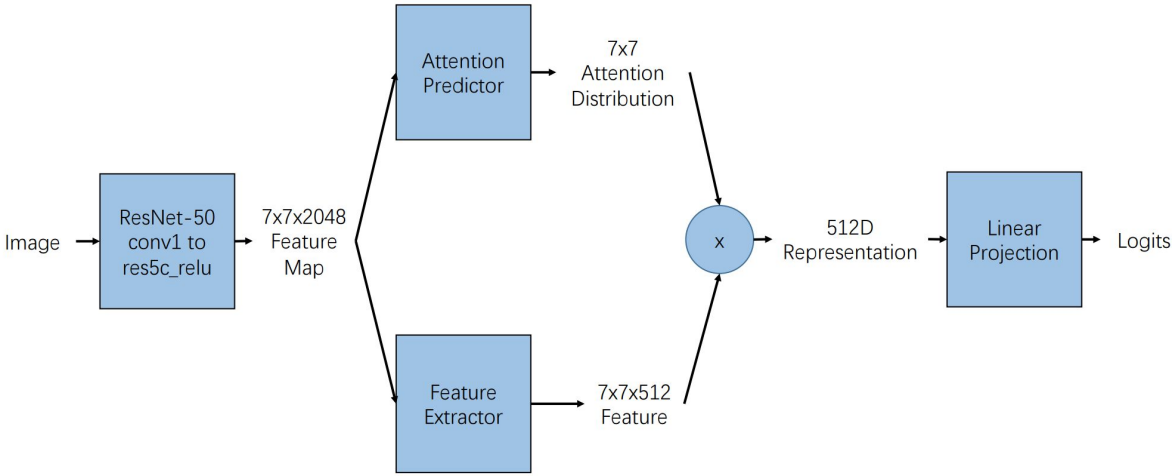
A: Because words can hurt as much as fists.

I should stop smoking because my lungs are extremely sensitive and could go up in smoke.

I should buy this candy because it is unique and rises above the rest, like the Swiss Alps.

So what did people do?

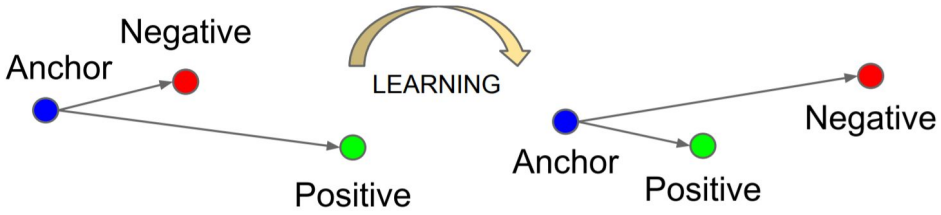
- (Hussein et al., 2018): in the original work, symbols are predicted with object detection
- But these results were obviously treated as a weak baseline



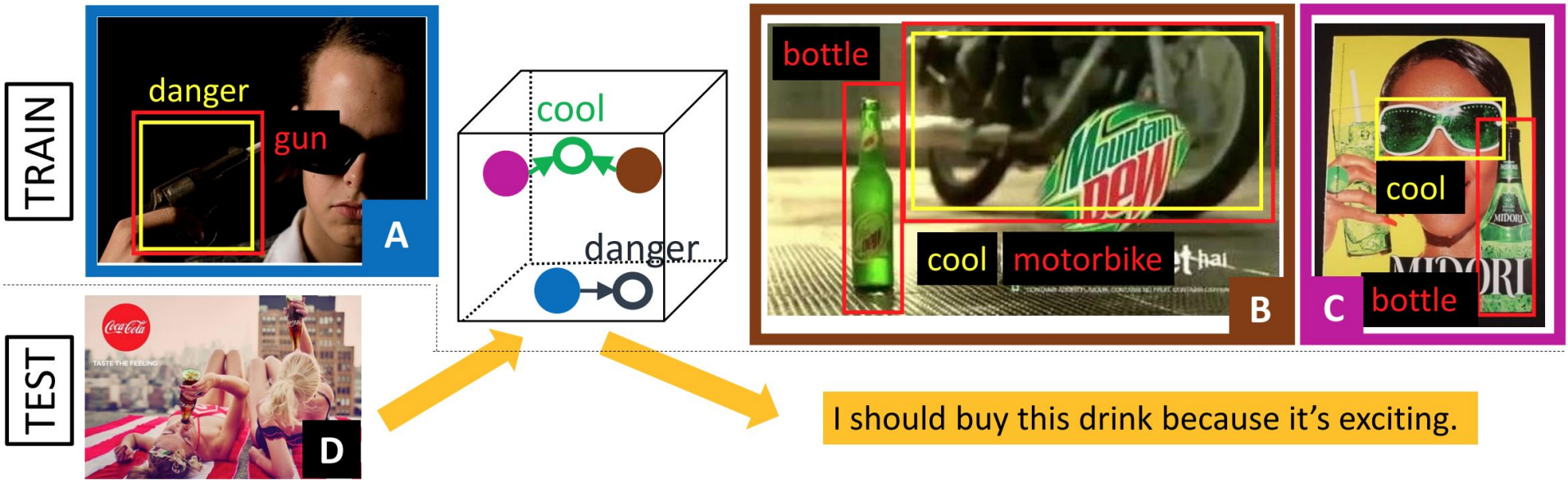
Speed (fast, quick, speed)

