深度学习与自然语言及视觉智能

Deep Learning for Language & Vision Intelligence

何晓冬 (Xiaodong He)

美国微软研究院　首席研究员
华盛顿大学　兼职教授
Principal Researcher　Deep Learning, Microsoft Research, Redmond, WA
Affiliate Professor　Electrical Engineering, University of Washington, Seattle, WA

Tutorial at CCL 2017, Nanjing, China

Tutorial Outline

Semantic Learning in Natural Language Processing

- Deep Structured Semantic Models (DSSM, a.k.a. sent2vec)
- Information Retrieval, Recommendation, Knowledge Graph, Question Answering

Multimodal Intelligence across Language & Vision

- Image-to-language Captioning
- Visual Question Answering
- Language-to-Image Synthesis

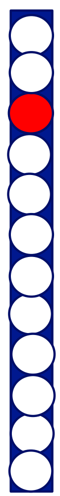
Continuous Word Representations

Project a word into a continuous space

e.g., word embedding

Captures the word meaning in a semantic space

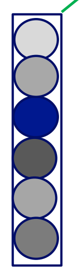
$f(\boldsymbol{cat}) =$



The index of "cat" in the vocabulary

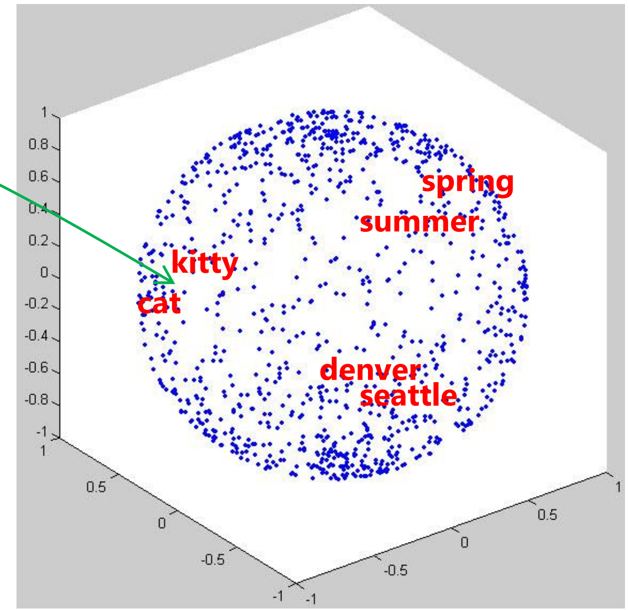
Dim=|V|=100K~100M

a.k.a the 1-hot word vector



Dim=100~1000

word embedding vector in the semantic space

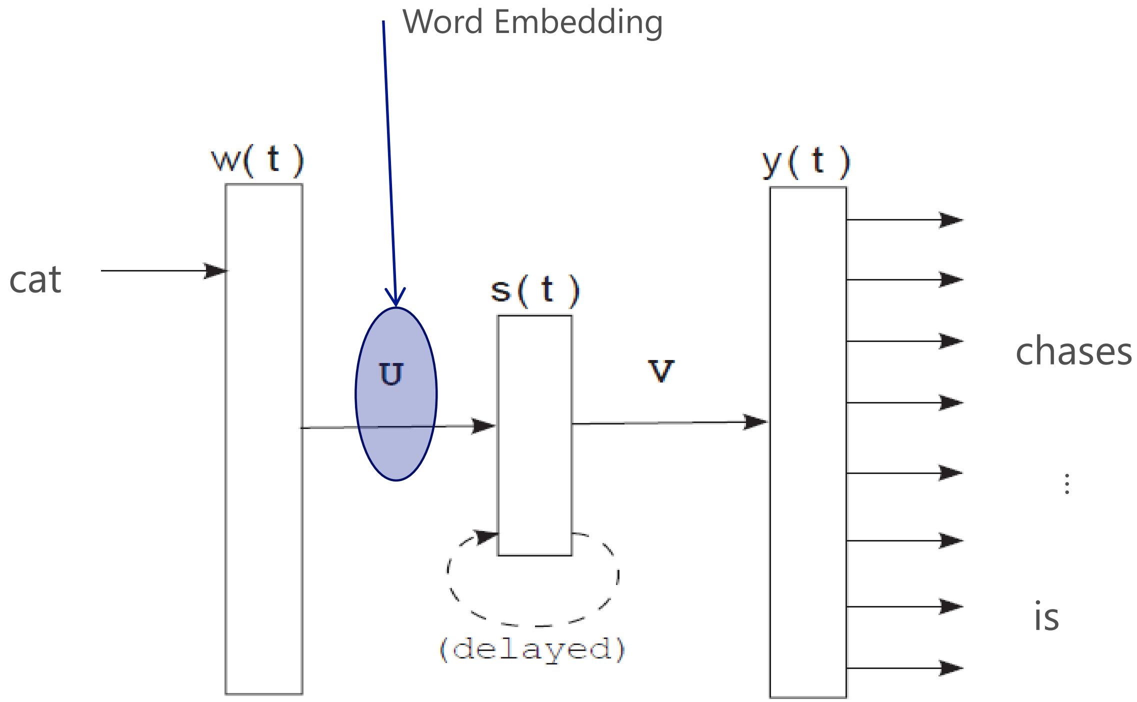


Deerwester, Dumais, Furnas, Landauer, Harshman, "Indexing by latent semantic analysis," JASIS 1990

Continuous Word Representations

- A lot of popular methods for creating word vectors!
  - Vector Space Model [Salton & McGill 83]
  - Latent Semantic Analysis [Deerwester+ 90]
  - Brown Clustering [Brown+ 92]
  - Latent Dirichlet Allocation [Blei+ 01]
  - Deep Neural Networks [Collobert & Weston 08]
  - Word2Vec [Mikolov+ 13]
  - GloVe [Pennington+ 14]
- Encode term co-occurrence information
- Measure semantic similarity well

RNN-LM Word Embedding



Mikolov, Yih, Zweig, "Linguistic Regularities in Continuous Space Word Representations," NAACL 2013

SENNA Word Embedding

Scoring:

$$Score(w_1, w_2, w_3, w_4, w_5) = U^T \sigma(W[f_1, f_2, f_3, f_4, f_5] + b)$$

Training:

$$J = \max(0, 1 + S^- - S^+)$$

Update the model until $S^+ > 1 + S^-$

Where

$$S^+ = Score(w_1, w_2, w_3, w_4, w_5)$$
$$S^- = Score(w_1, w_2, w^-, w_4, w_5)$$

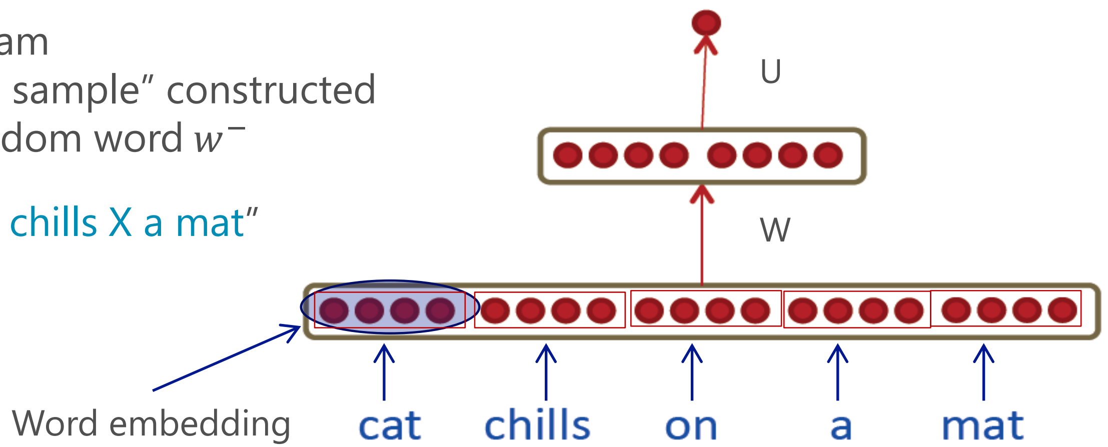
And

$< w_1, w_2, w_3, w_4, w_5 >$ is a valid 5-gram

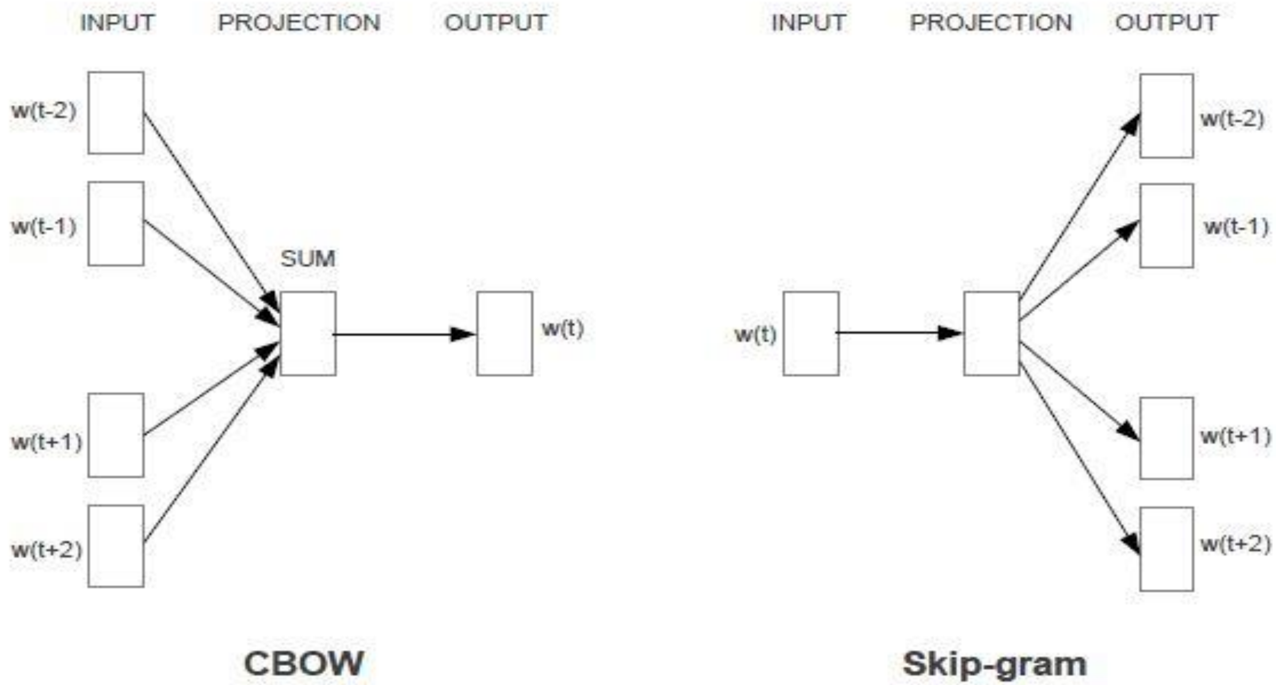
$< w_1, w_2, w^-, w_4, w_5 >$ is a "negative sample" constructed by replacing the word w_3 with a random word w^-

e.g., a negative example: "cat chills X a mat"

Collobert, Weston, Bottou, Karlen, Kavukcuoglu, Kuksa, "Natural Language Processing (Almost) from Scratch," JMLR 2011



CBOW/Skip-gram Word Embeddings



Continuous Bag-of-Words

The CBOW architecture (a) on the left, and the Skip-gram architecture (b) on the right. [Mikolov et al., 2013 ICLR].

GloVe: Global Vectors for Word Representation [Pennington+ EMNLP-14]

- Semantic relatedness can be observed from word co-occurrence counts and ratios

Context words

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	$\boxed{8.9}$	8.5×10^{-2}	1.36	0.96

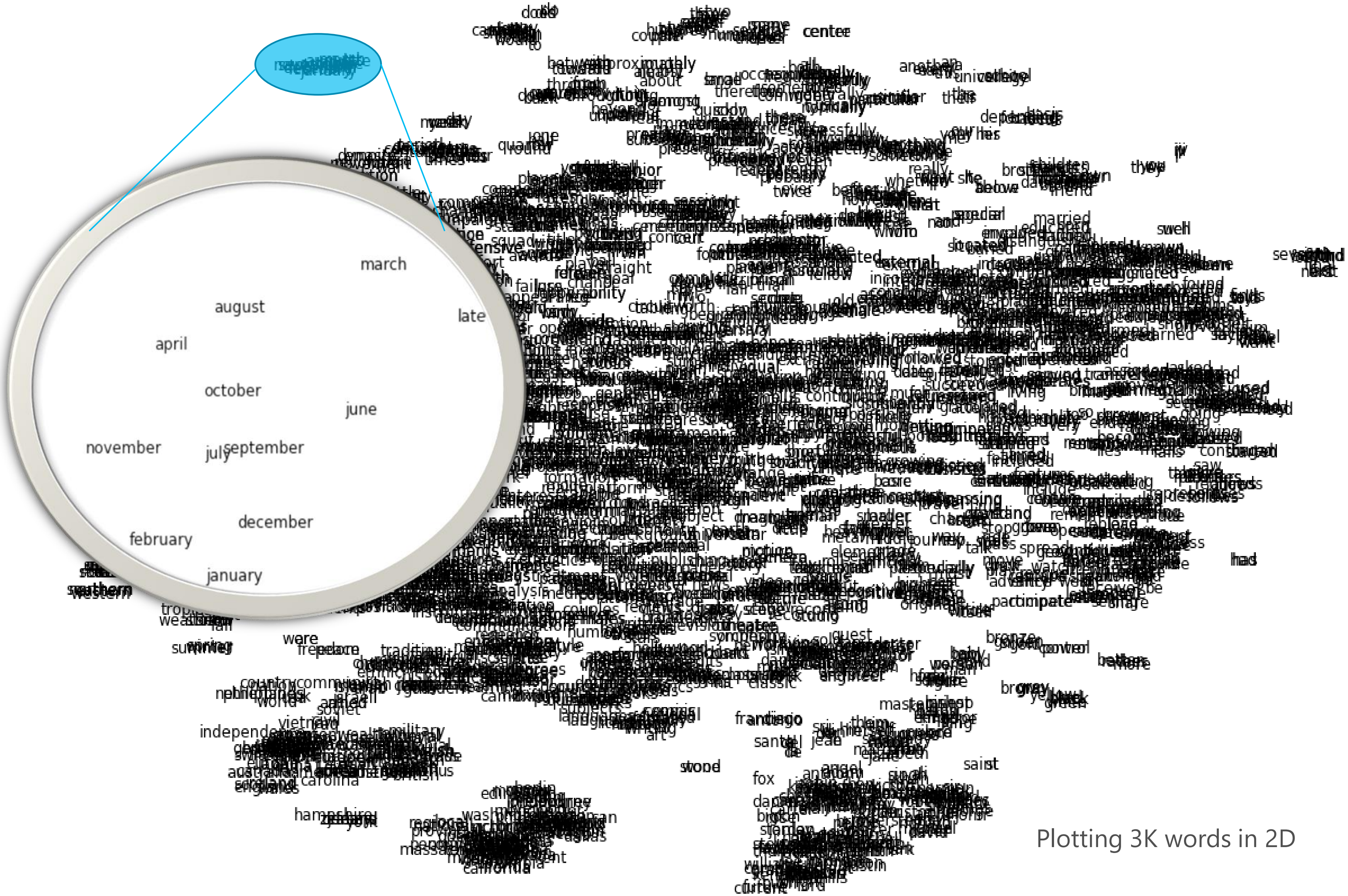
"solid" is more related to "ice"

GloVe: Global Vectors for Word Representation [Pennington+ EMNLP-14]

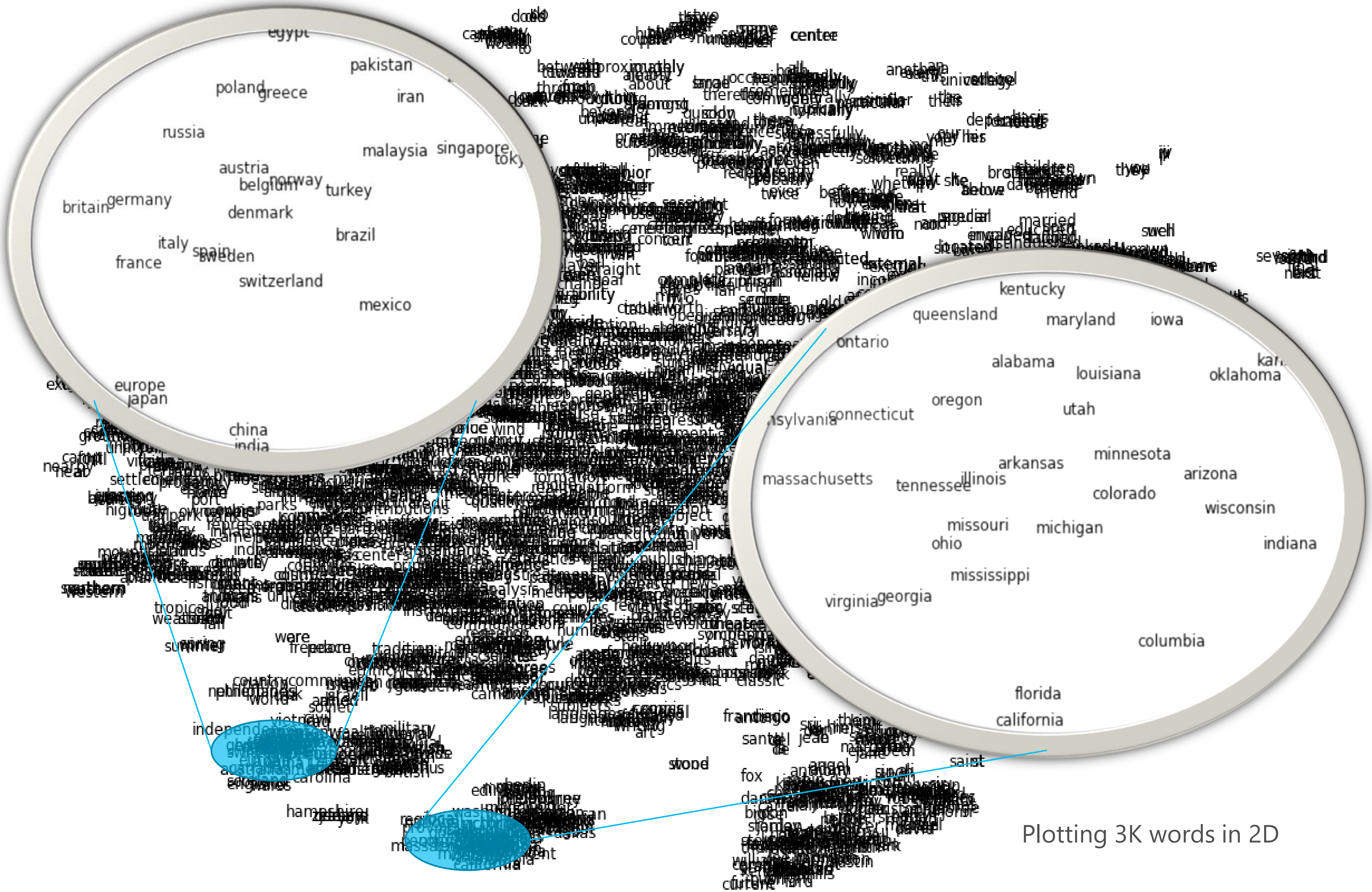
- Word embedding model design principle:

$F(w_i, w_j, \widetilde{w}_k) = \frac{P(k|i)}{P(k|j)}$ (e.g., i = ice, j = steam, k = solid/gas)

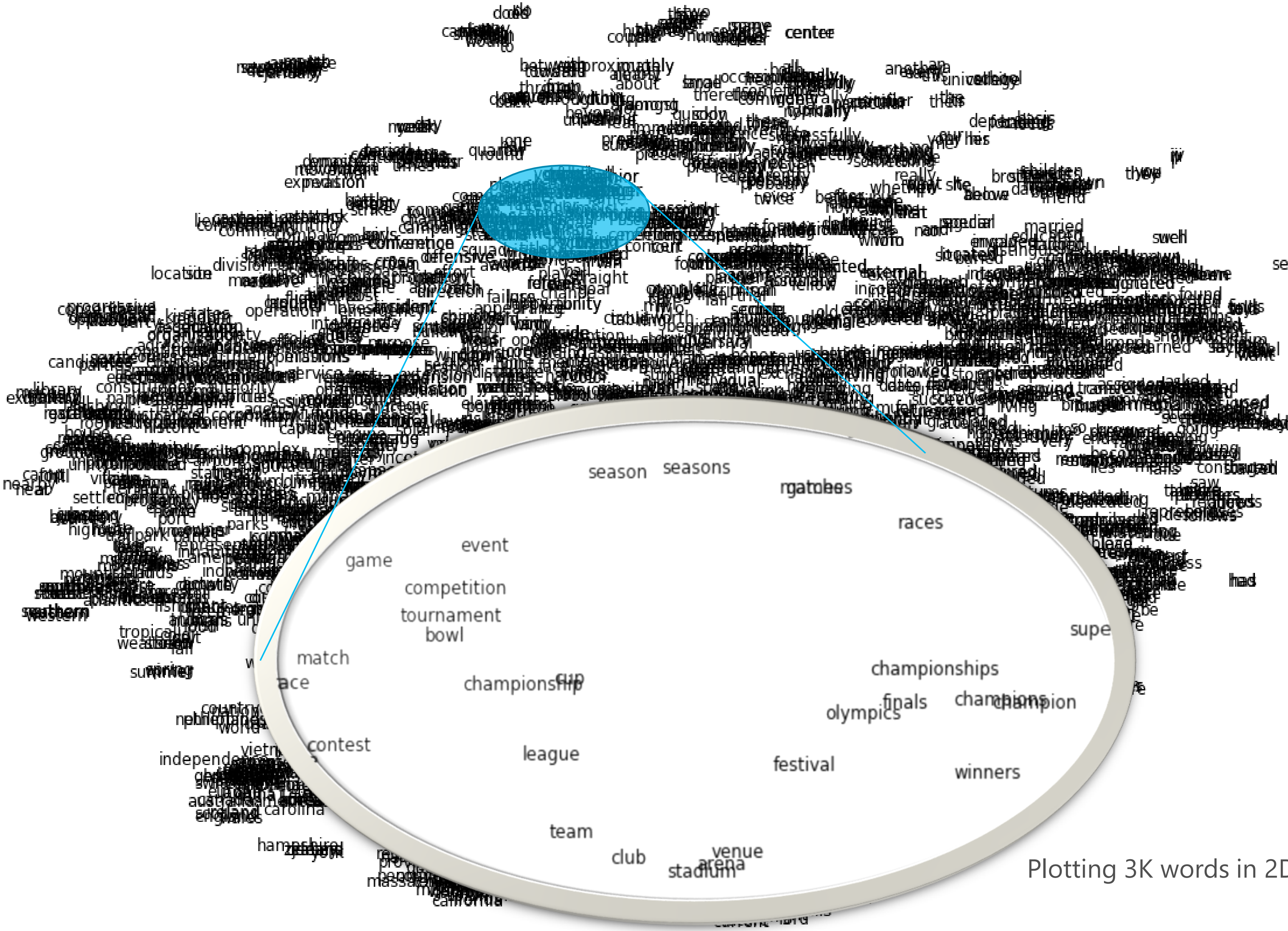
- Objective: $$J = \sum_{i,j=1}^{V} \underbrace{f\left(X_{ij}\right)}_{\text{Down weight low co-occurrences}} \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log \underbrace{X_{ij}}_{\text{co-occurrence counts}}\right)^2$$