Symbolism detection by joint embeddings

- (Ye, Kovashka, 2018): later, joint multimodal representations were used to map symbols into a joint latent space with statements describing the ad's message

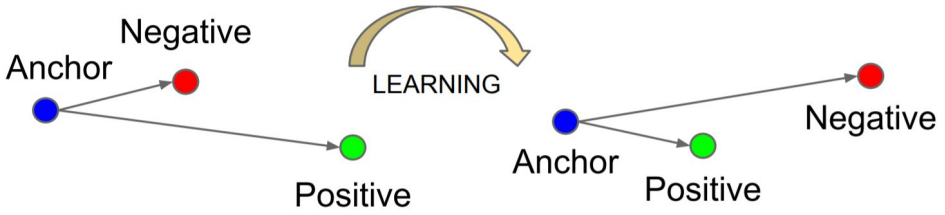


- Siamese networks with triplet loss and a knowledge base for symbols

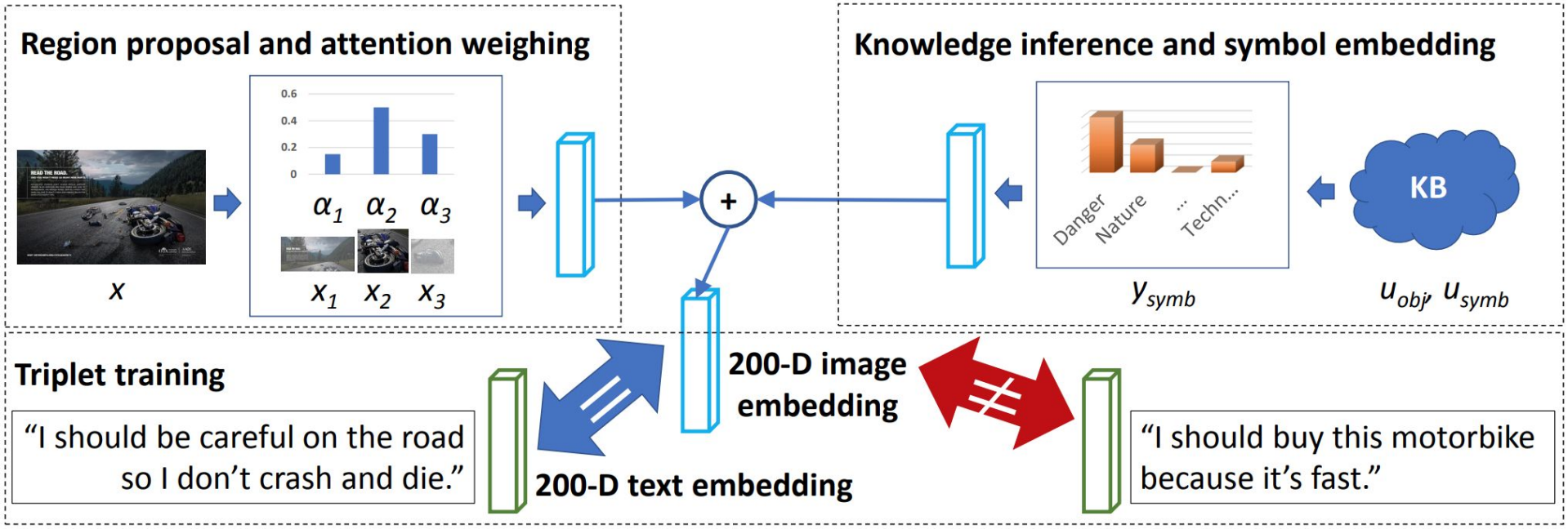


Symbolism detection by joint embeddings

- They have an object detection framework with attention, knowledge inference that checks objects corresponding to symbols, and a text embedding

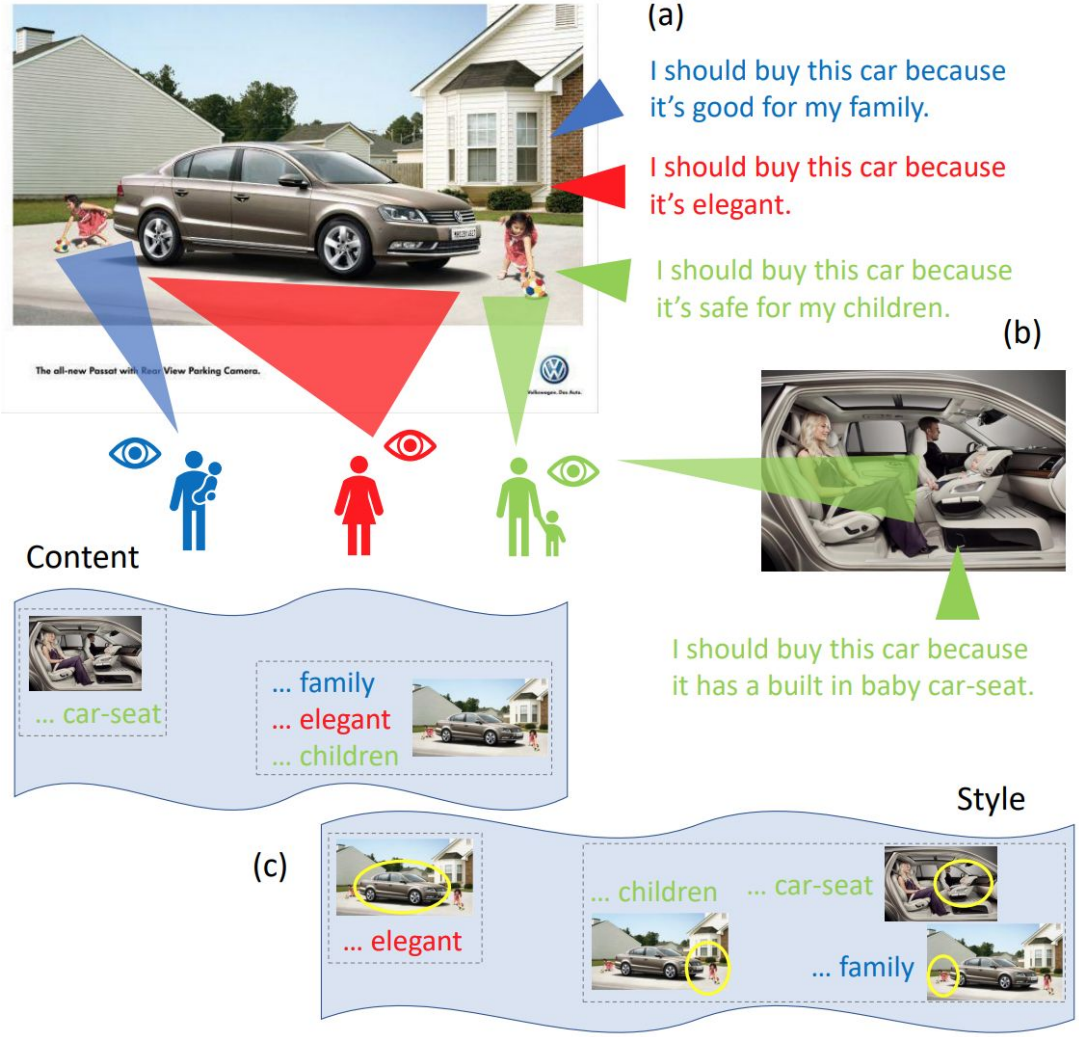


- So this work was “very multimodal” from the start



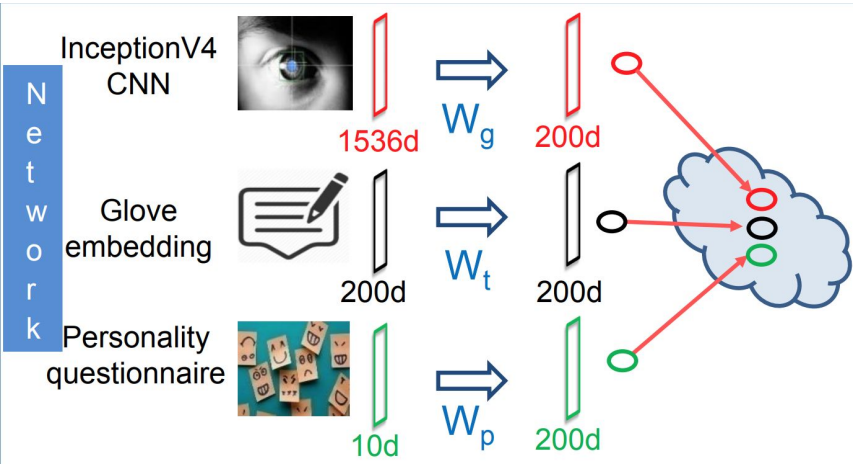
Personalized symbolism detection

- (Murrugarra-Llerena, Kovashka, 2019): let's add the viewer's personality! They collect caption-gaze samples from users

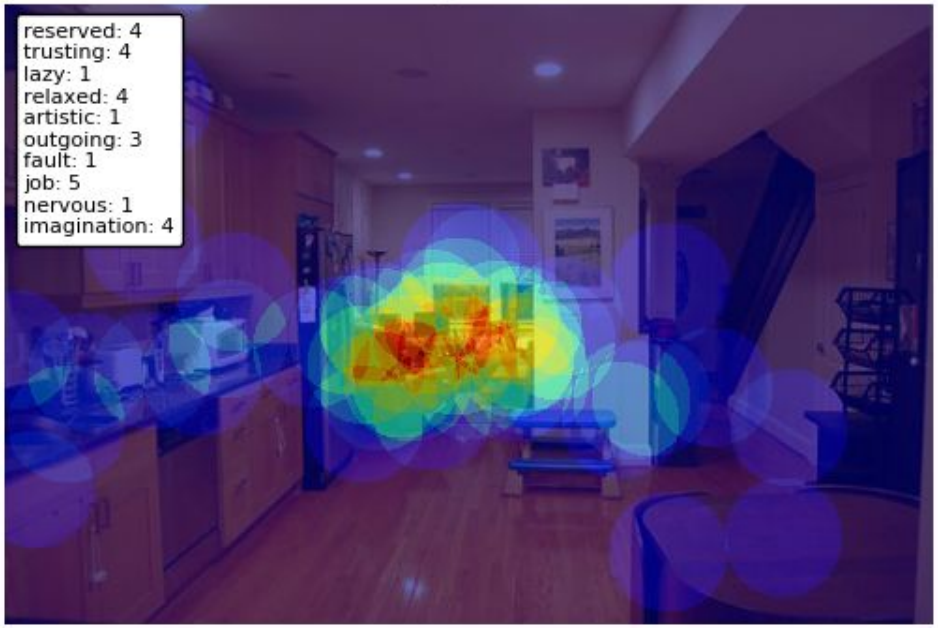


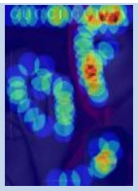
Personalized symbolism detection

- The dataset contains text annotations and the annotator's personality
- The approach is again to learn joint embeddings, this time with three modalities

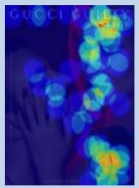


A home with a kitchen and living room with a recliner, chair and tv.