Plotting 3K words in 2D



Plotting 3K words in 2D

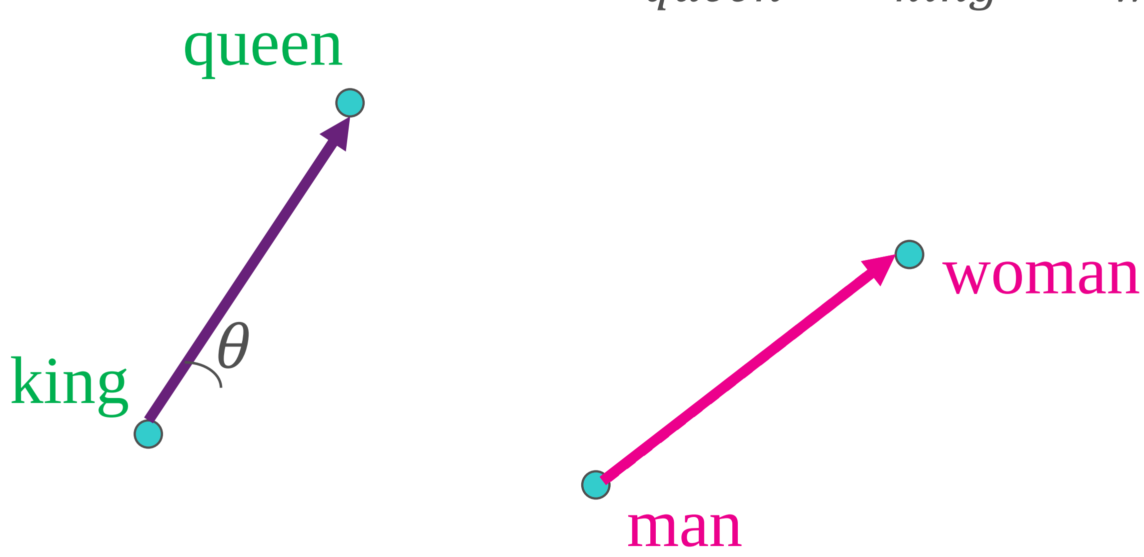


Plotting 3K words in 2D

Unexpected Finding: Directional Similarity

- Word embedding taken from recurrent neural network language model (RNN-LM) [Mikolov+ 2011]

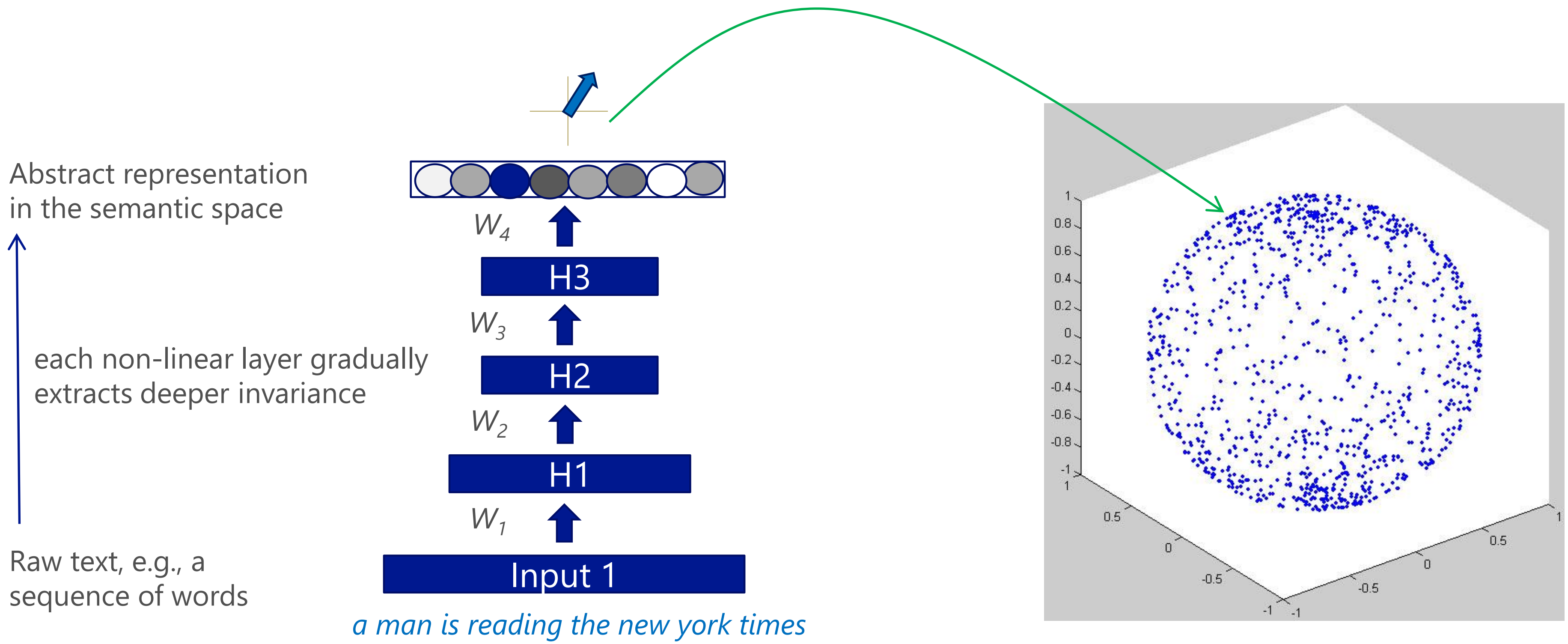
$$V_{queen} - V_{king} \cong V_{woman} - V_{man}$$



- Relational similarity is derived by the cosine score

Semantic representations for **sentences**

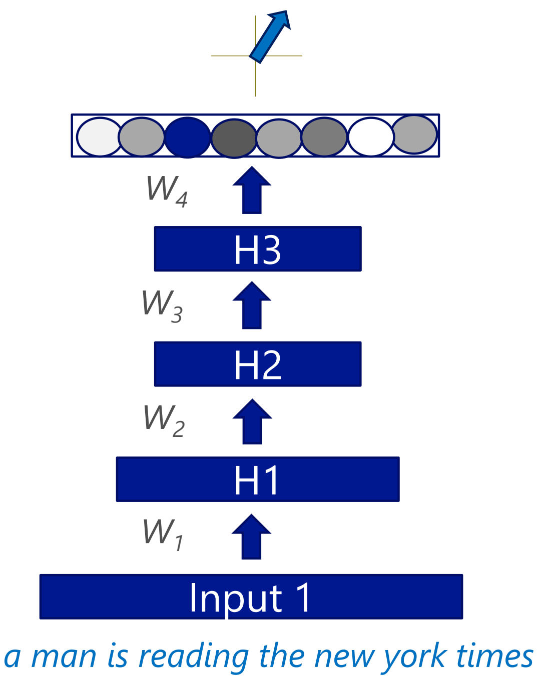
e.g., from a raw sentence to an abstract semantic vector (Sent2Vec)



Sent2Vec is crucial in many NLP tasks

Tasks	***Source***	***Target***
Web search	*search query*	*web documents*
Ad selection	*search query*	*ad keywords*
Contextual entity ranking	*mention (highlighted)*	*entities*
Online recommendation	*doc in reading*	*interesting things / other docs*
Machine translation	*phrases in language S*	*phrases in language T*
Knowledge-base construction	*entity*	*entity*
Question answering	*pattern | mention*	*relation | entity*
Personalized recommendation	*user*	*app, movie, etc.*
Image search	*query*	*image*
Image captioning	*image*	*text*
…		

The supervision problem:



However

- the semantic meaning of texts – to be learned – is latent
- no clear target for the model to learn
- How to do back-propagation?

Fortunately

- we usually know if two texts are "similar" or not.
- That's the signal for semantic representation learning.

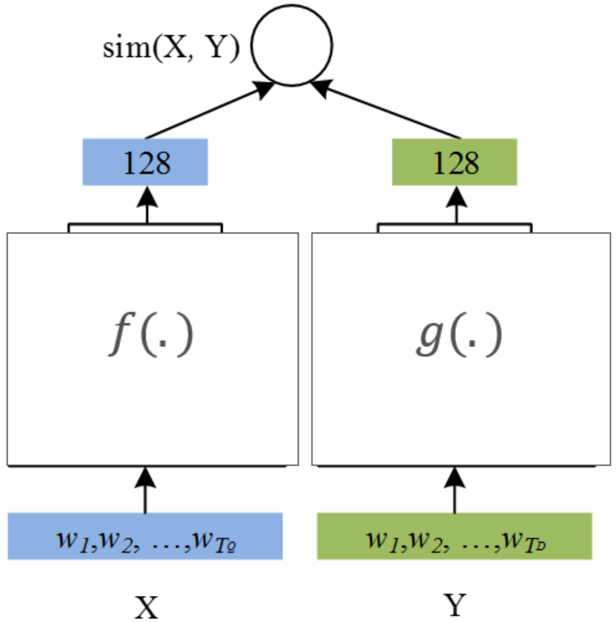
Deep Structured Semantic Model

Deep Structured Semantic Model/Deep Semantic Similarity Model (**DSSM**)

project the whole sentence to a continuous semantic space – e.g., *Sentence to Vector*.

The DSSM is built upon **characters** (rather than words) for scalability and generalizability

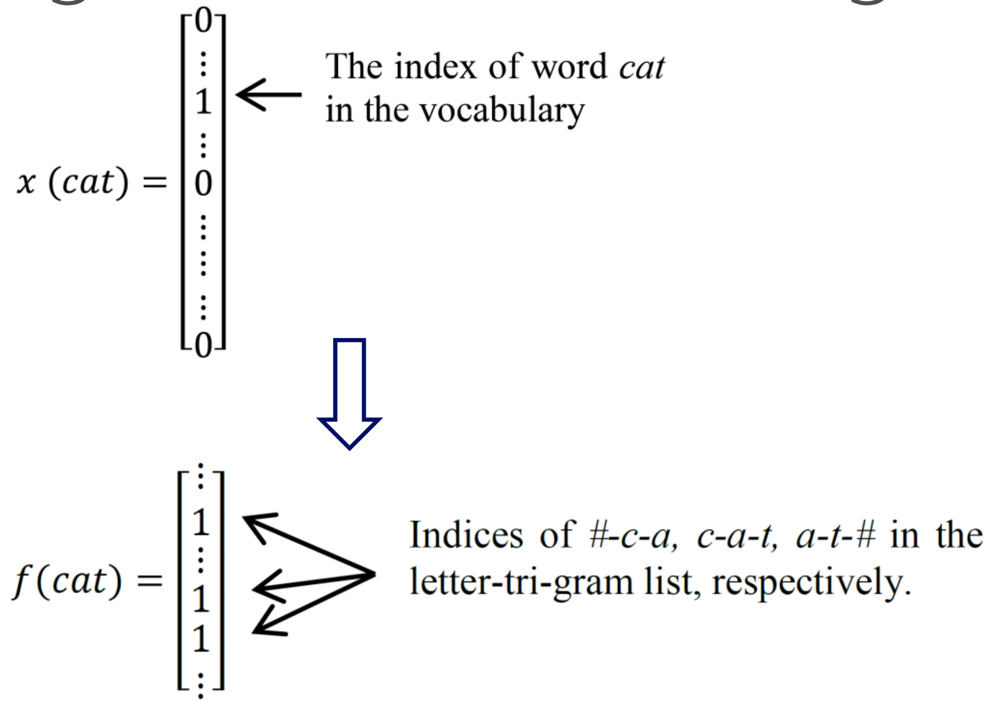
The DSSM is trained by optimizing an **similarity-driven** objective



Huang, He, Gao, Deng, Acero, Heck, "Learning deep structured semantic models for web search using clickthrough data," CIKM, October, 2013

Character-level coding (a.k.a. word hashing)

- E.g., character-trigram based *Word Hashing* of "cat"
  - -> #cat#
  - Tri-characters: #-c-a, c-a-t, a-t-#.
- Compact representation
  - |Voc| (500K) → |Char-trigram| (30K)
- Generalize to unseen words
- Robust to misspelling, inflection, etc.

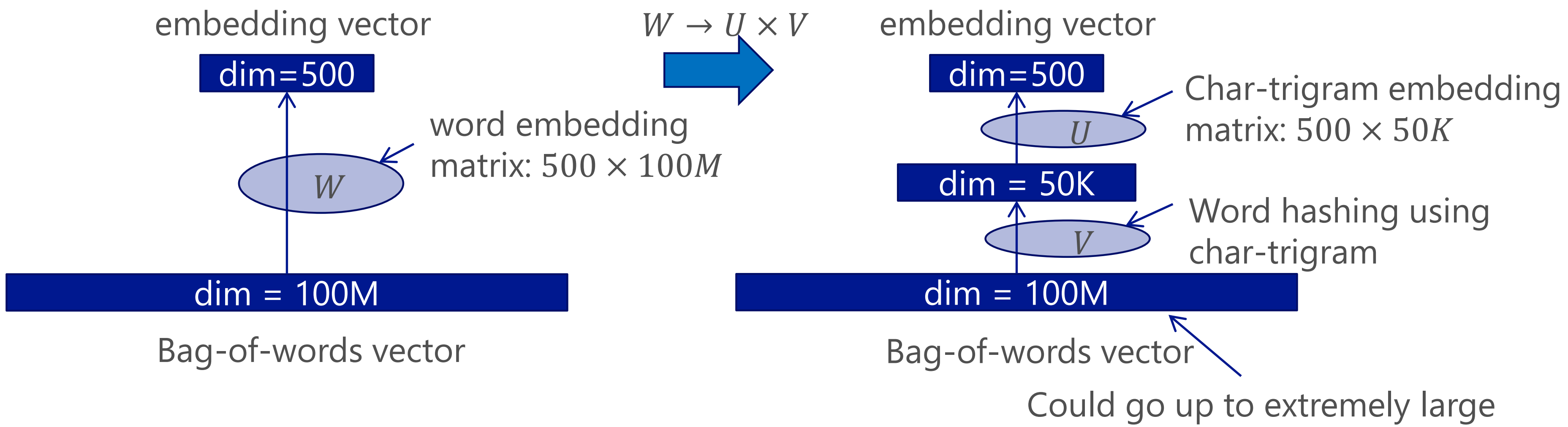


What if different words have the same word hashing code (collision)?

Vocabulary size	Unique letter-tg observed in voc	Number of Collisions
40K	10306	2 (0.005%)
500K	30621	22 (0.004%)

DSSM: built at the character-level

Decompose *any* word into set of context-dependent characters



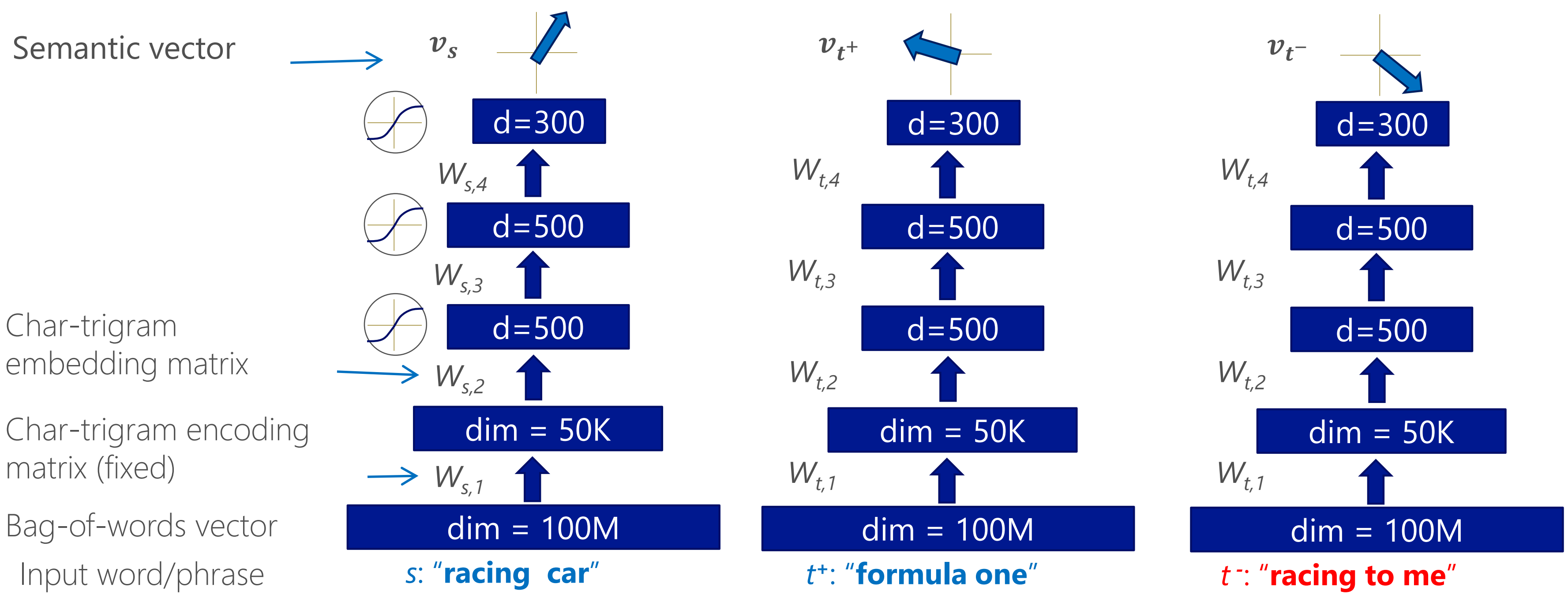
Preferable for large scale NL tasks

- Arbitrary size of vocabulary (*scalability*)
- Misspellings, word fragments, new words, etc. (*generalizability*)

DSSM: a similarity-driven Sent2Vec model

Initialization:

Neural networks are initialized with random weights



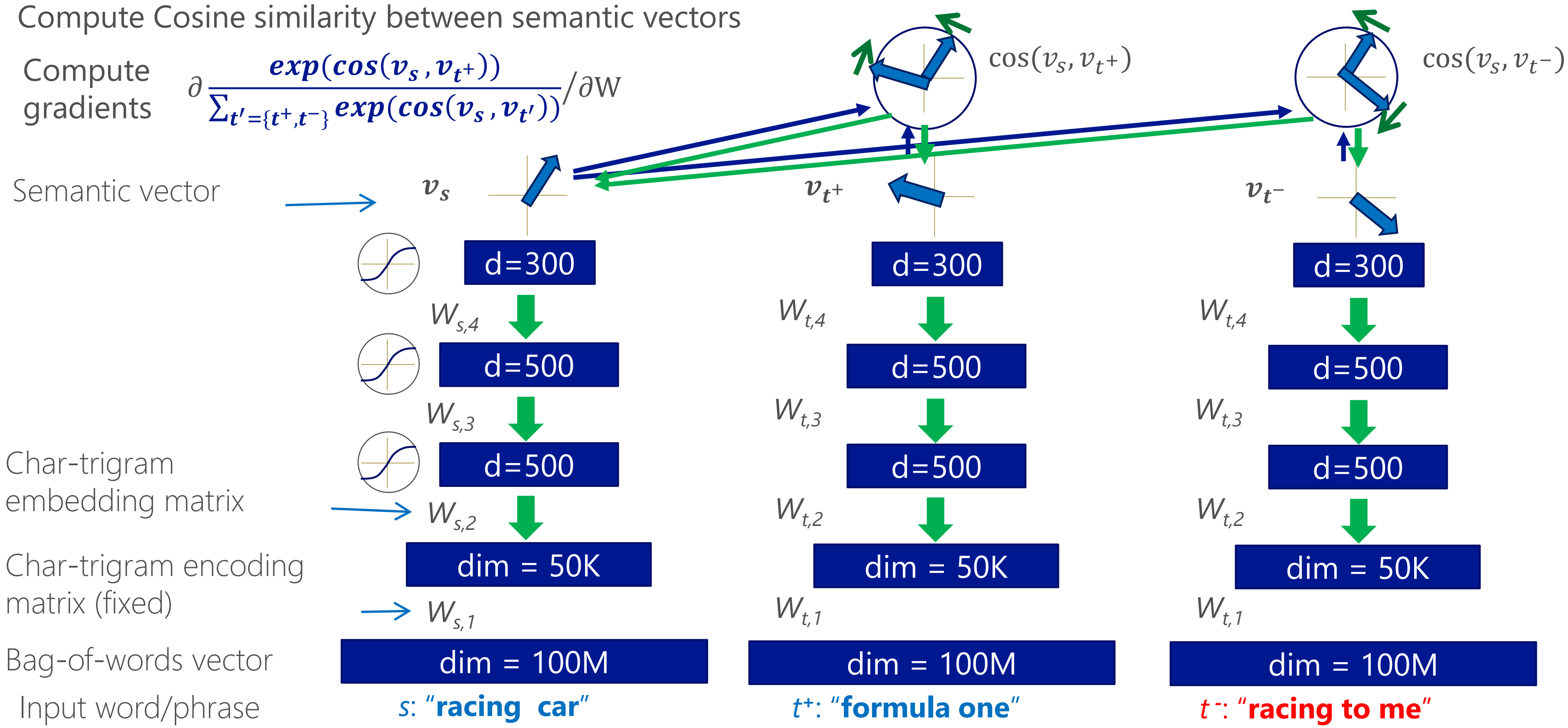
[Huang, He, Gao, Deng, Acero, Heck, "Learning DSSM for web search using clickthrough data," CIKM, 2013]