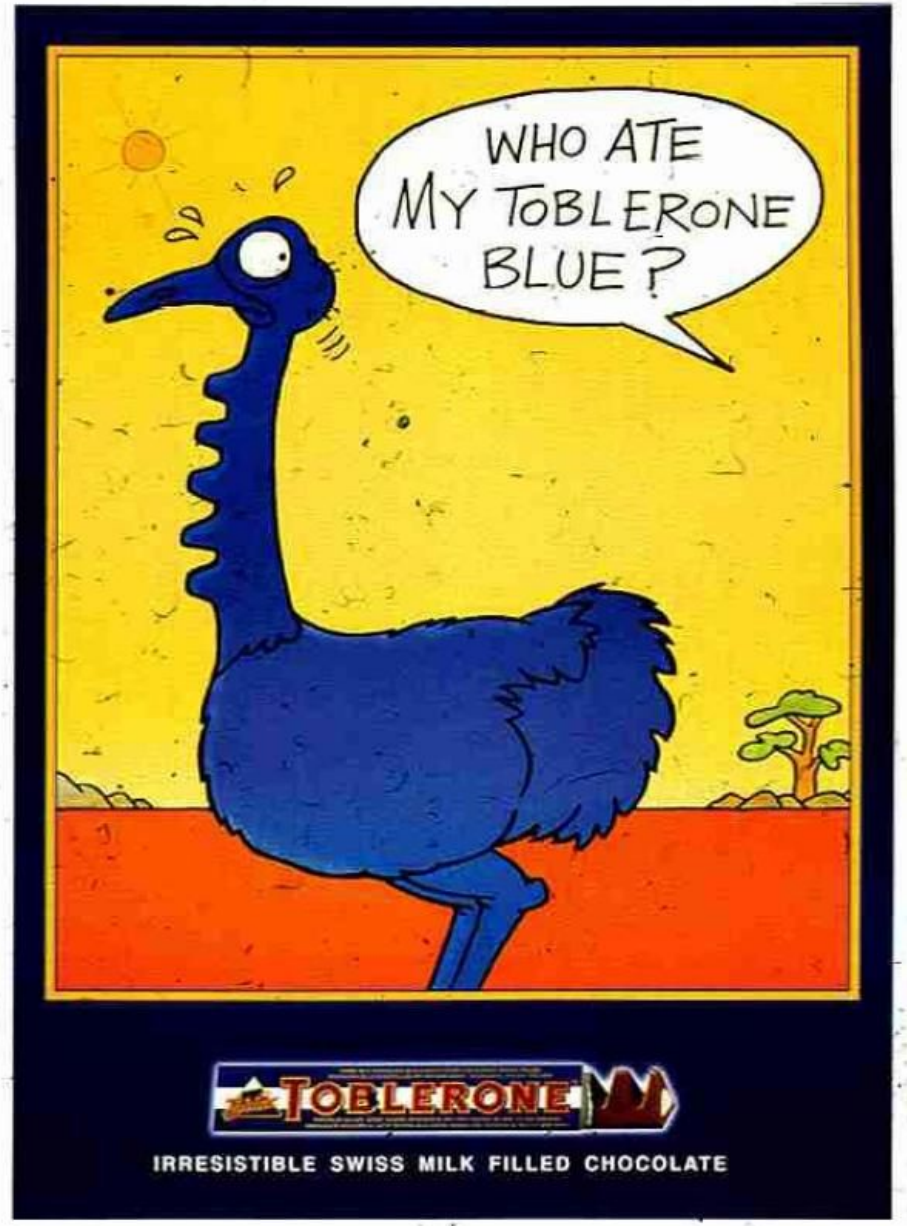
I should take more time to put myself together because it will allow me to find more **appropriate companionship.**