DSSM: a similarity-driven Sent2Vec model

Training:

Compute Cosine similarity between semantic vectors

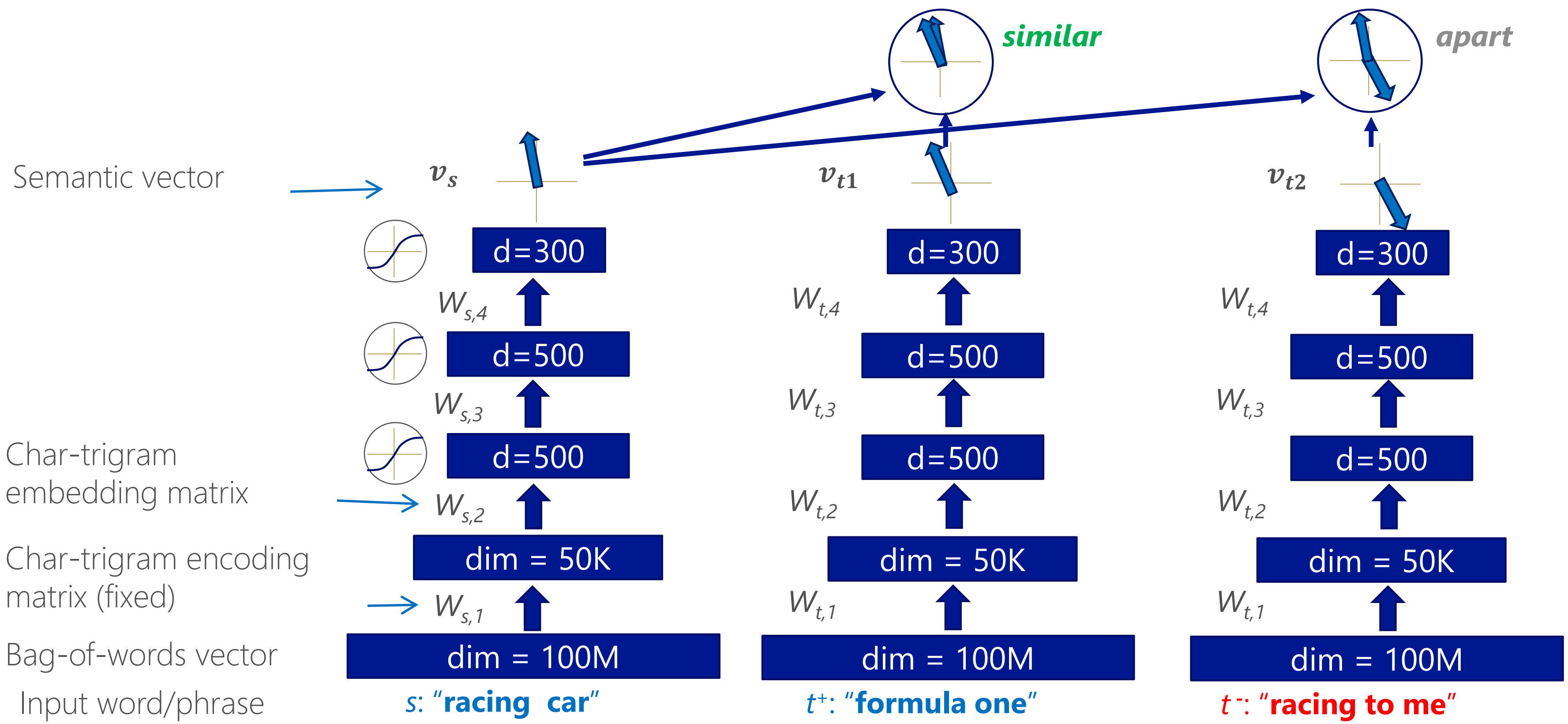
Compute gradients $\partial \dfrac{\boldsymbol{exp(cos(v_s, v_{t^+}))}}{\sum_{t'=\{t^+,t^-\}} \boldsymbol{exp(cos(v_s, v_{t'}))}} / \partial W$



[Huang, He, Gao, Deng, Acero, Heck, "Learning DSSM for web search using clickthrough data," CIKM, 2013]

DSSM: a similarity-driven Sent2Vec model

Runtime:



[Huang, He, Gao, Deng, Acero, Heck, "Learning DSSM for web search using clickthrough data," CIKM, 2013]

Training objectives

Objective: cosine similarity based loss

Using web search as an example:

- a query q and a list of docs $D = \{d^+, d_1^-, \ldots d_K^-\}$
  - d^+ positive doc; $d_1^-, \ldots d_K^-$ are negative docs to q (e.g., sampled from not clicked docs)
- Objective: the posterior probability of the clicked doc given the query

$$P_\theta(d^+|q) = \frac{\exp\left(\gamma\, cos(v_{\boldsymbol{\theta}}(q), v_{\boldsymbol{\theta}}(d^+))\right)}{\sum_{d\in\boldsymbol{D}} \exp\left(\gamma\, cos(v_{\boldsymbol{\theta}}(q), v_{\boldsymbol{\theta}}(d))\right)}$$

e.g., $v_{\boldsymbol{\theta}}(q) = \sigma(W_{s,4} \times \sigma(W_{s,3} \times \sigma(W_{s,2} \times ltg(q)))$

$v_{\boldsymbol{\theta}}(d) = \sigma(W_{t,4} \times \sigma(W_{t,3} \times \sigma(W_{t,2} \times ltg(d)))$

where $\theta = \{W_{s,2\sim4}, W_{t,2\sim4}\}, \sigma(\,)$ is a tanh function.

Using Convolutional Neural Net in DSSM

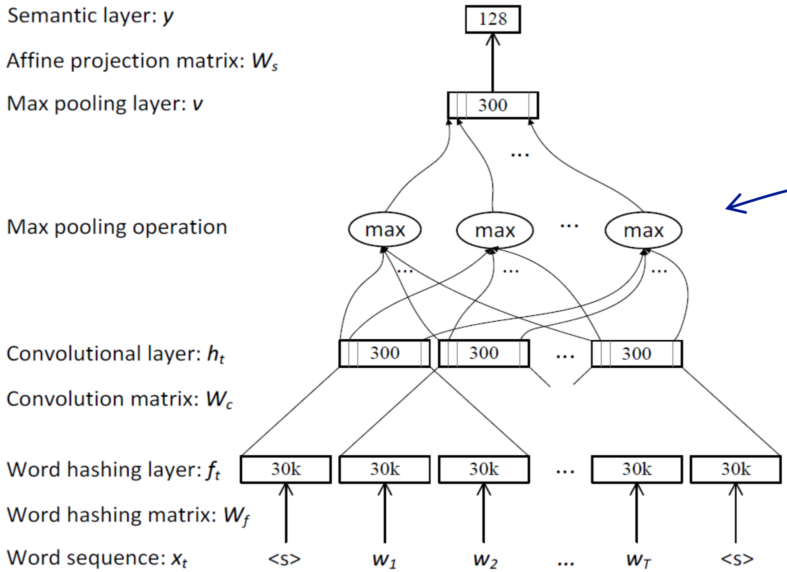


Figure 1: Illustration of the C-DSSM. A convolutional layer with the window size of three is illustrated.

Model local context at the convolutional layer

Model global context at the pooling layer

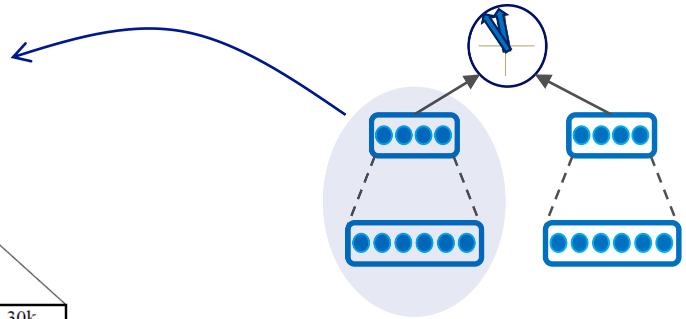


Figure credit [Shen, He, Gao, Deng, Mesnil, WWW2014]

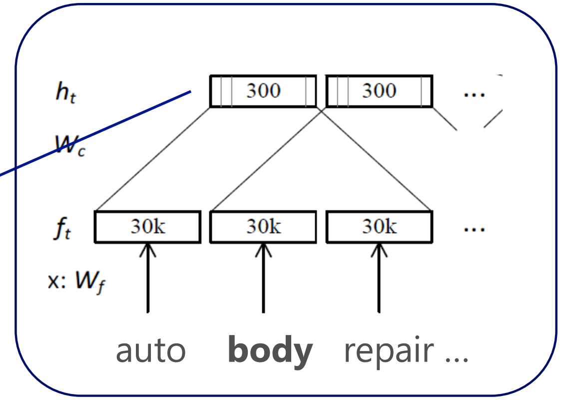
Strong performance on many NLP tasks

Information Retrieval: [Shen, He, Gao, Deng, Mesnil, WWW2014 & CIKM2014], Entity Ranking: [Gao, Pantel, Gamon, He, Deng, Shen, EMNLP2014], Question answering: [Yih, He, Meek, ACL2014; Yih, Chang, He, Gao, ACL2015], Recommendation [Elkahky, Song, He, WWW2015], Spoken language understanding [Chen, Hakkani-Tür, He, ICASSP2016]…

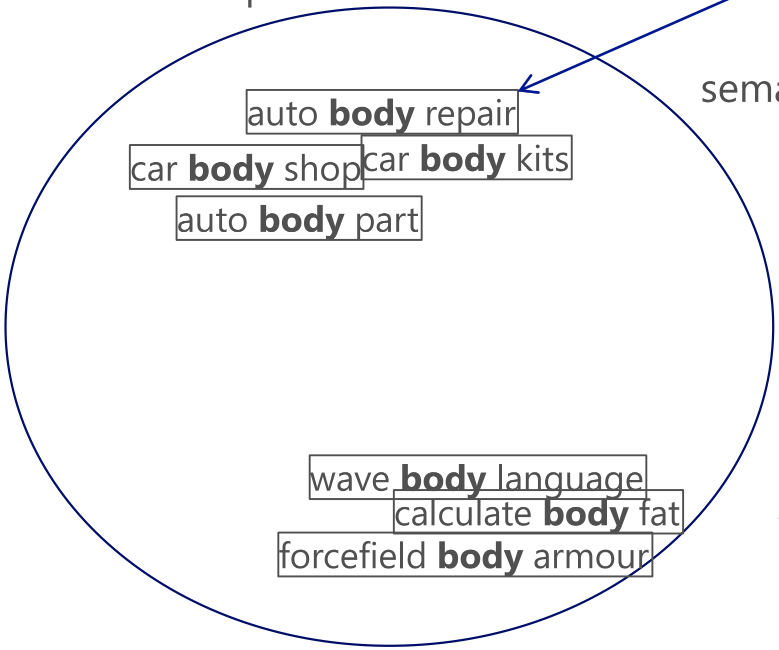
– What does the model learn at the convolutional layer?

Capture the local context dependent word sense

- Learn one embedding vector for each local context-dependent word



$$h_t = W_c \times [f_{t-1}, f_t, f_{t+1}]$$



semantic space

The similarity between different "***body***" within contexts

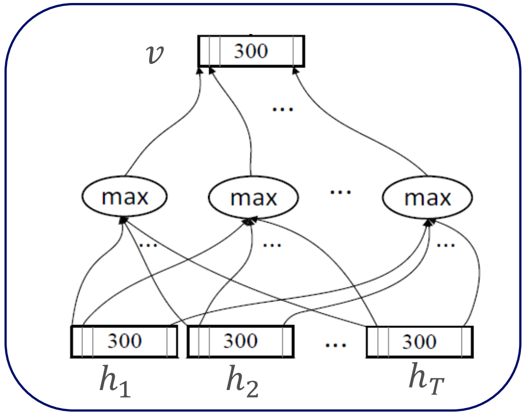
car **body** shop	**cosine similarity**
car **body** kits	0.698
auto **body** repair	0.578
auto **body** parts	0.555
wave **body** language	0.301
calculate **body** fat	0.220
forcefield **body** armour	0.165

high similarity

low similarity

CDSSM: What happens at the max-pooling layer?

- Aggregate *local topics* to form the *global intent*
- Identify salient words/phrase at the max-pooling layer



$$v(i) = \max_{t=1,\ldots,T} \{h_t(i)\}$$

where $i = 1, \ldots, 300$

Words that win the most active neurons at the **max-pooling layers:**

auto body repair cost calculator software

Usually, those are salient words containing clear intents/topics

DSSM for Information Retrieval

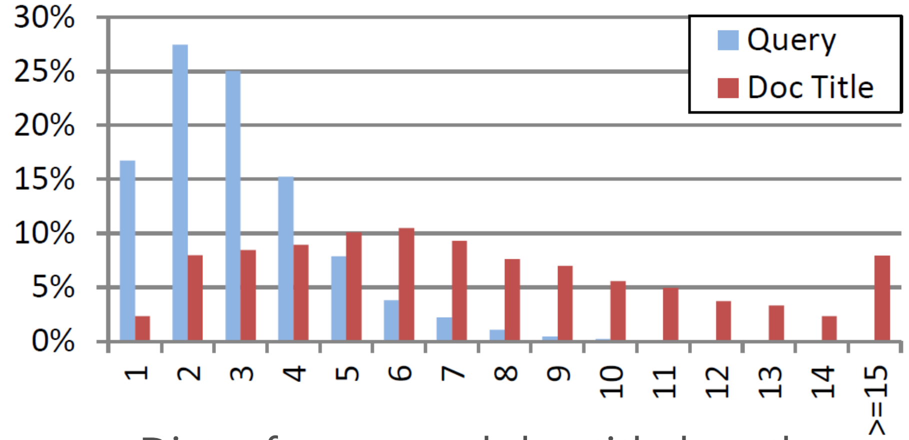
- Training Dataset
  - Mine semantically-similar text pairs from Search Logs, e.g., 30 Million (Query, Document) Click Pairs

QUERY (Q)	Clicked Doc Title (T)
how to deal with stuffy nose	best home remedies for cold and flu
stuffy nose treatment	best home remedies for cold and flu
cold home remedies	best home remedies for cold and flu
… …	… …
go israel	forums goisrael community
skate at wholesale at pr	wholesale skates southeastern skate supply
breastfeeding nursing blister baby	clogged milk ducts babycenter

[Gao, He, Nie, CIKM2010]

Experimental Setting

- Testing Dataset
  - **12,071** English queries
  - around 65 web document associated to each query in average
  - Human gives each <query, doc> pair the label, with range **0 to 4**
  - 0: Bad 1: Fair 2: Good 3: Perfect 4: Excellent
- Evaluation Metric: (higher the better)
  - NDCG
- Using NVidia GPU K40 for training

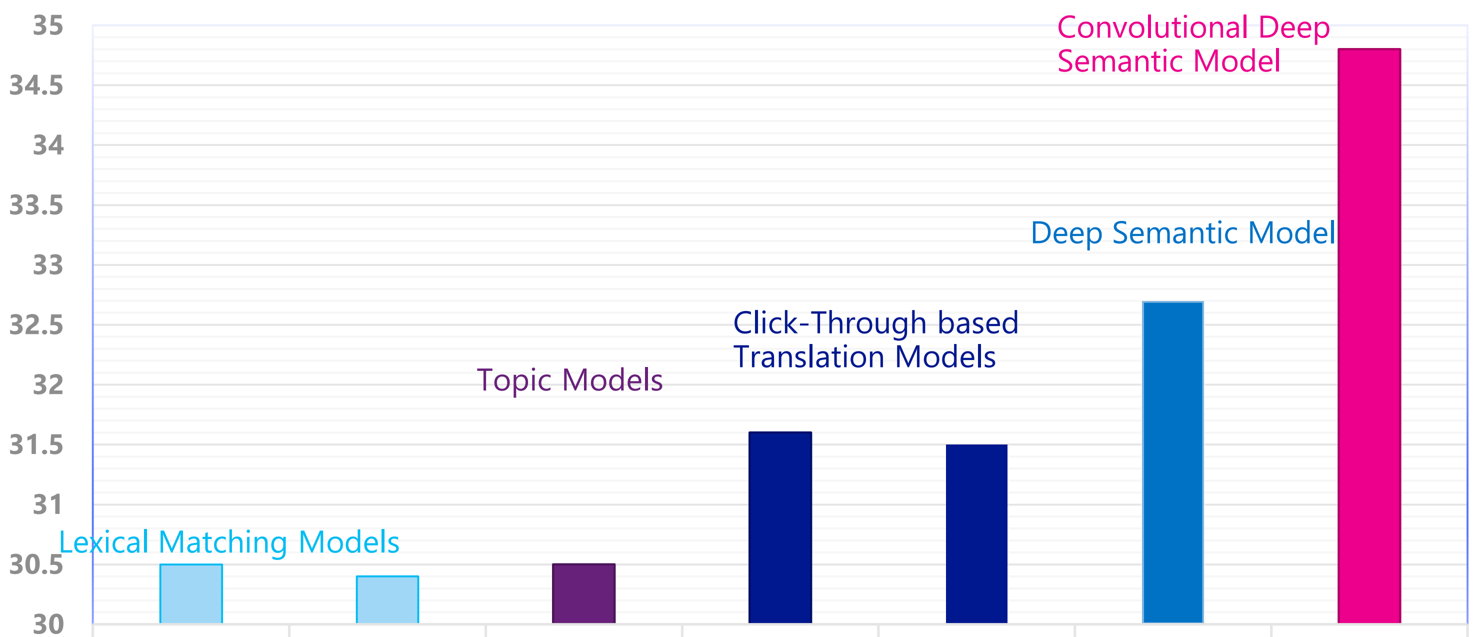


Dist. of query and doc title length

Results

[Shen et al. CIKM2014]

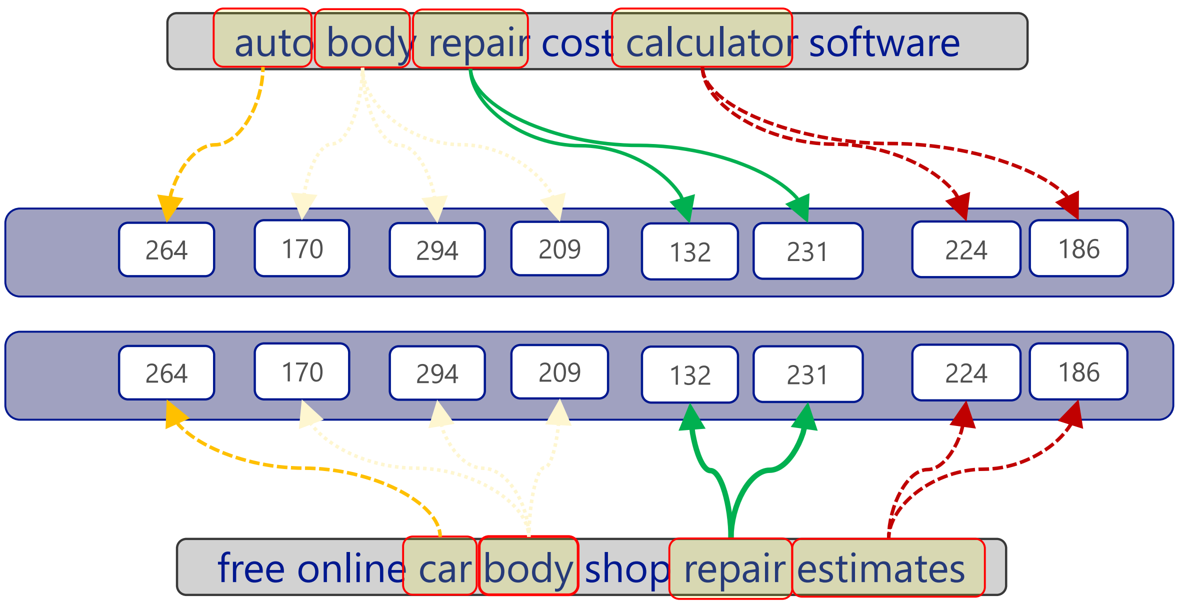
NDCG@1 Results



	BM25	ULM	PLSA	BLTM	WTM	DSSM	CDSSM
■ NDCG@1	30.5	30.4	30.5	31.6	31.5	32.7	34.8

Example: semantic matching

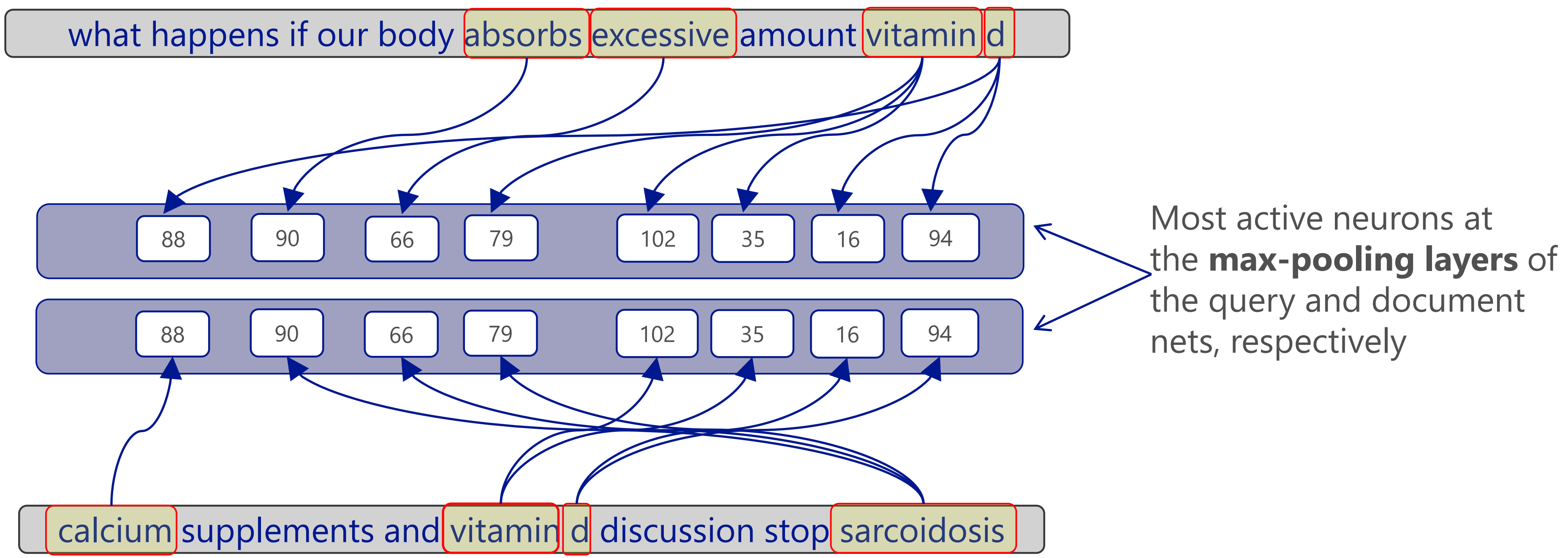
- Semantic matching of query and document