I should buy Gucci Guilty perfume because it will make me a **sexier person**

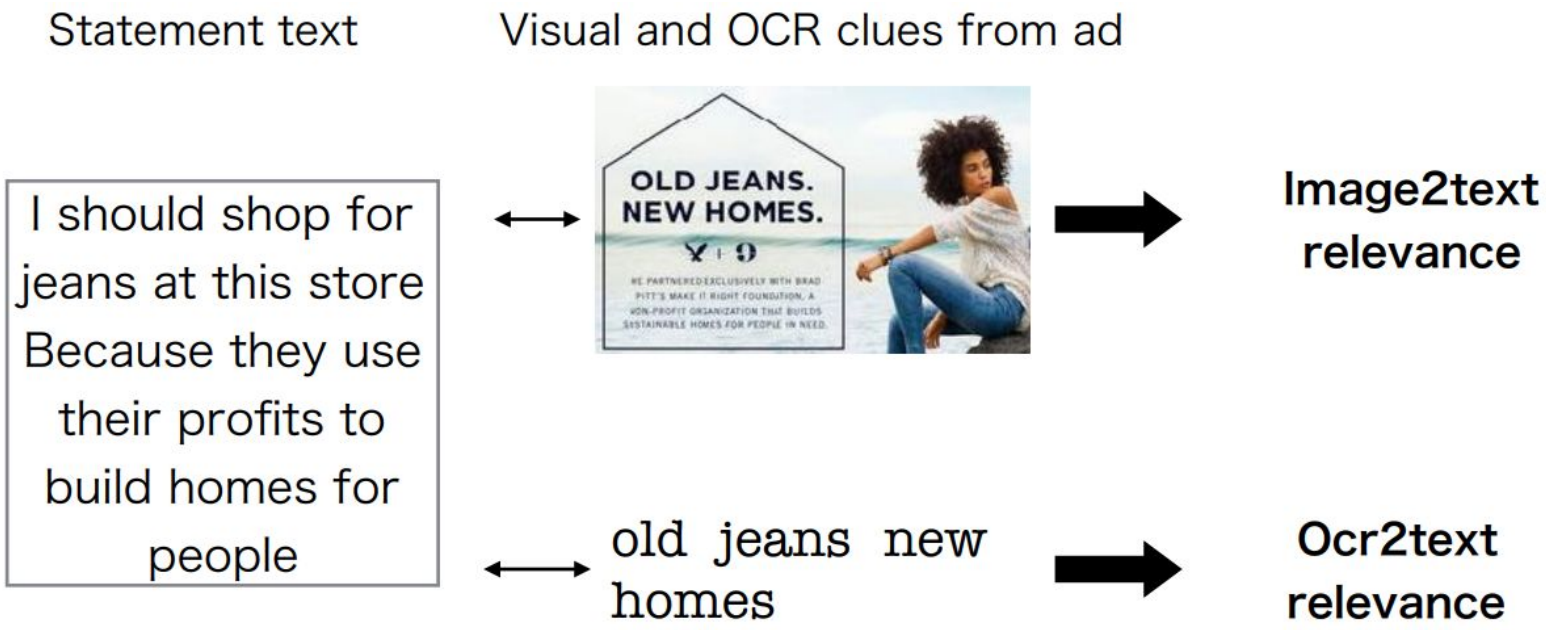
But what about the text?

- All of these works treat the ad as an image; they use NLP only to process Q&A statements
- But often an ad is completely incomprehensible without reading the text!



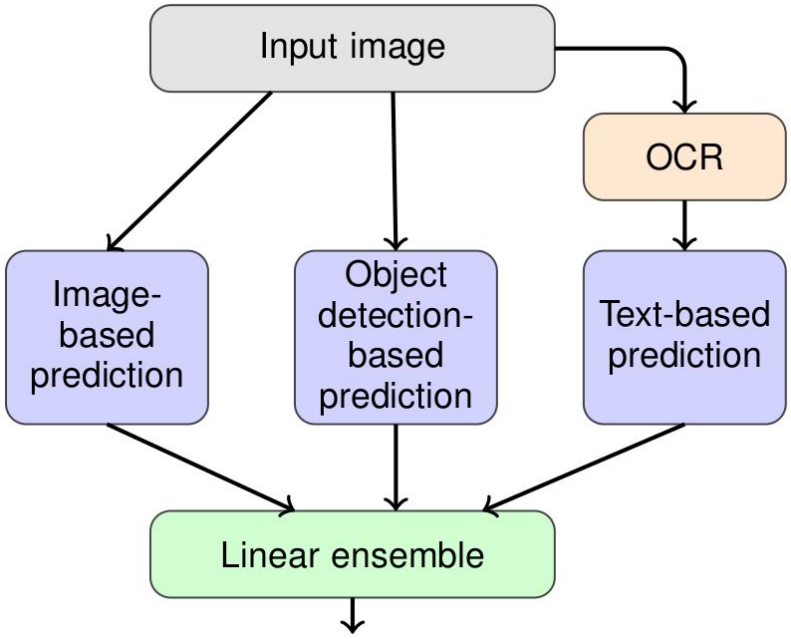
But what about the text?

- This was noted already in the original competition on this dataset, in a work by Otani et al. (2018)
- They applied it to a different problem: matching text statements to ads; their pipeline is actually rather straightforward IR



What we did

- Our approach was to combine text-based models that read the text with OCR and image processing:
  - **image-based classifier**: we already exceeded state of the art just with an EfficientNet-based classifier!
  - **object detection**: Faster R-CNN with Inception-Resnet backbone
  - **text-based models**: Transformer-based models (BERT, RoBERTa) and simpler classifiers



An aside: OCR is hard!

- By the way, OCR is hard! (later we had some more luck with CharNet)

Tesseract (Smith, 2007)	Bite the boredom, TCT am 5 bite-sized treats with a crunchy outside PUR RCo e eB Pree ee Rar ed oe 2 ; RN good
EAST + *Tesseract* (Kopeykina and Savchenko, 2019)	Bite dale boredom, unleash the fun! bite sized treats rh Pe tah outside and delicious aay filled Sita bo ii te STi) me KFc macaroni and cheese. Sl PPPS m)/7-) UC So good
Google Android Vision API	FC ma heese A Bite the boredom, unleash the fun! 5 bite-sized treats with a cr good chy and a delicious center filled with ni and creamy cheese. soft
CloudVision (Otani et al., 2018)	Bite the boredom, unleash the fun! 5 bite-sized treats with a crunchy outside and a delicious center filled with soft macaroni and creamy cheese. KFC ma and heese good KFC
AR-Net Caption	a poster for the film

Our results

- By reading the text and linear blending, we improve over state of the art results; even OCR quality is not as crucial as it might seem

OCR	Model	Dataset: 221 labels Image-only results: **0.1967**			Dataset: 53 labels Image-only results: **0.2814**		
		Text-based (w/text)	Ensemble (w/text)	Ensemble + Backoff (all)	Text-based (w/text)	Ensemble (w/text)	Ensemble + Backoff (all)
Tesseract	BoN	**0.0881**	**0.1996**	**0.2002**	0.1345	0.2889	0.2865
	BERT	0.0220	0.1932	0.1967	**0.1794**	**0.2935**	**0.2892**
	RoBERTa	0.0220	0.1934	0.1967	0.1765	0.2918	0.2882
EAST + Tesseract	BoN	**0.1198**	**0.2087**	**0.2122**	0.1457	0.2955	0.2957
	BERT	0.0689	0.1949	0.1989	0.1933	**0.3050**	**0.3046**
	RoBERTa	0.0225	0.1975	0.2013	**0.1994**	0.2914	0.2919
AR-Net Captions	BoN	0.1014	0.1918		0.1586	0.2862	
	BERT	**0.1089**	0.1975		0.1615	0.2834	
	RoBERTa	0.0226	**0.2012**		**0.1733**	**0.2871**	
Google Android Vision API	BoN	**0.1407**	**0.2102**	**0.2106**	0.2076	0.3026	0.2994
	BERT	0.0971	0.1975	0.2000	0.2278	0.3054	0.3017
	RoBERTa	0.1201	0.1992	0.2014	**0.2409**	**0.3189**	**0.3128**
CloudVision (Otani et al., 2018)	BoN	**0.1830**	**0.2310**	**0.2292**	0.2420	**0.3186**	**0.3177**
	BERT	0.1520	0.2014	0.2023	0.2653	0.3048	0.3051
	RoBERTa	0.1580	0.2006	0.2016	**0.2898**	0.3074	0.3075