Most active neurons at the **max-pooling layers** of the query and document nets, respectively

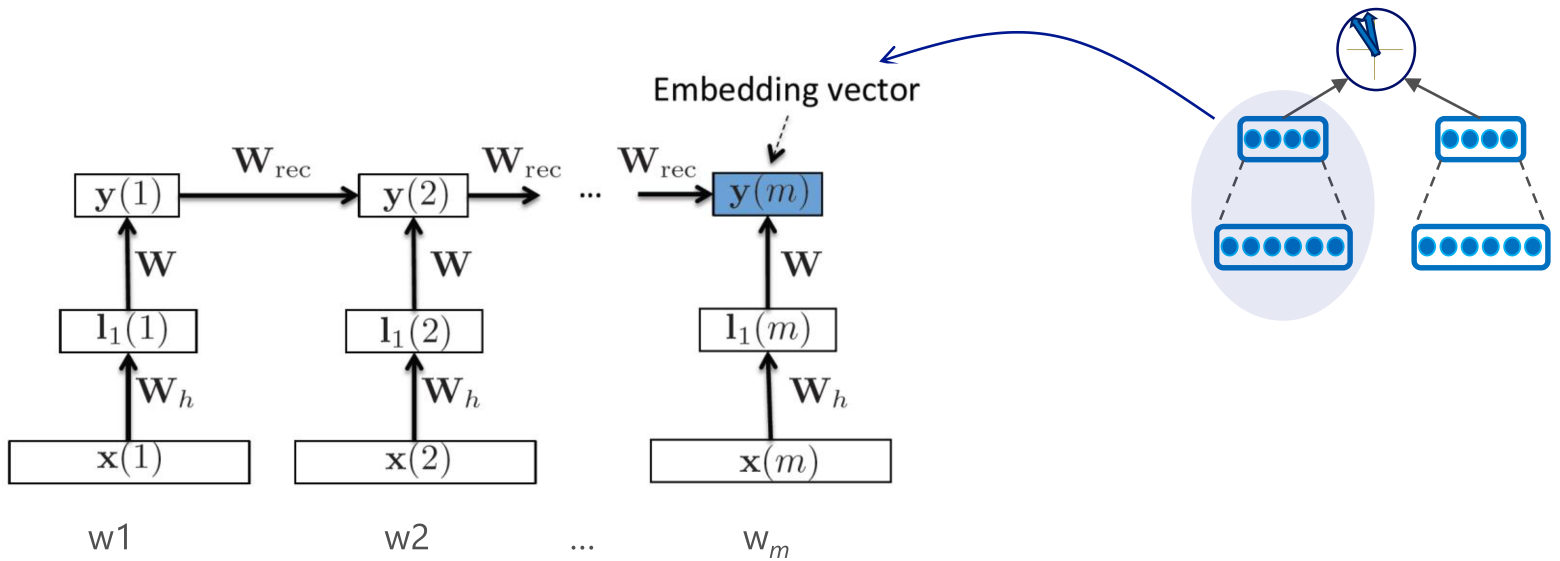
More complex semantic matching example

sarcoidosis is a disease, a symptom is excessive amount of calcium in one's urine and blood. So medicines that increase the absorbing of calcium should be avoid. While Vitamin d is closely associated to calcium absorbing.
We observed that "sarcoidosis" in the document title and "absorbs" "excessive" and "vitamin (d)" in the query have high activations at neurons 90, 66, 79, indicating that the model knows that "sarcoidosis" share similar semantic meaning with "absorbs" "excessive" "vitamin (d)", collectively.



Recurrent DSSM

- Encode the word one by one in the recurrent hidden layer
- The hidden layer at the last word codes the semantics of the full sentence
- Model is trained by a cosine similarity driven objective

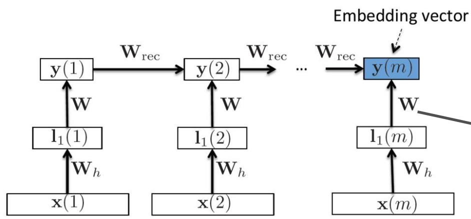


[Palangi, Deng, Shen, Gao, He, Chen, Song, Ward, Deep Sentence Embedding Using the LSTM network: Analysis and Application to IR, IEEE TASL, 2016]

Using LSTM cells

LSTM (long short term memory) uses special cells in RNN

[Hochreiter and J. Schmidhuber, 1997]



$$
\begin{aligned}
\mathbf{y}_g(t) &= g(\mathbf{W}_4\mathbf{l}_1(t) + \mathbf{W}_{\mathrm{rec}4}\mathbf{y}(t-1) + \mathbf{b}_4) \\
\mathbf{i}(t) &= \sigma(\mathbf{W}_3\mathbf{l}_1(t) + \mathbf{W}_{\mathrm{rec}3}\mathbf{y}(t-1) + \mathbf{W}_{p3}\mathbf{c}(t-1) + \mathbf{b}_3) \\
\mathbf{f}(t) &= \sigma(\mathbf{W}_2\mathbf{l}_1(t) + \mathbf{W}_{\mathrm{rec}2}\mathbf{y}(t-1) + \mathbf{W}_{p2}\mathbf{c}(t-1) + \mathbf{b}_2) \\
\mathbf{c}(t) &= \mathbf{f}(t) \circ \mathbf{c}(t-1) + \mathbf{i}(t) \circ \mathbf{y}_g(t) \\
\mathbf{o}(t) &= \sigma(\mathbf{W}_1\mathbf{l}_1(t) + \mathbf{W}_{\mathrm{rec}1}\mathbf{y}(t-1) + \mathbf{W}_{p1}\mathbf{c}(t) + \mathbf{b}_1) \\
\mathbf{y}(t) &= \mathbf{o}(t) \circ h(\mathbf{c}(t)) \qquad (2)
\end{aligned}
$$

where $\circ$ denotes Hadamard (element-wise) product.

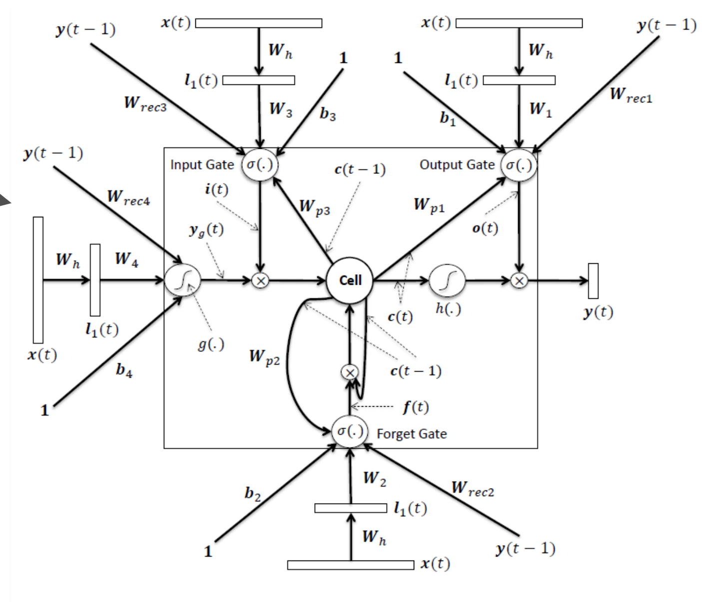


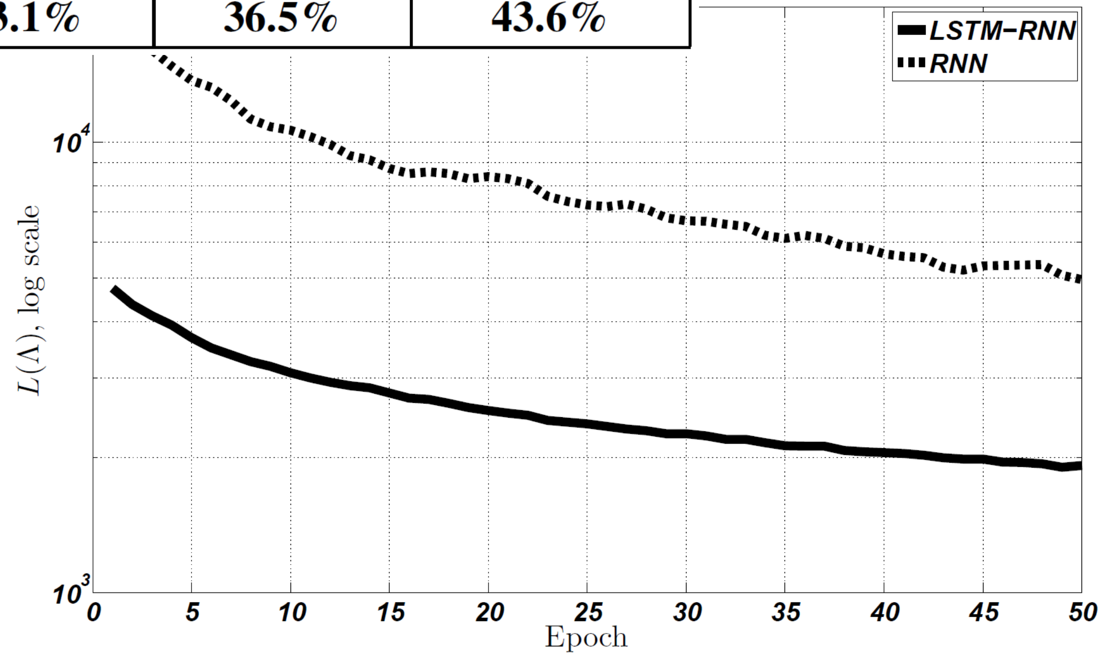
Figure 2. The basic LSTM architecture used for sentence embedding

Results

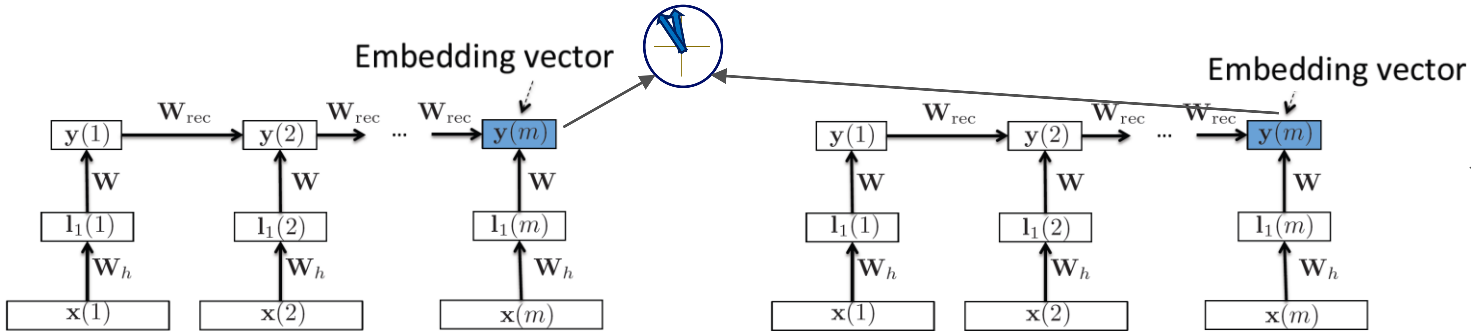
Model	NDCG@1	NDCG@3	NDCG@10
BM25	30.5%	32.8%	38.8%
PLSA (T=500)	30.8%	33.7%	40.2%
DSSM (nhid = 288/96), 2 Layers	31.0%	34.4%	41.7%
CLSM (nhid = 288/96), 2 Layers	31.8%	35.1%	42.6%
RNN (nhid = 288), 1 Layer	31.7%	35.0%	42.3%
LSTM-RNN (ncell = 96), 1 Layer	**33.1%**	**36.5%**	**43.6%**

LSTM learns much faster than regular RNN

LSTM effectively represents the semantic information of a sentence using a vector



Related work: DSSM vs. Seq2Seq



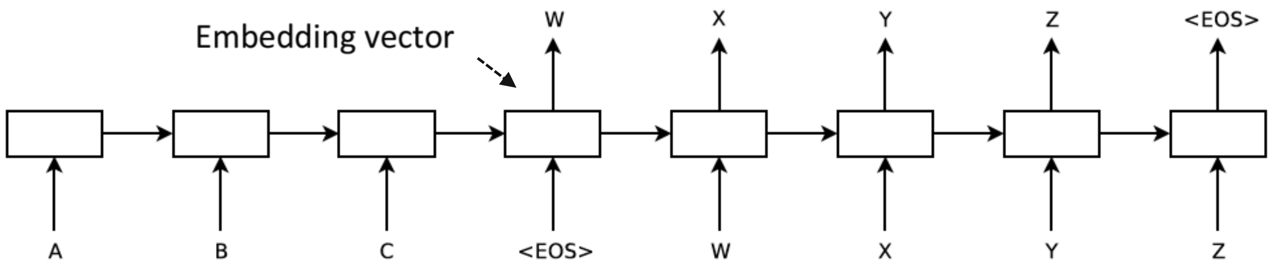
{Palangi, Deng, Shen, Gao, He, Chen, Song, Ward, 2016]

Source sentence

Target sentence

DSSM optimizes *sentence-level* semantic similarity

vs.



Source sentence

Target sentence

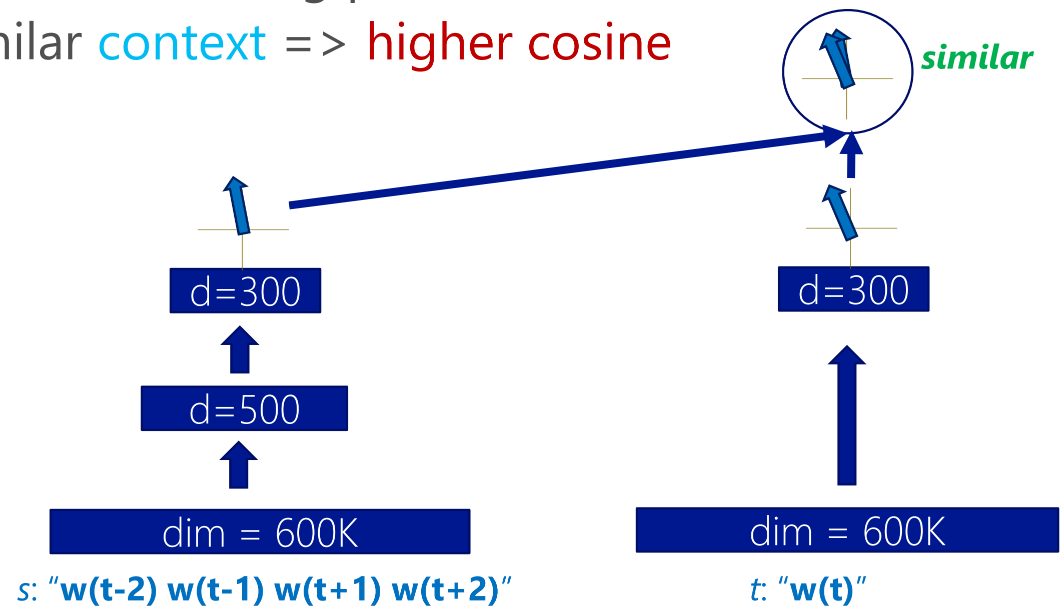
Seq2Seq optimizes *word-level* cross-entropy

[Sutskever, Vinyals, Le, 2014. Sequence to Sequence Learning with Neural Networks]

DSSM for Word embedding learning

- Learn a word's semantic meaning by means of its neighbors (context)
  - Construct context <-> word training pair for DSSM
  - Similar words with similar context => higher cosine
- Training Condition:
  - 600K vocabulary size
  - 1B words from Wikipedia
  - 300-dimentional vector

You shall know a word by the company it keeps (J. R. Firth 1957: 11)



[Song, He, Gao, Deng, 2014]

Evaluation on Word Analogy

The dataset contains 19,544 word analogy questions:

Semantic questions, e.g.,: “Athens is to Greece as Berlin is to ?”

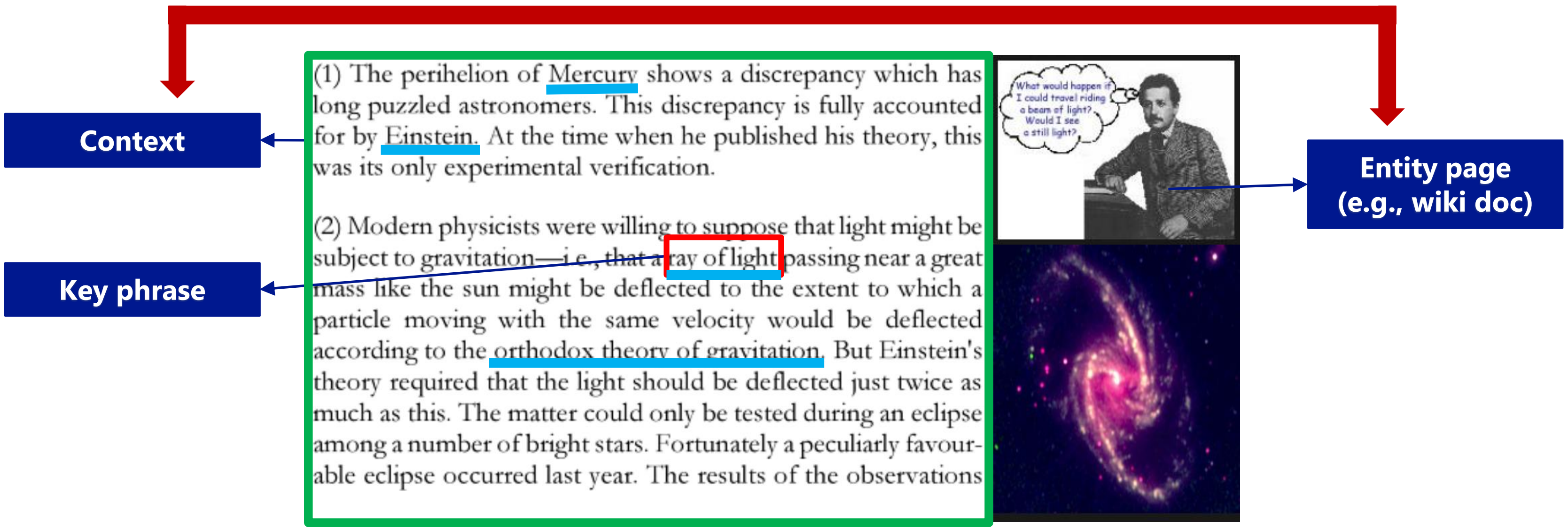
Syntactic questions, e.g.,: “dance is to dancing as fly is to ?”