OCR	CONVERSE There's no lace like home	LIZER BIKE HELMETS	Put your shirt and join our team WWF
Text-based	sex;love;fashion	danger;safety;protection	nature;environment
Image-based	nature;fun;love;sports	danger;violence;death;love;injury	nature;danger;death;environment
Blend	None	danger;violence;death;safety	nature
Ground truth	fun;safety;comfort;humor	injury;safety	environment

Takeaway point from this example

- **Key takeaway**: a simple straightforward approach can be more useful than complex models (in this case, multimodal joint embeddings)
- This paper was published at COLING 2020

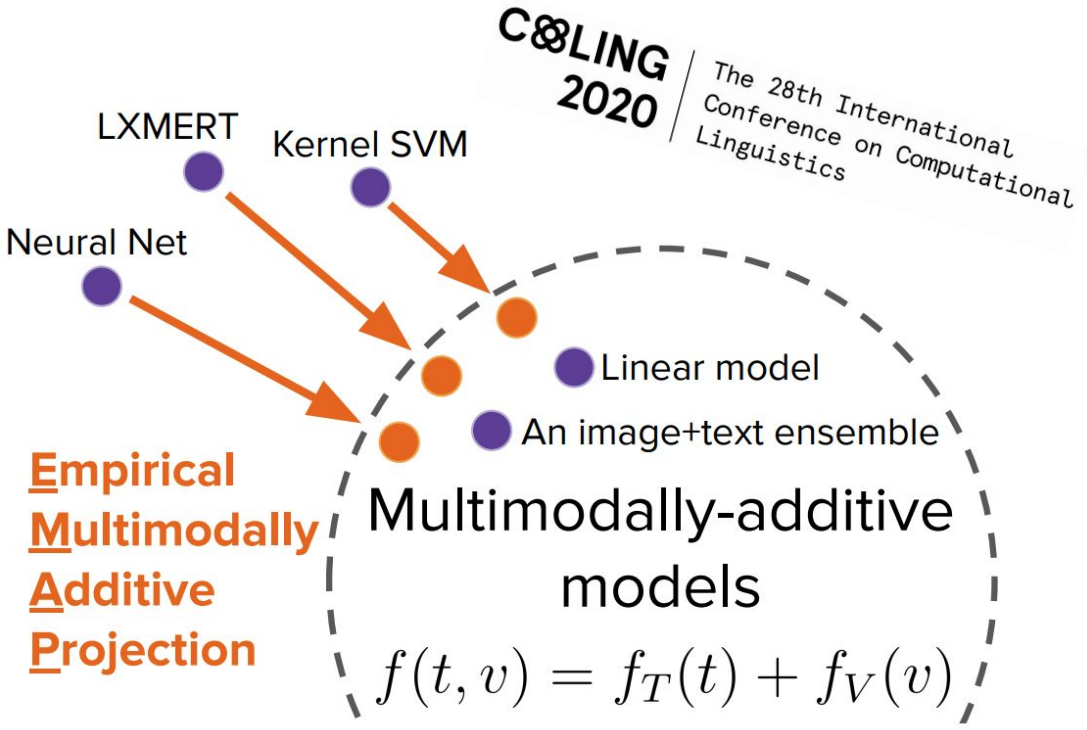
Ad Lingua: Text Classification Improves Symbolism Prediction in Image Advertisements

Andrey Savchenko[1,3], Anton Alekseev[1], Sejeong Kwon[2], Elena Tutubalina[1], Evgeniy Miasnikov[1] and Sergey Nikolenko[1,4]

- Interestingly, at the same conference we saw...

Empirically Multimodally Additive Projection

- (Hessel, Lee, 2020): "Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!"
- They try to approximate multimodal models with **linear** image/text ensembles, removing cross-modal interactions
- And they find very little or no loss in performance!



So what's next? Let's find out together!

THANK YOU
FOR YOUR ATTENTION!