Model	**Dim**	**Size**	**Accuracy Avg.(sem+syn)**
SG	300	1B	61.0%
CBOW	300	1.6B	36.1%
vLBL	300	1.5B	60.0%
ivLBL	300	1.5B	64.0%
GloVe	300	1.6B	70.3%
DSSM	300	1B	71.9%

(i)vLBL from (Mnih et al., 2013); skip-gram (SG) and CBOW from (Mikolov et al., 2013a,b); GloVe from (Pennington+, 2014)

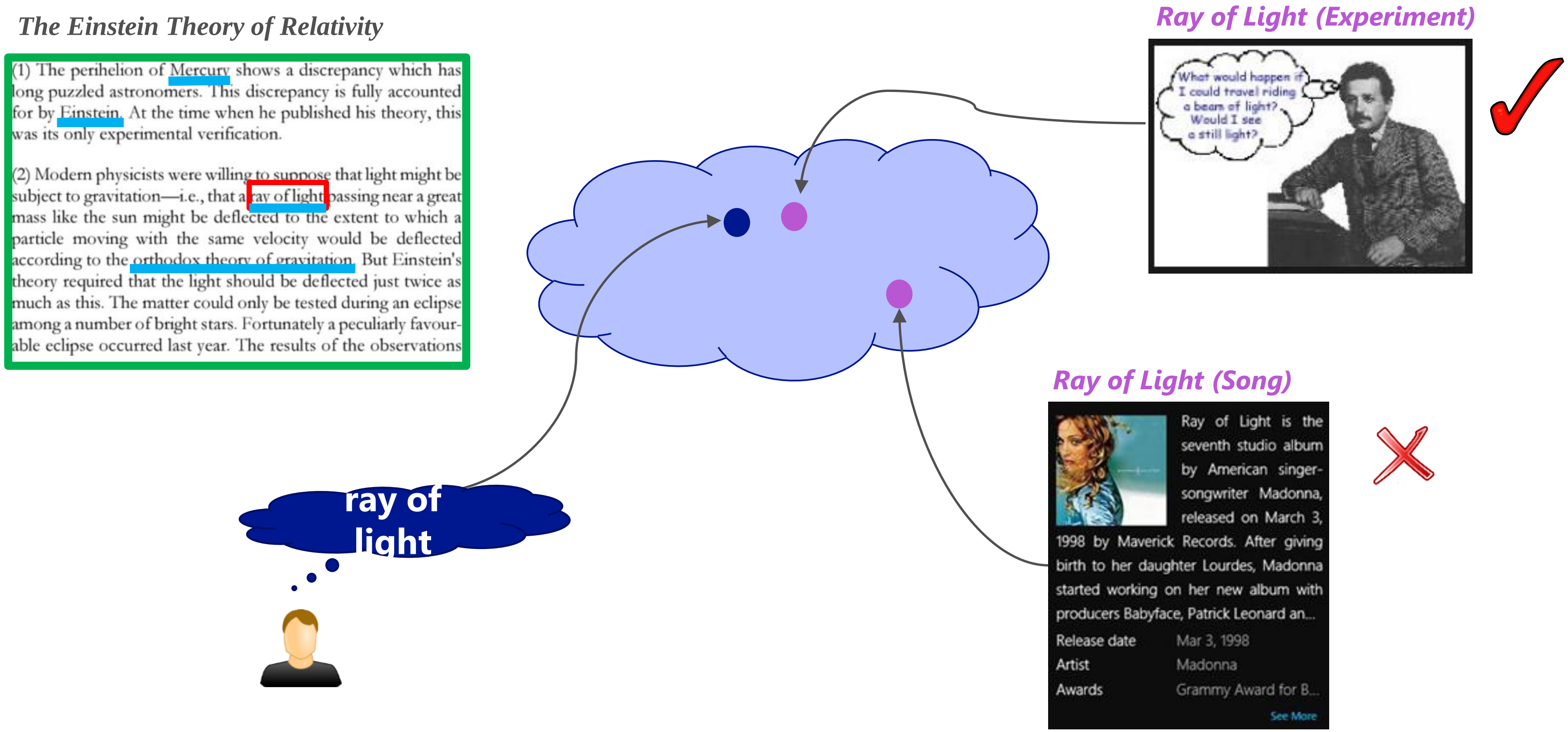
Contextual based Recommendation

Given a user-highlighted text span representing an entity of interest, recommend supplementary document for the entity



Gao, Pantel, Gamon, He, Deng, Shen, "Modeling interestingness with deep neural networks." EMNLP2014

Learning DSSM for contextual recommendaiton



Extract Labeled Pairs from Web Browsing Logs
Contextual Entity Search

- When a hyperlink H points to a Wikipedia P'

http://runningmoron.blogspot.in/

...
I spent a lot of time finding music that was motivating and that I'd also want to listen to through my phone. I could find none. None! I wound up downloading three Metallica songs, a Judas Priest song and one from Bush.
...

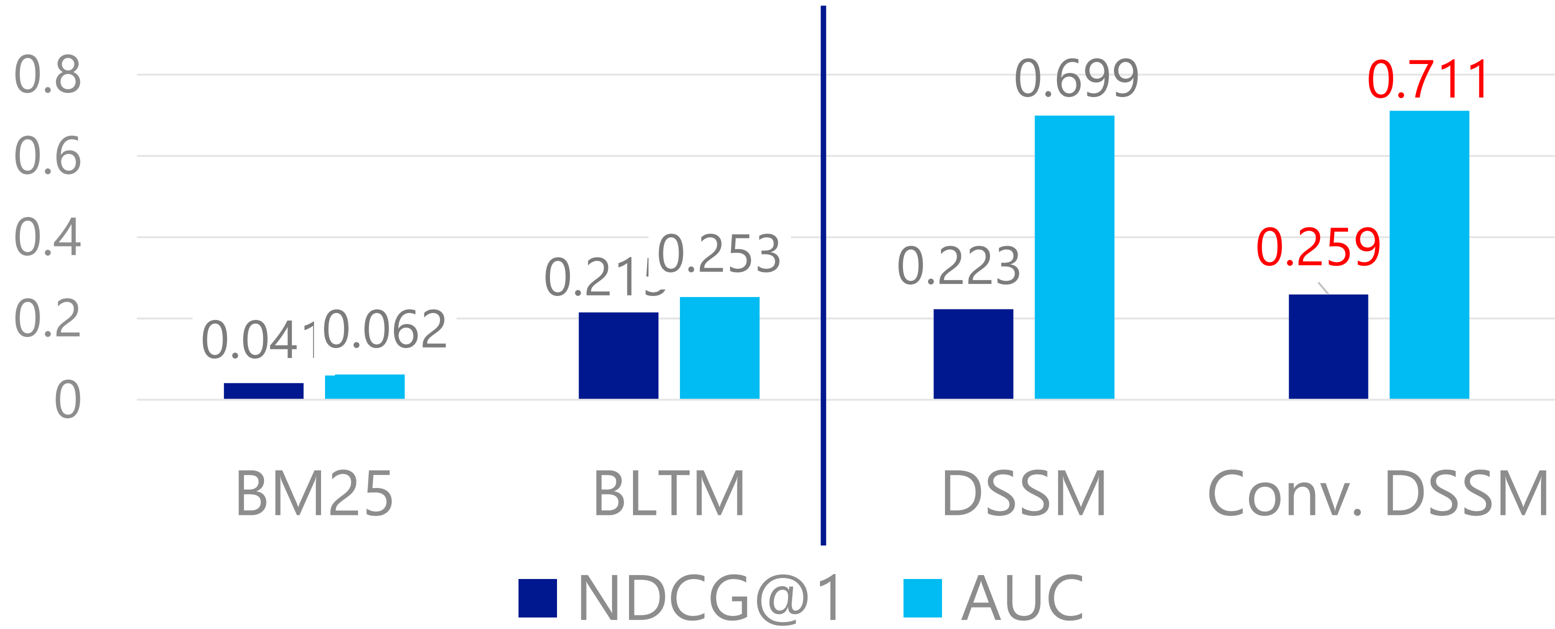
http://en.wikipedia.org/wiki/Bush_(band)

- (anchor text of H & surrounding words, text in P')

Experimental Settings

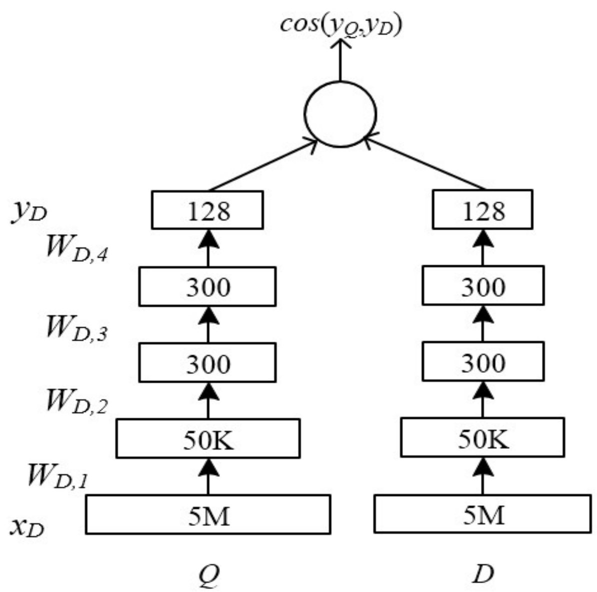
- Training/validation data: 18M of user clicks in wiki pages
- Evaluation data
  - Sample 10k Web documents as the source documents
  - Use named entities in the doc as query; retain up to 100 returned documents as target documents
  - Manually label whether each target document is a good page describing the entity
  - 870k labeled pairs in total
- Evaluation metric: NDCG and AUC

Results: DSSM

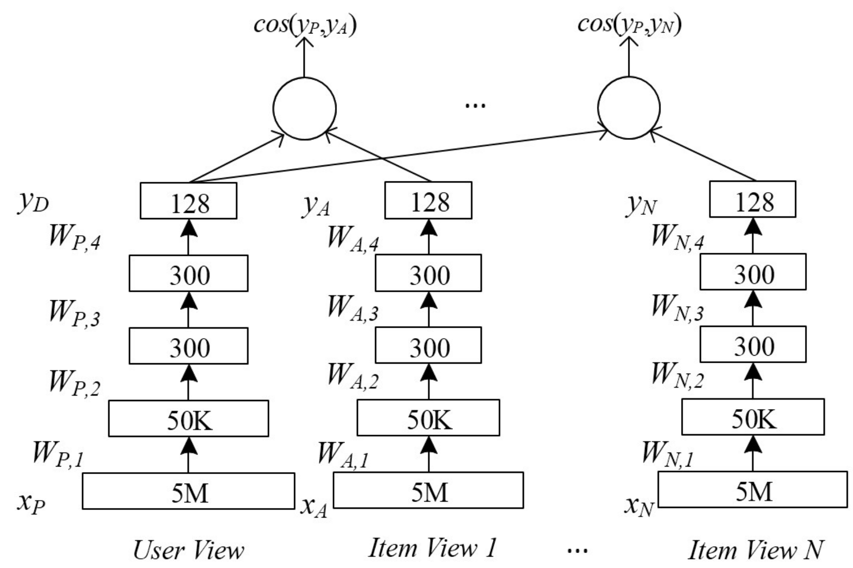


- DSSM: bag-of-words input
- Conv. DSSM: convolutional DSSM

Multi-View DSSM for Recommendation



Single-view DSSM



Multi-view DSSM

Type	DataSet	UserCnt	Feature Size	Joint Users
User View	Search	20M	3.5M	/
Item View	News	5M	100K	1.5M
	Apps	1M	50K	210K
	Movie/TV	60K	50K	16K

[Ali Mamdouh Elkahky , Yang Song , Xiaodong He, "A Multi-View Deep Learning Approach for Cross Domain User Modeling in Recommendation Systems," in WWW 2015]

Experiments

- Multi-View DSSM works the best
  - Much better results than Collaborative Filtering (CF) etc.
  - Outperform Single-View DSSM too
  - Works for cold-start scenarios (New Users)

	Algorithm	All Users		New Users	
		MRR	P@1	MRR	P@1
I	Most Frequent	0.298	0.103	0.303	0.119
	CF	0.337	0.142	/	/
	CCA (TopK) [29]	0.295	0.105	0.295	0.104
	CTR [32]	0.448	0.277	0.319	0.142
II	SV- Kmeans	0.359	0.159	0.336	0.154
	SV-LSH	0.372	0.169	0.339	0.158
	SV-TopK	**0.497**	**0.315**	**0.436**	**0.268**
III	MV-Kmeans	0.362	0.16	0.339	0.156
	MV-TopK	0.517	0.335	0.466	0.297
	MV-TopK w/ Xbox	**0.527**	**0.347**	**0.473**	**0.306**

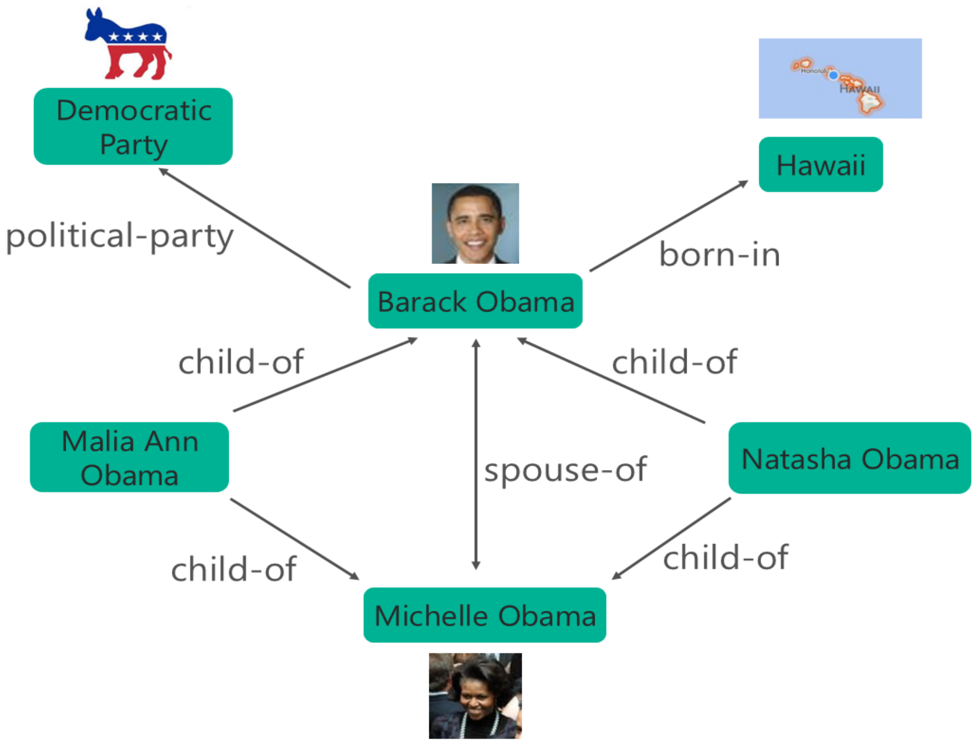
Table 3: Results for different algorithms on Windows Apps Data Set. Type *I* algorithms are baseline methods we compare with. Type *II* are single user-item view methods trained using the original DSSM framework. Type *III* are multi-view DNN models we proposed. The best performance is achieved by training a MV-DNN on all three user-item views with TopK as feature selection method.

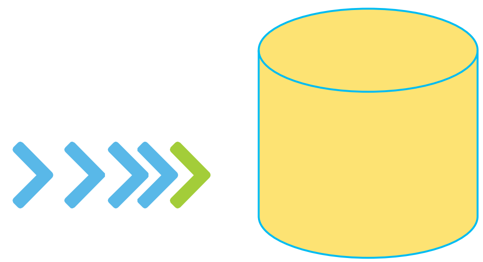
Some related work

Deep CNN for text input Mainly classification tasks in the paper	[Kalchbrenner, Grefenstette, Blunsom, A Convolutional Neural Network for Modelling Sentences, ACL2014]
Sequence to sequence learning	[Sutskever, Vinyals, Le, 2014. Sequence to Sequence Learning with Neural Networks]
Paragraph Vector Learn a vector for a paragraph	Quoc Le, Tomas Mikolov, Distributed Representations of Sentences and Documents, in ICML 2014
Recursive NN (ReNN) Tree structure, e.g., for parsing	[Socher, Lin, Ng, Manning, "Parsing natural scenes and natural language with recursive neural networks", 2011]
Tensor product representation (TPR) Tree representation	[Smolensky and Legendre: The Harmonic Mind, From Neural Computation to Optimality-Theoretic Grammar, MIT Press, 2006]
Tree-structured LSTM Network Tree structure LSTM	[Tai, Socher, Manning. 2015. Improved Semantic Representations From Tree-Structured LSTM Networks.]

Embedding Knowledge Base / Knowledge Graph

- Captures world knowledge by storing properties of millions of entities, as well as relations among them





Freebase
DBpedia
YAGO
NELL
OpenIE/ReVerb

Current KB Applications in NLP & IR

- Question Answering

  "What are the names of Obama's daughters?"

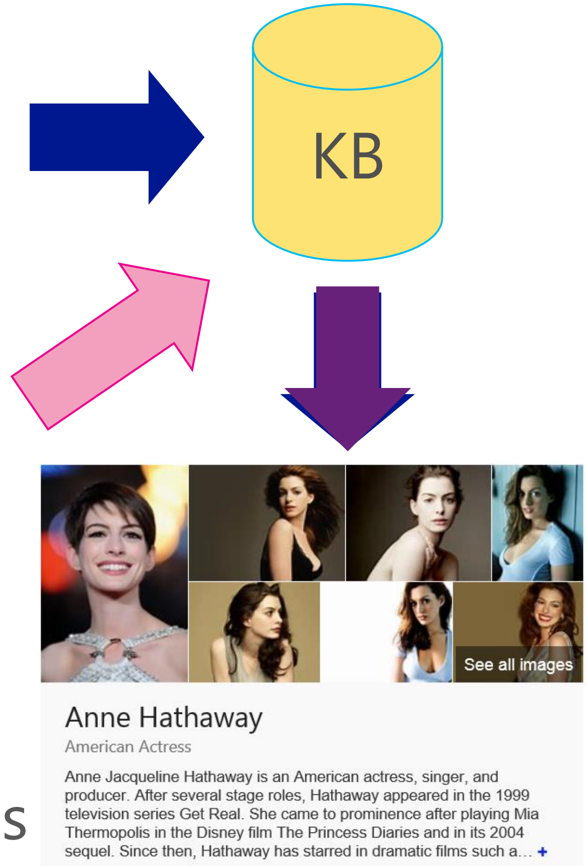
  $\lambda x. parent(Obama, x) \wedge gender(x, Female)$

- Information Extraction
  - *"Hathaway was born in Brooklyn, New York."*

  $bornIn(Hathaway, Brooklyn)$

  $contains(New\ York, Brooklyn)$

- Web Search
  - Identify entities and relationships in queries



Reasoning with Knowledge Base

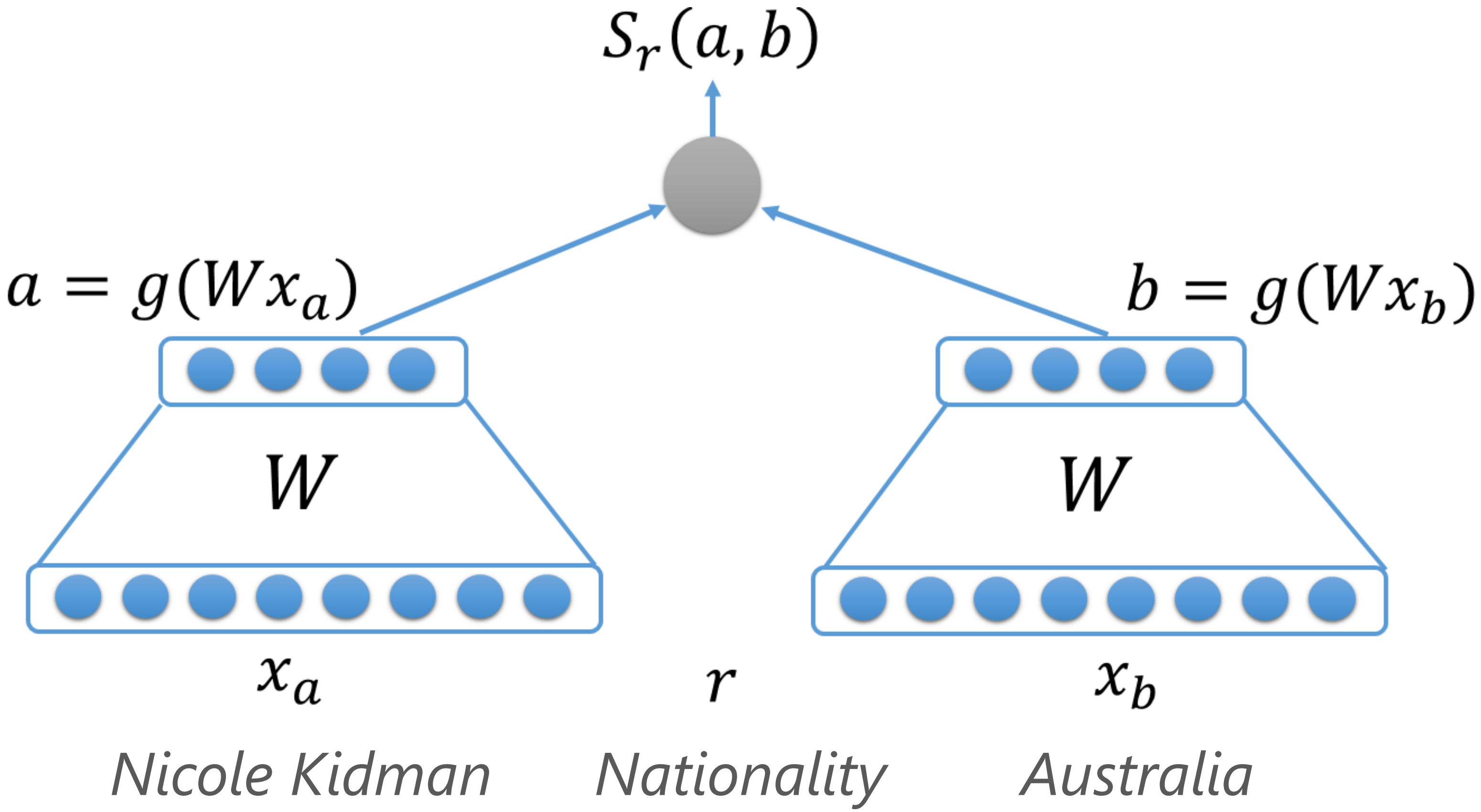
- Knowledge base is never complete!
  - Predict new facts: $Nationality(Natasha\ Obama, ?)$
  - Mine rules: $BornInCity(a, b) \wedge CityInCountry(b, c) \Rightarrow Nationality(a, c)$
- Modeling multi-relational data
  - Statistical relational learning [Getoor & Taskar, 2007]
  - Path ranking methods (e.g., random walk) [e.g., Lao+ 2011]
  - Knowledge base embedding
    - Very efficient
    - Better prediction accuracy

Knowledge Base Embedding

- Each entity in a KB is represented by an R^d vector
- Predict whether (e_1, r, e_2) is true by $f_r(\boldsymbol{v}_{e_1}, \boldsymbol{v}_{e_2})$

- Recent neural network based KB embedding
  - SME [Bordes+, AISTATS-12], NTN [Socher+, NIPS-13], TransE [Bordes+, NIPS-13], Bilinear-Diag [Yang+, ICLR2015]

Neural Knowledge Base Embedding



Relation Operators

Relation representation	Scoring Function $S_r(a, b)$	# Parameters
Vector (TransE) (Bordes+ 2013)	$\|\|a - b + V_r\|\|_{1,2}$	$O(n_r \times k)$
Matrix (Bilinear) (Bordes+ 2012, Collobert & Weston 2008)	$a^T M_r b$ $u^T f(M_{r1}a + M_{r2}b)$	$O(n_r \times k^2)$
Tensor (NTN) (Socher+ 2013)	$u^T f(a^T T_r b + M_{r1}a + M_{r2}b)$	$O(n_r \times k^2 \times d)$
Diagonal Matrix (Bilinear-Diag) (Yang+ 2015)	$a^T diag(M_r) b$	$O(n_r \times k)$

n_r: #predicates, k: #dimensions of entity vectors, d: #layers

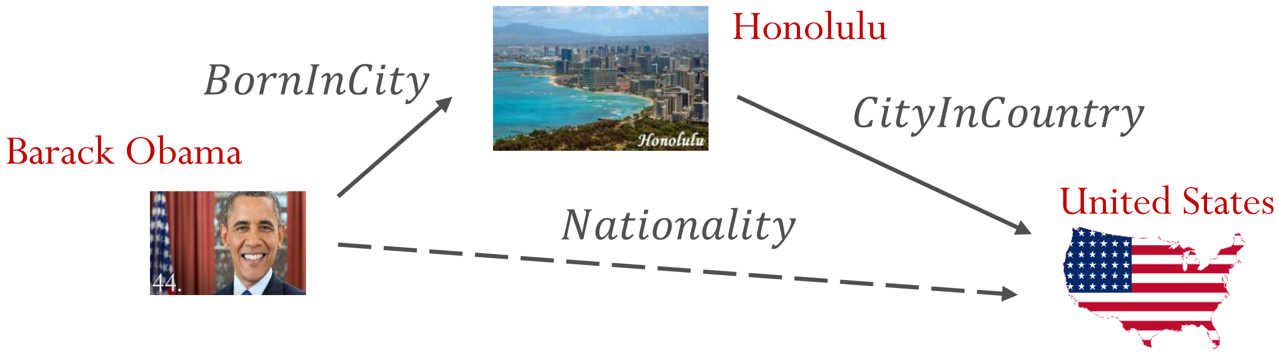
Empirical Comparisons of NN-based KB Embedding Methods [Yang+ ICLR-2015]

- Models with fewer parameters tend to perform better (for the datasets FB-15k and WN).
- The bilinear operator ($\boldsymbol{a}^T \boldsymbol{M}_r \boldsymbol{b}$) plays an important role in capturing entity interactions.
- With the same model complexity, multiplicative operations are superior to additive operations in modeling relations.
- Initializing entity vectors with pre-trained phrase embedding vectors can significantly boost performance.

Mining Horn-clause Rules [Yang+ ICLR-2015]

- Can relation embedding capture relation composition?

$$BornInCity(a,b) \wedge CityInCountry(b,c) \Rightarrow Nationality(a,c)$$



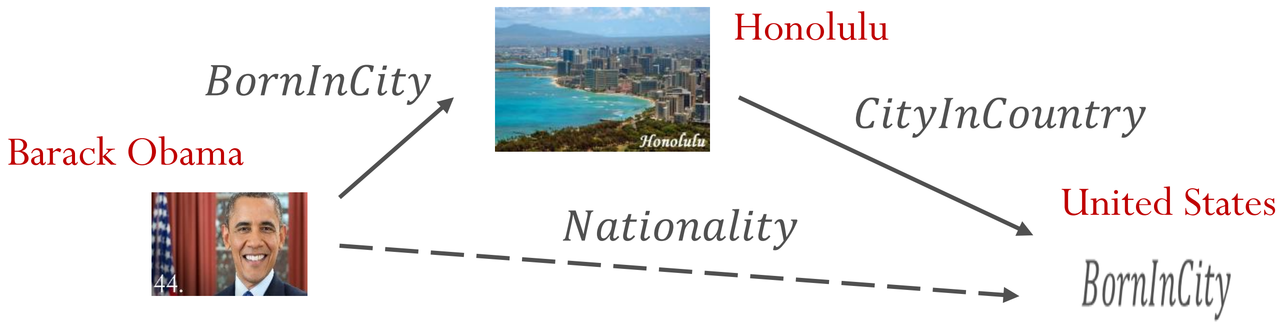
- Embedding-based Horn-clause rule extraction
  - For each relation r, find a chain of relations $r_1 \cdots r_n$, such that:

    $$dist(M_r, M_1 \circ M_2 \circ \cdots \circ M_n) < \theta$$

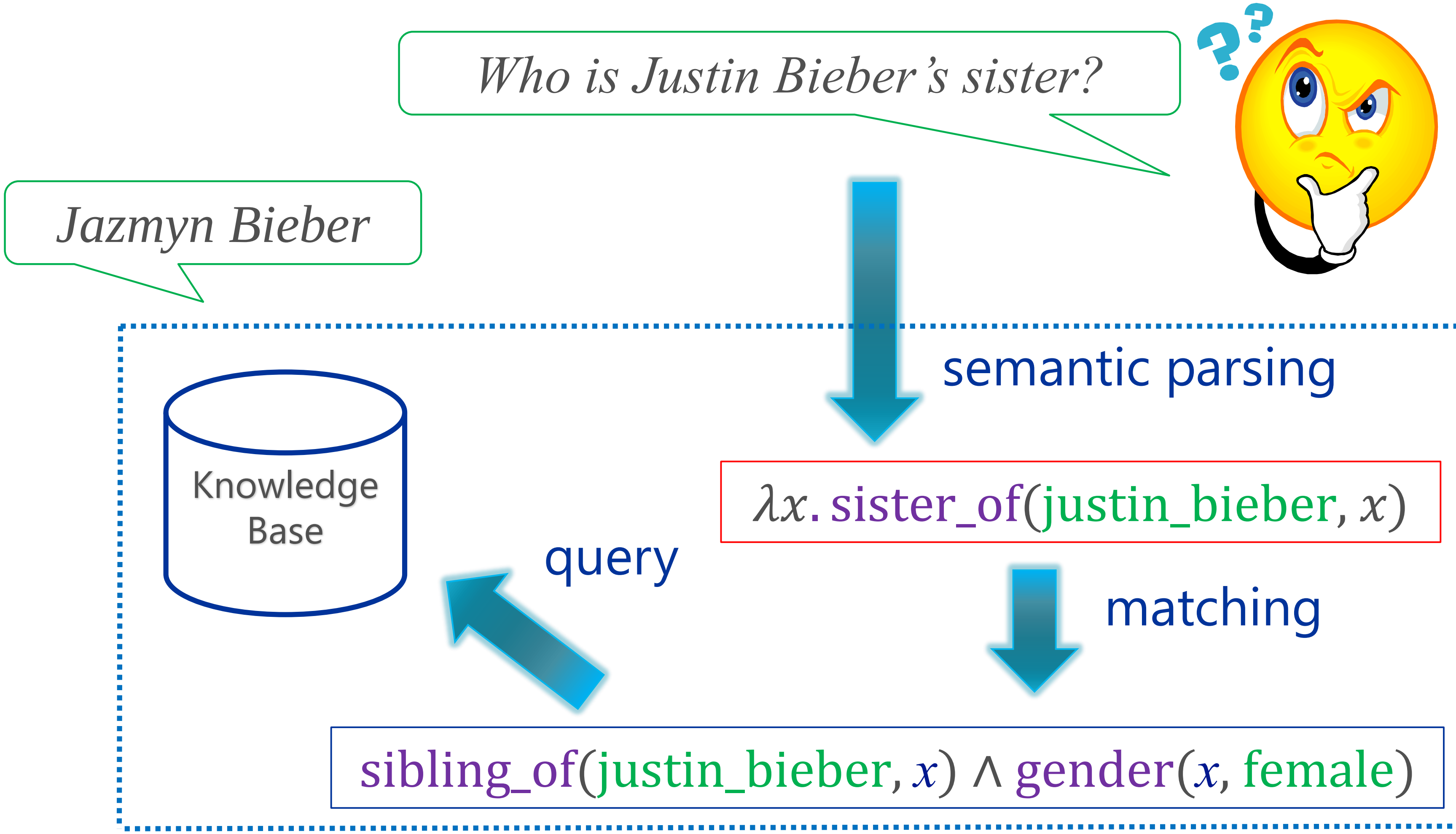
  - $r_1(e_1, e_2) \wedge r_2(e_2, e_3) \cdots \wedge r_n(e_n, e_{n+1}) \rightarrow r(e_1, e_{n+1})$

Learning from Relational Paths [Guu+ EMNLP-15, Garcia-Duran+ EMNLP-15, Toutanova+ ACL-16]

- Single-edge path: $\mathbf{score}(s, r, t) = v_s^{\mathrm{T}} M_r v_t$
  - (Obama, Nationality, USA)
- Multi-edge path: $\mathbf{score}(s, r_1, \cdots, r_k, t) = v_s^{\mathrm{T}} M_{r_1} \cdots M_{r_k} v_t$
  - (Obama, BornInCity, CityInCountry, USA)



KB-based Question Answering

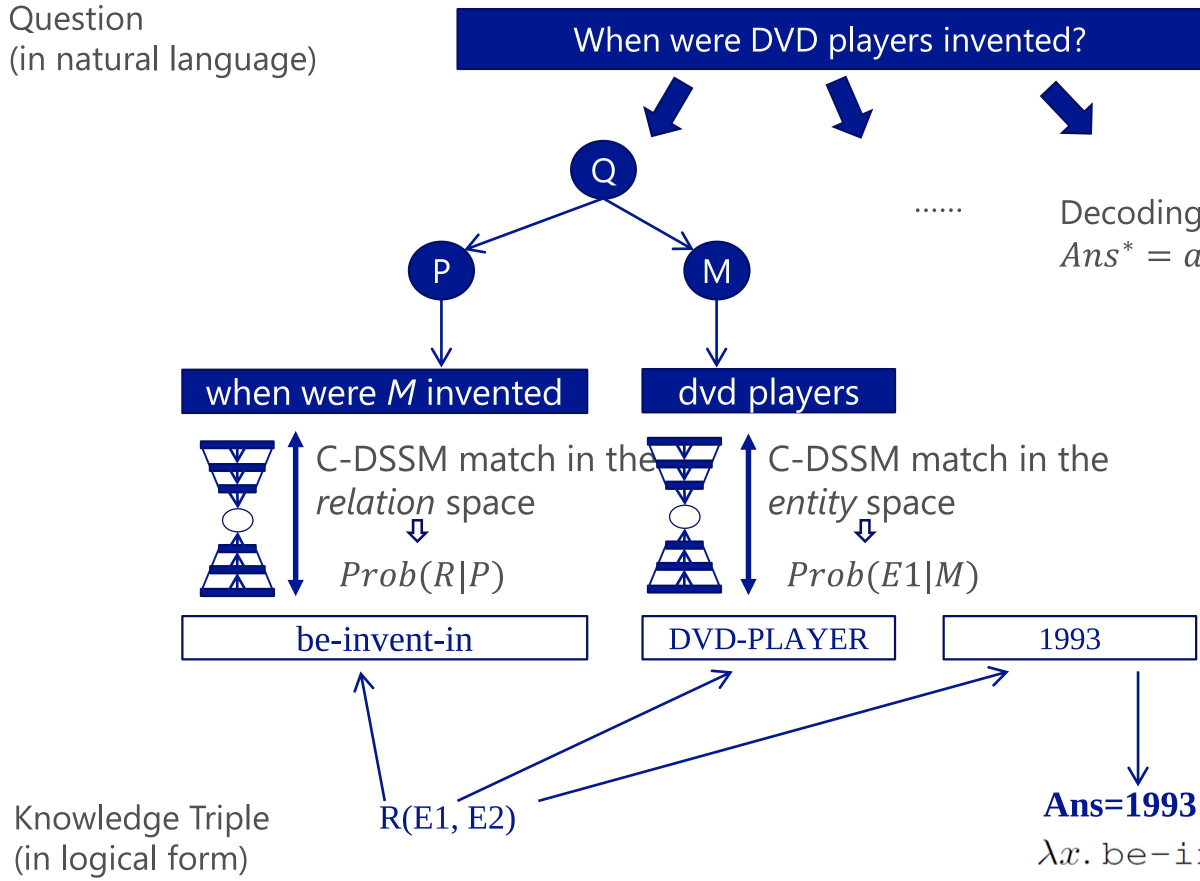


Key Challenge – Language Mismatch

- Lots of ways to ask the same question
  - *"What was the date that Minnesota became a state?"*
  - *"Minnesota became a state on?"*
  - *"When was the state Minnesota created?"*
  - *"Minnesota's date it entered the union?"*
  - *"When was Minnesota established as a state?"*
  - *"What day did Minnesota officially become a state?"*

- Need to map them to the predicate defined in KB
  - location.dated_location.date_founded

DSSM in question answering



Decoding the best answer:
$Ans^* = argmax_{Ans} P(Ans|KB,Q)$

$$\begin{aligned} P(Ans|KB,Q) &= \sum_{SP} P(Ans,SP|KB,Q) \\ &\approx \max_{SP,Triple} P(Ans|SP,KB,Q)P(SP|Q) \\ &\approx \max_{SP,Triple} Prob(R|P) \times Prob(E1|M) \end{aligned}$$

$\lambda x.\, \texttt{be-invent-in(dvd-player,}x\texttt{)}$

Yih, He, Meek, "Semantic parsing for single-relation question answering," ACL 2014

Experiments: Data

Paralex dataset [Fader et al., 2013]

- 1.8M (question, single-relation queries)

  When were DVD players invented?

  $\lambda x.\mathrm{be{-}invent{-}in}(\mathrm{dvd{-}player}, x)$

- 1.2M (relation pattern, relation)

  When were X invented?

  $\mathrm{be{-}invent{-}in}_2$

- 160k (mention, entity)

  Saint Patrick day

  $\mathrm{st{-}patrick{-}day}$

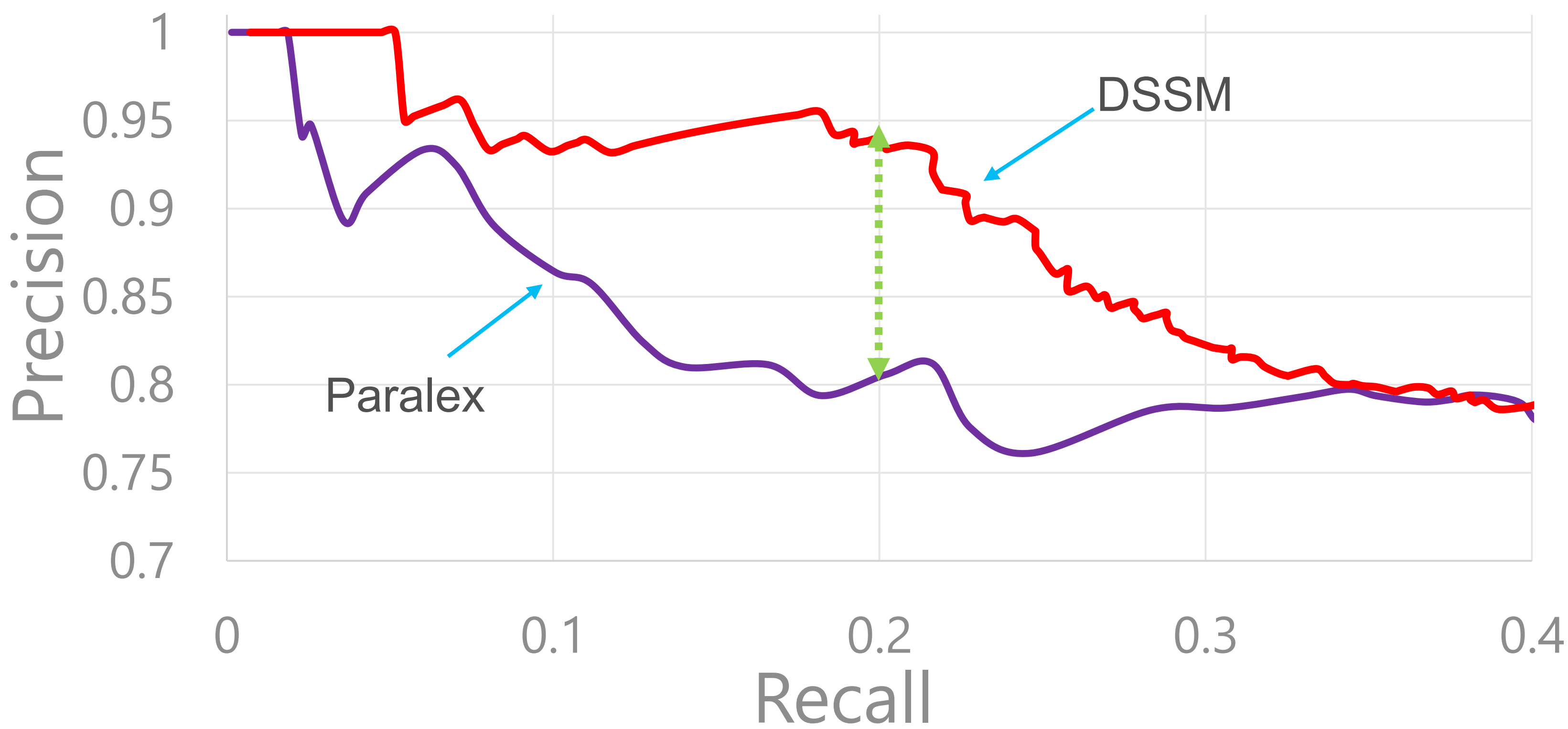
Experiments: Task – Question Answering

- Same test questions in the Paralex dataset
- 698 questions from 37 clusters

  - *What language do people in Hong Kong use?*
    be−speak−in(english, hong−kong)
    be−predominant−language−in
    (cantonese, hong−kong)

  - *Where do you find Mt Ararat?*
    be−highest−mountain−in(ararat, turkey)
    be−mountain−in(ararat, armenia)

Experiments: Results



Answering more complicated questions

WebQuestions Dataset [Berant+ EMNLP-2013]

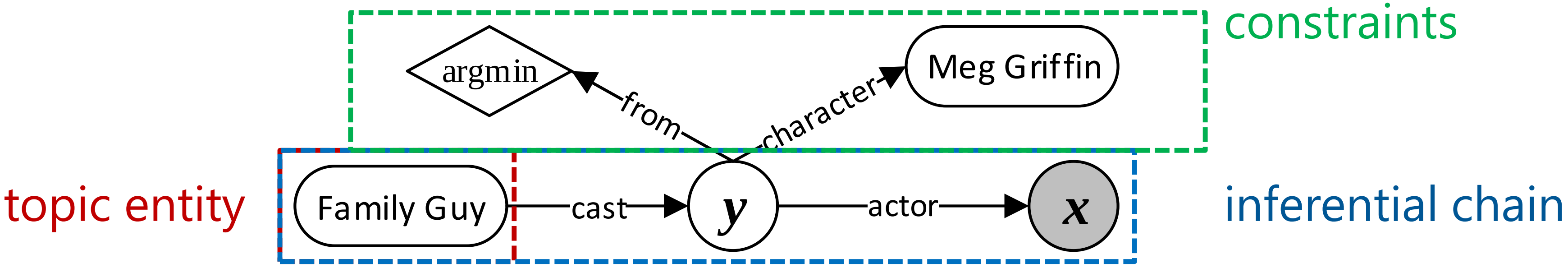
- *What character did Natalie Portman play in Star Wars?* ⇒ Padme Amidala
- *What kind of money to take to Bahamas?* ⇒ Bahamian dollar
- *What currency do you use in Costa Rica?* ⇒ Costa Rican colon
- *What did Obama study in school?* ⇒ political science
- *What do Michelle Obama do for a living?* ⇒ writer, lawyer
- *What killed Sammy Davis Jr?* ⇒ throat cancer

[Examples from Berant]

- 5,810 questions crawled from Google Suggest API and answered using Amazon MTurk
  - 3,778 training, 2,032 testing
  - A question may have multiple answers → using Avg. F1 (~accuracy)

Staged Query Graph Generation

- Query graph
  - Resembles subgraphs of the knowledge base
  - Can be directly mapped to a logical form in λ-calculus
  - Semantic parsing: a search problem that *grows* the graph through actions
- Who first voiced Meg on Family Guy?
- $\lambda x. \exists y. \mathrm{cast}(\mathrm{FamilyGuy}, y) \wedge \mathrm{actor}(y, x) \wedge \mathrm{character}(y, \mathrm{MegGriffin})$

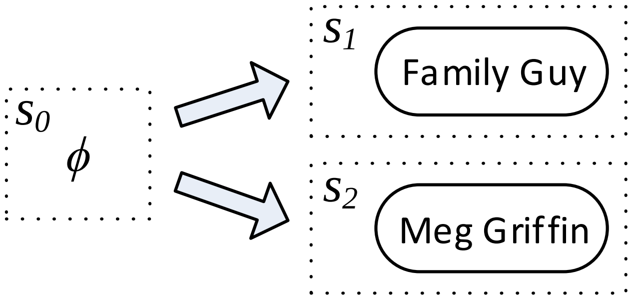


[Yih, Chang, He, Gao, "Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base," ACL 2015]

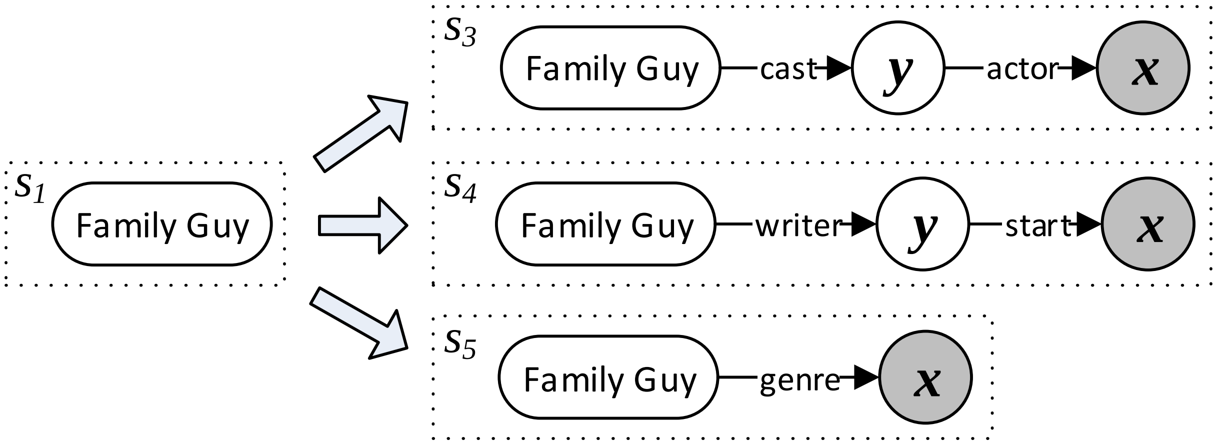
Graph Generation Stages

- Who first voiced Meg on Family Guy?

1. Topic Entity Linking [Yang&Chang ACL-15]



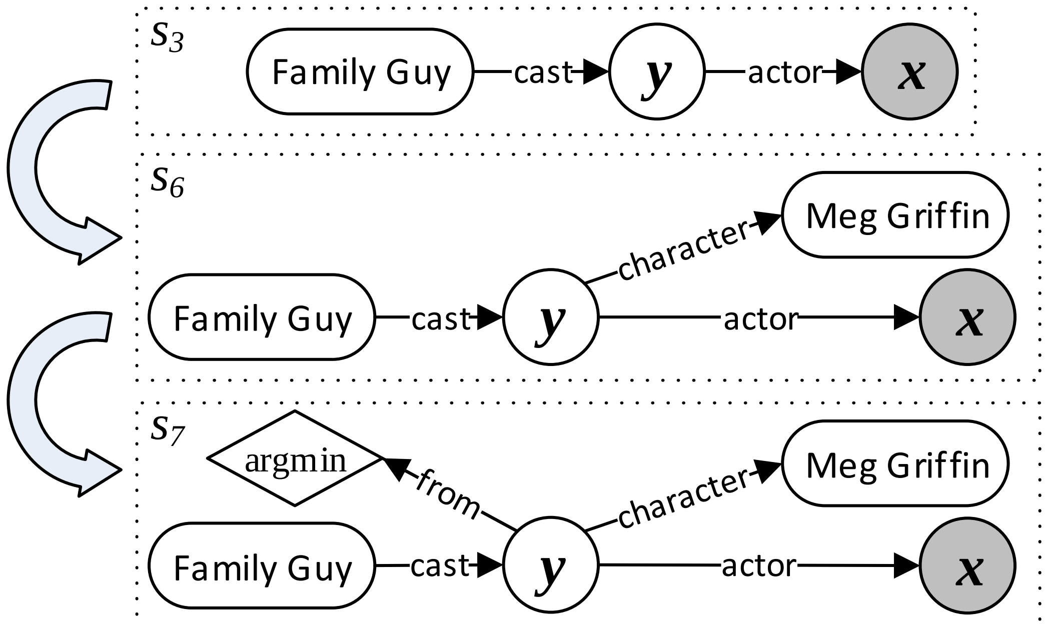
2. Identify the core inferential chain



Graph Generation Stages (cont'd)

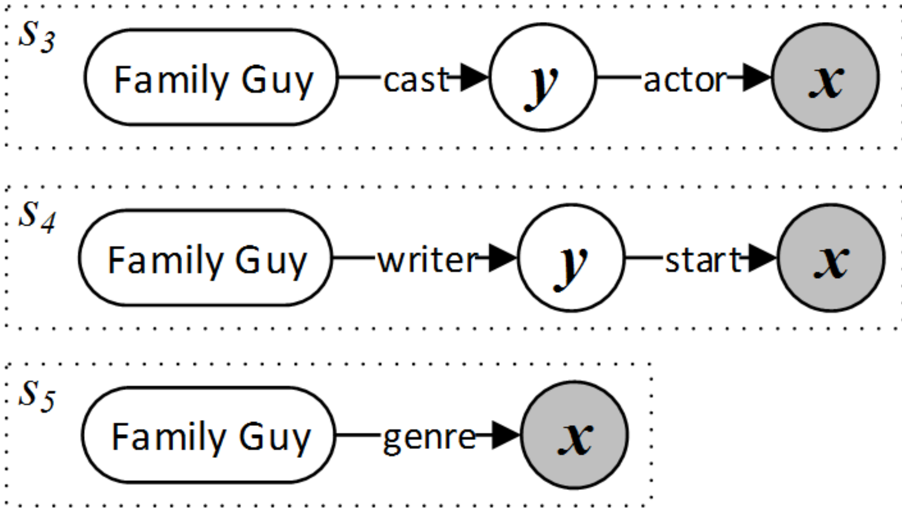
- Who first voiced Meg on Family Guy?

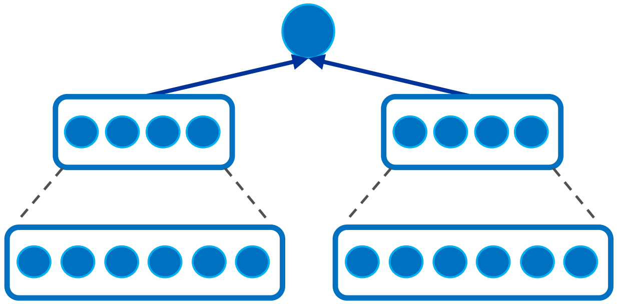
3. Augment constraints



Identify Inferential Chain using DSSM

- Who first voiced Meg on Family Guy?

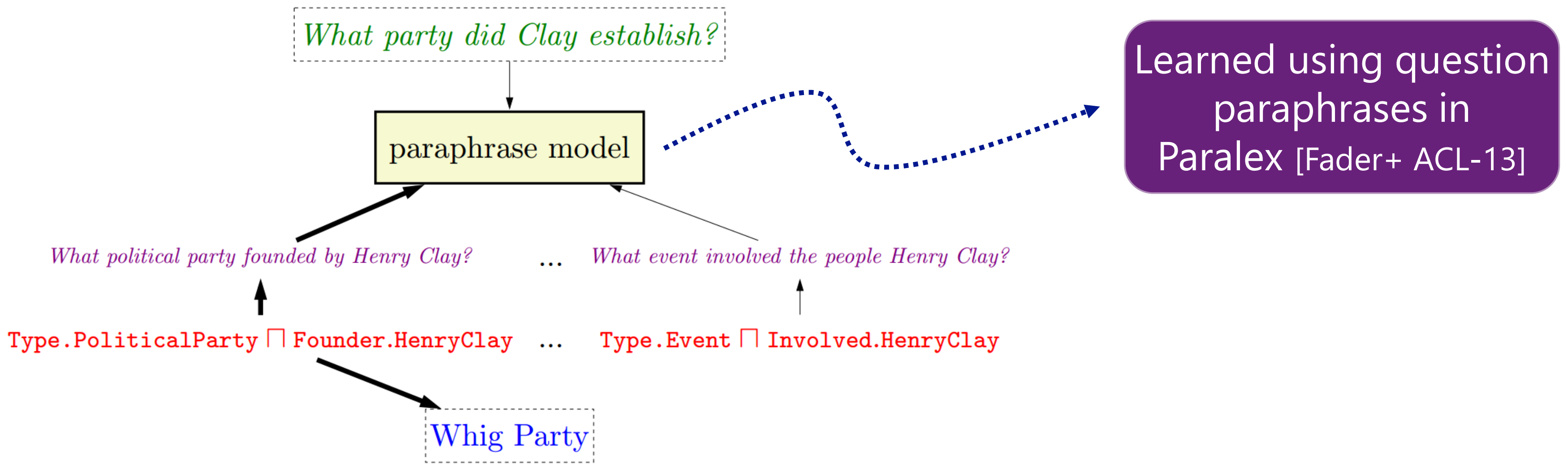




- Semantic match **(“Who first voiced Meg on ⟨*e*⟩”,** “cast-actor”**)**
- Single pattern/relation matching model: 49.6% F_1 (vs. 52.5% F_1 Full)

Matching Questions

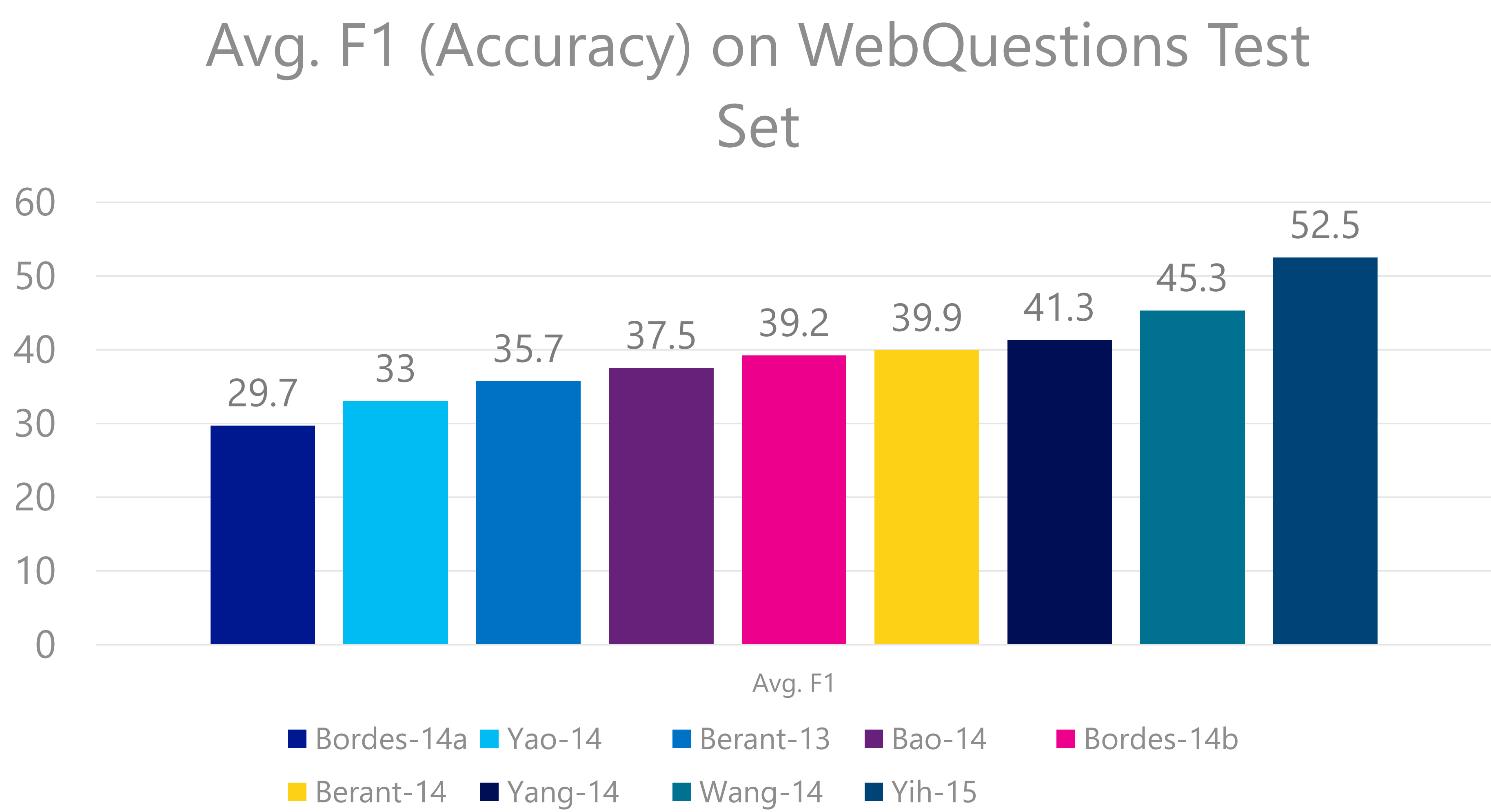
- Semantic Parsing via Paraphrasing [Berant&Liang ACL-14]



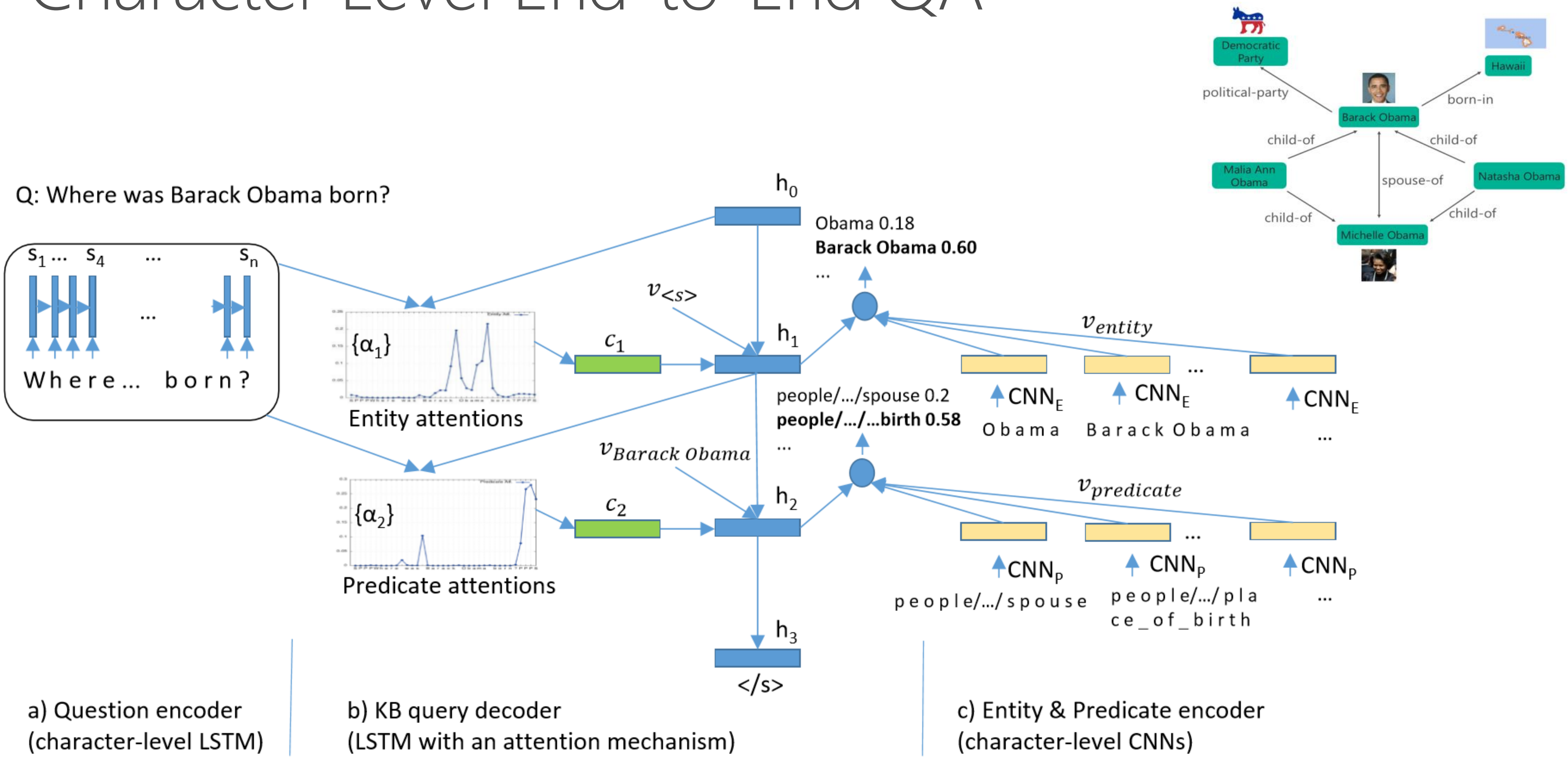
- Create phrase matching features using phrase table derived from word alignment results
- Represent questions as vectors (avg. of word vectors)

Subgraph Embedding [Bordes+ EMNLP-2014]

- Basic idea: map question and answer to vectors
  - q: question (Who did Clooney marry in 1987?)
  - a: answer candidate (K. Preston)
  - $S(q, a) = f(q)^{\mathrm{T}} g(a)$, where $f(q) = \mathbf{W}\phi(q), g(a) = \mathbf{W}\psi(a)$
- Answer candidate generation
  - Assume the topic entity (Clooney → G. Clooney) in q is given
  - All neighboring entities 1 or 2 edges away from topic entity
- Input encoding
  - $\phi(q)$: bag-of-word binary vectors
  - $\psi(a)$: binary encoding of the answer entity


Avg. F1 (Accuracy) on WebQuestions Test Set
60
50
40
30
20
10
0
29.7
33
35.7
37.5
39.2
39.9
41.3
45.3
52.5
Avg. F1
Bordes-14a
Yao-14
Berant-13
Bao-14
Bordes-14b
Berant-14
Yang-14
Wang-14
Yih-15

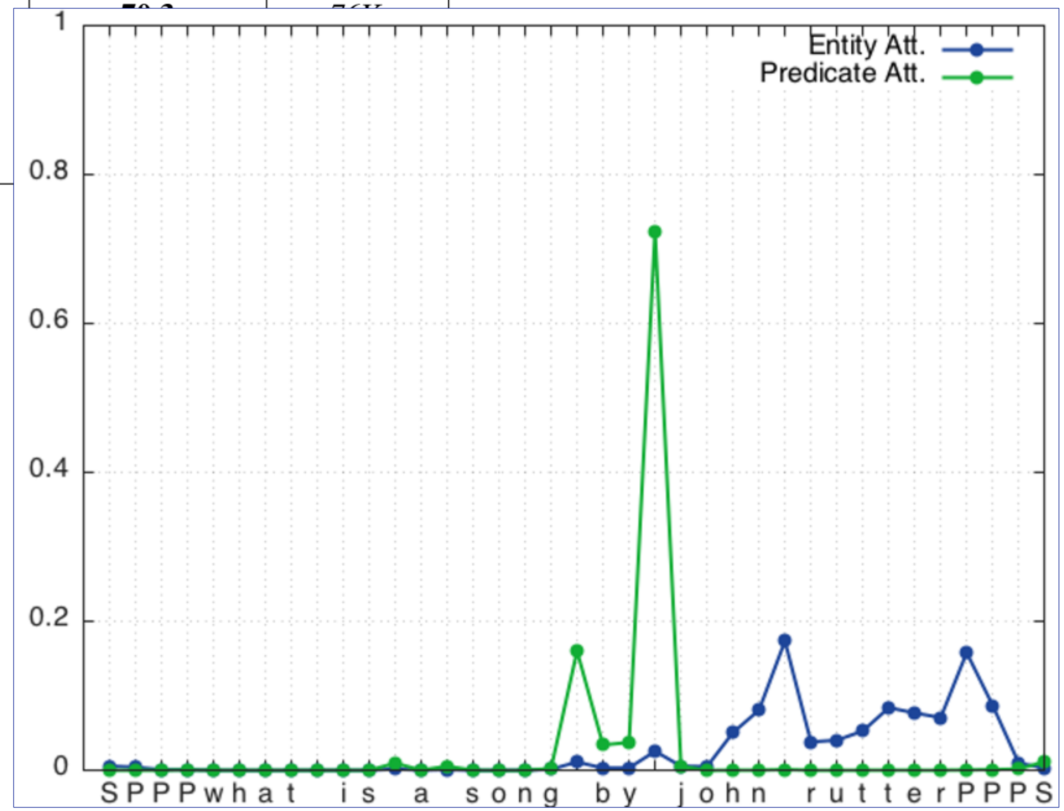
Character Level End-to-End QA



David Golub, Xiaodong He, Character-Level Question Answering with Attention, in EMNLP 2016

Experimental Results & Attention Analysis

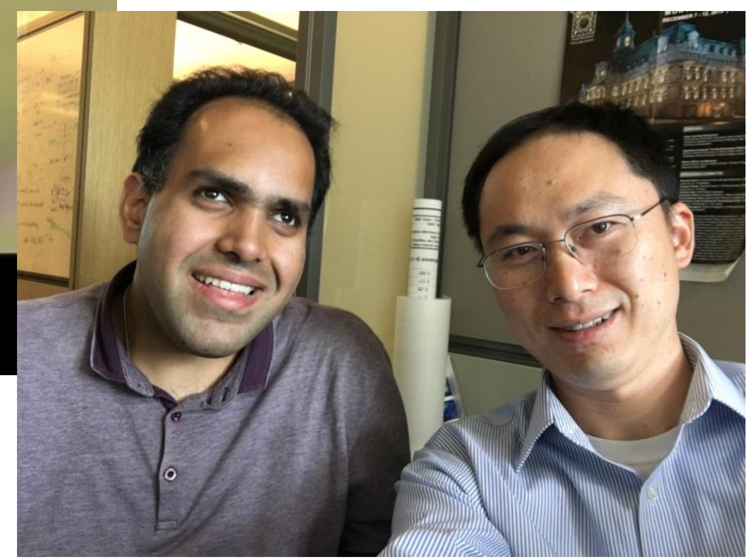
RESULTS ON SIMPLEQUESTIONS DATASET									
KB	TRAIN SOURCES WQ	SIQ	PRP	AUTOGEN. QUESTIONS	EMBED TYPE	MODEL	ENSEMBLE	SQ ACCURACY	# TRAIN EXAMPLES
FB2M	no	yes	no	no	Char	Ours	1 model	**70.9**	76K
FB2M	no	yes	no	no	Word	Ours	1 model	53.9	76K
FB2M	yes	yes	yes	yes	Word	MemNN	1 model	62.7	26M
FB5M	no	yes	no	no	Char	Ours	1 model	[illegible]	[illegible]
FB5M	no	yes	no	no	Word	Ours	1 model	[illegible]	[illegible]
FB5M	yes	yes	yes	yes	Word	MemNN	5 models	[illegible]	[illegible]
FB5M	yes	yes	yes	yes	Word	MemNN	Subgraph	[illegible]	[illegible]
FB5M	yes	yes	yes	yes	Word	MemNN	1 model	[illegible]	[illegible]



[David Golub, Xiaodong He, Character-Level Question Answering with Attention, in EMNLP 2016]

微软 Seeing AI:
计算机帮助盲人“**看见**”世界，并和世界**交流**

https://youtu.be/R2mC-NUAmMk



[Liquid pouring]

Outline

Vision

- Empower **intelligent communications** between humans, computers, and the world.
- Enable next-generation scenarios such as **universal chatbot** and **mixed reality**.

Technical Challenge

- Teach machines to perform true **perception**, **reasoning**, and **generating** responses and interpretations **across language and vision**

Driving Tasks

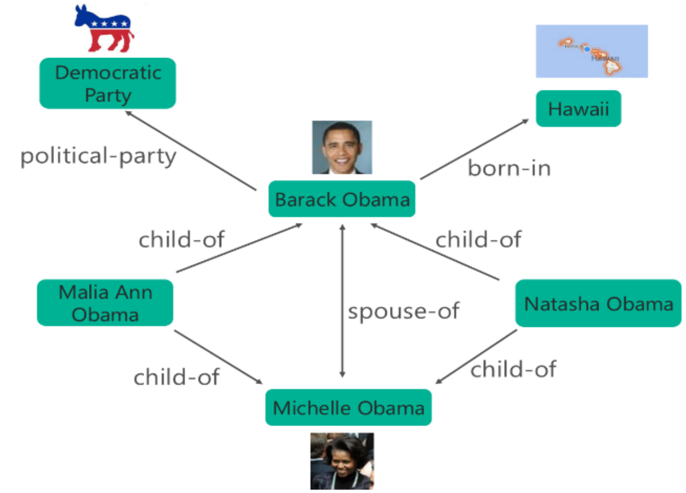
- Describe contents of images in natural language (e.g., **image captioning**)
- Answer natural language questions about images (e.g., **visual QA/Dialog**)
- Synthesize images given natural language description (e.g., **image synthesis**)

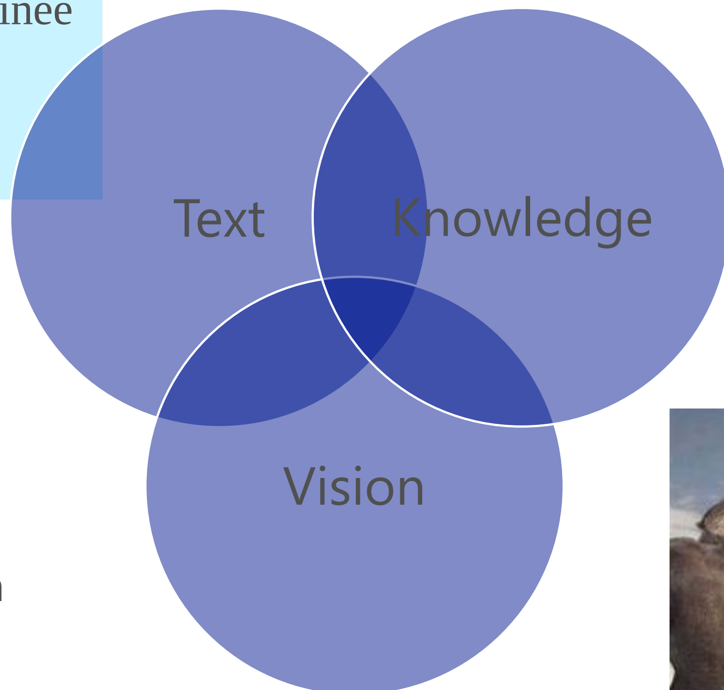
Deep Attention Mechanism

- Model the inherent **structure** of AI problems in an end-to-end framework
- Provide **interpretability** of the reasoning process in performing AI tasks

Multimodal Intelligence: process *language, vision, and knowledge* jointly

Barack Obama is an American politician serving as the 44th President of the United States. Born in Honolulu, Hawaii. In 2008, he defeated Republican nominee and was inaugurated as president on January 20, 2009.
(Wikipedia.org)





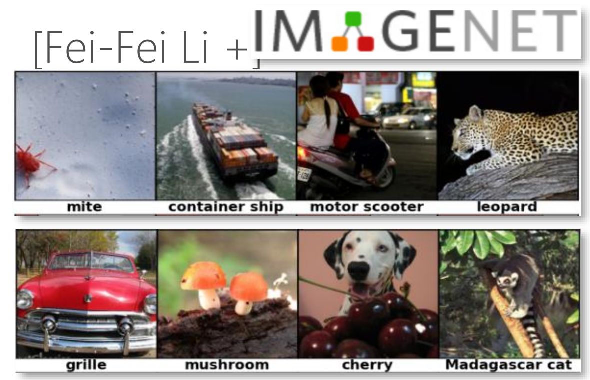
- Machine
  - concept level recognition
- Human:
  - entity level knowledge
    → **who, what, where**
  - Learning/reasoning/planning/explanation
    → **why, how, when**

http://s122.photobucket.com/user/bmeuppls/media/stampede.jpg.html

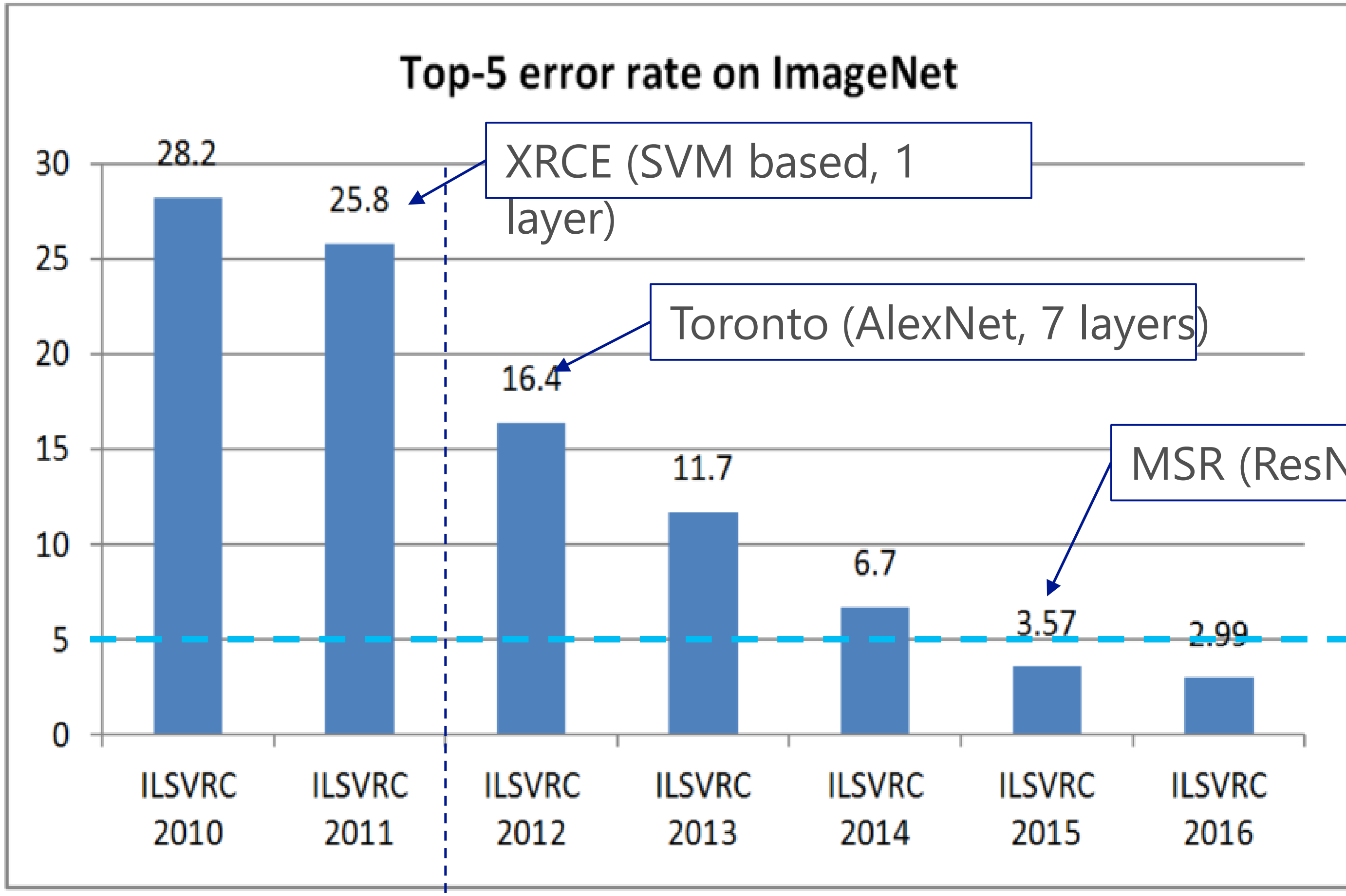
Image Object Recognition

Reached human parity on ImageNet in 2015

[Fei-Fei Li +]



(1K categories)



Human performance is about 5% error rate

pre 2012, most models are based on linear classifiers

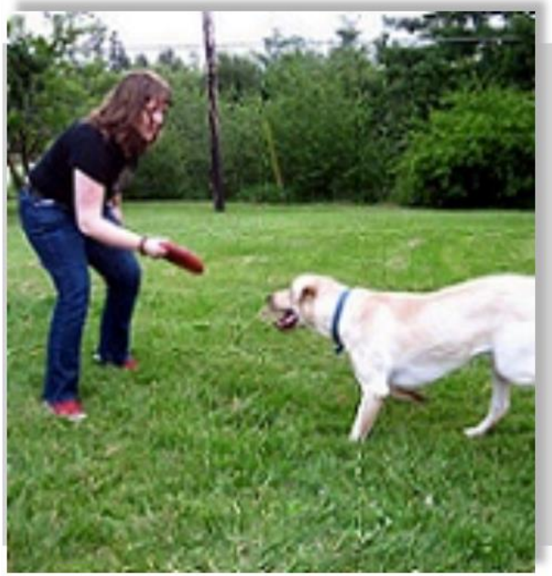
After 2012, most models are based on deep learning algorithms

However, true understanding of the world is much more challenging

Microsoft COCO
Common Objects in Context

E.g., describe the scene with natural language

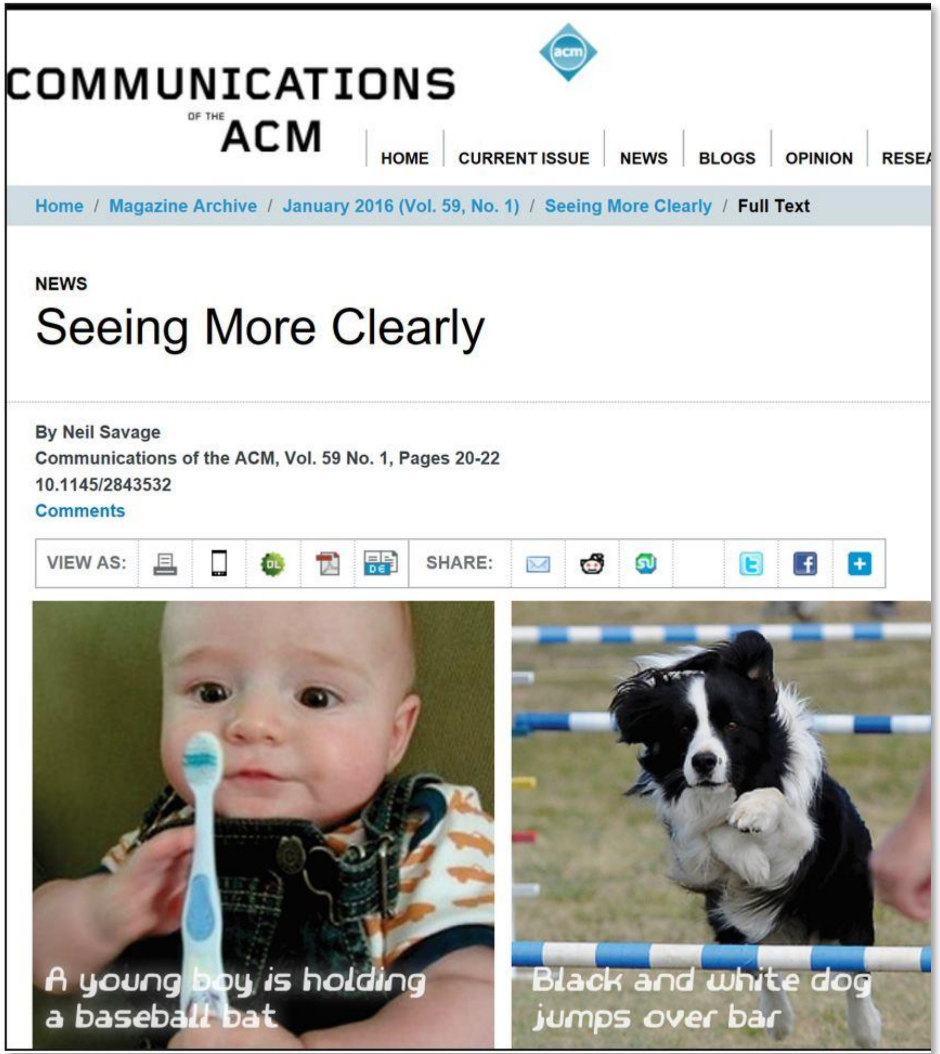
1. Understanding the image's content
2. Reasoning relationships among objects & concepts
3. Generate a story in natural language



- a woman is playing a frisbee with a dog.
- a woman is playing frisbee with her large dog.
- a girl holding a frisbee with a dog coming at her.
- a woman kneeling down holding a frisbee in front of a white dog.
- a young lady is playing frisbee with her dog.

[Lin, et al., 2014]

Problems in Vision & Language Intelligence


COMMUNICATIONS OF THE ACM

HOME | CURRENT ISSUE | NEWS | BLOGS | OPINION | RESEA

Home / Magazine Archive / January 2016 (Vol. 59, No. 1) / Seeing More Clearly / Full Text

NEWS

Seeing More Clearly

By Neil Savage

Communications of the ACM, Vol. 59 No. 1, Pages 20-22

10.1145/2843532

Comments

VIEW AS: SHARE:

CACM January 2016 (Vol. 59,No. 1)

"With careful training, these things (object recognition) actually work very well," *Rob Fergus says*

"The complete level, on par with an adult, I think is going to be a long way off," *said Fei-Fei Li*

"The overall picture should have the same semantic value as the description," *Xiaodong He says*

"If you really understood the image, you could answer a question about it." *- Richard Zemel*

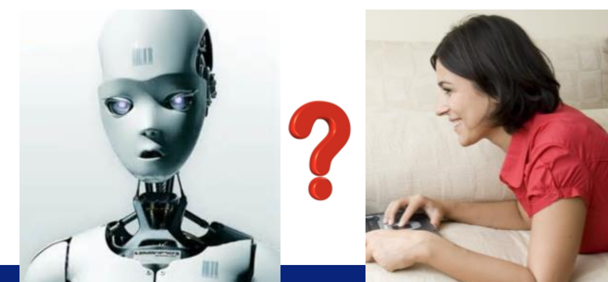
Image Captioning

describe objects, attributes, and relationship in an image, in a natural language form

a man holding a tennis racquet on a tennis court

the man is on the tennis court playing a game

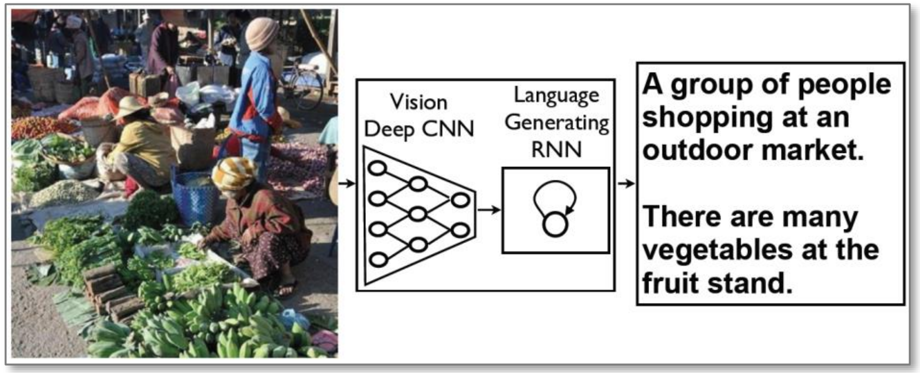
-- Let's do a Turing Test!



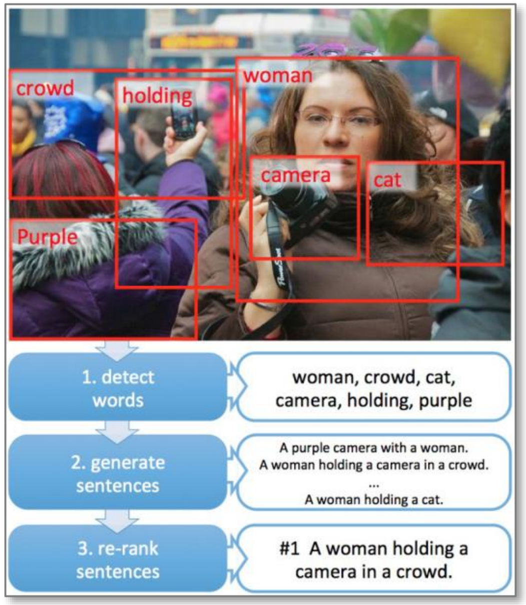
Two major paradigms for image captioning

Vector-to-Sequence

Adopted **encoder-decoder** framework from machine translation, Popular: Google, Montreal, Stanford, Berkeley



[Vinyals, Toshev, Bengio, Erhan, "Show and Tell: A Neural Image Caption Generator," CVPR, June 2015]



Compositional framework

Visual **concept detection** => caption **candidates generation** => Deep **semantic ranking**

Compositional framework can potentially exploit non paired image-caption data more effectively

[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "From Captions to Visual Concepts and Back," CVPR, June 2015]

Vector-to-Sequence Approach

E.g., Google uses a CNN to generate a whole-image feature vector, then feed it into a LSTM-based language model to generate the caption (system is an ensemble of 20 LSTMs).

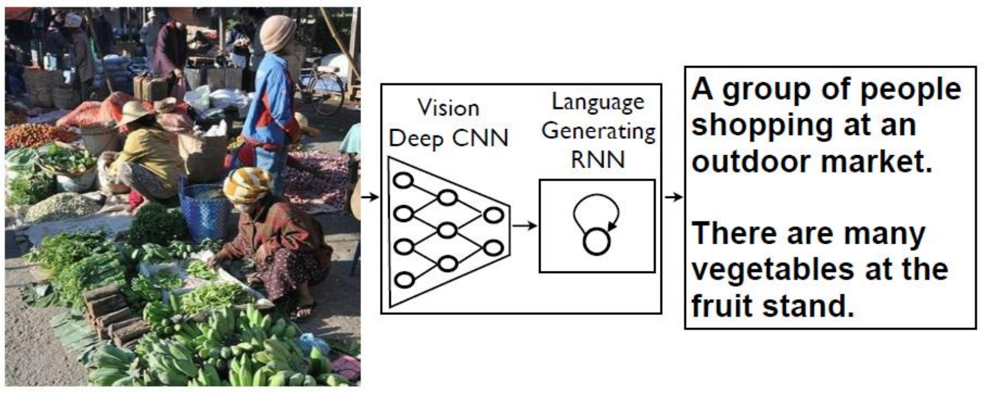


Figure 1. NIC, our model, is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. It generates complete sentences in natural language from an input image, as shown on the example above.

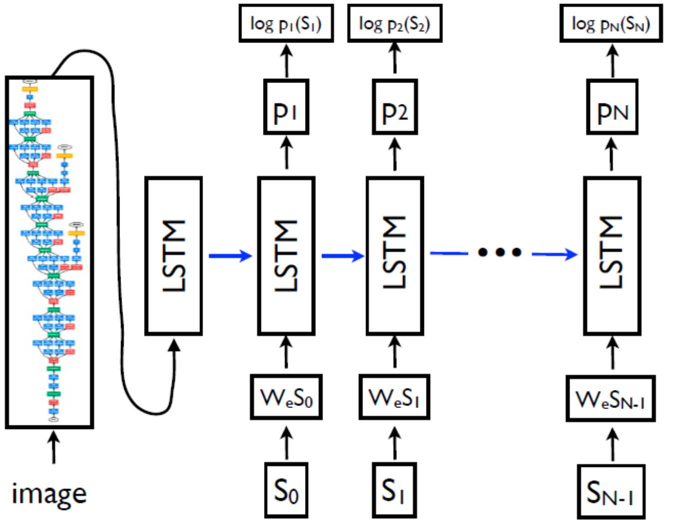


Figure 3. LSTM model combined with a CNN image embedder (as defined in [30]) and word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections in Figure 2. All LSTMs share the same parameters.

Vinyals, Toshev, Bengio, Erhan, "Show and Tell: A Neural Image Caption Generator", CVPR 2015

Compositional Approach

- Word detection
  - Deep-learned model to detect key concepts in the image
- Language model generates caption candidates
  - Maxent language model conditional on words detected from the image
- Deep multi-modal semantic model re-ranking
  - Hypothetical captions re-ranked by deep-learned multimodal semantic model looking at the entire image

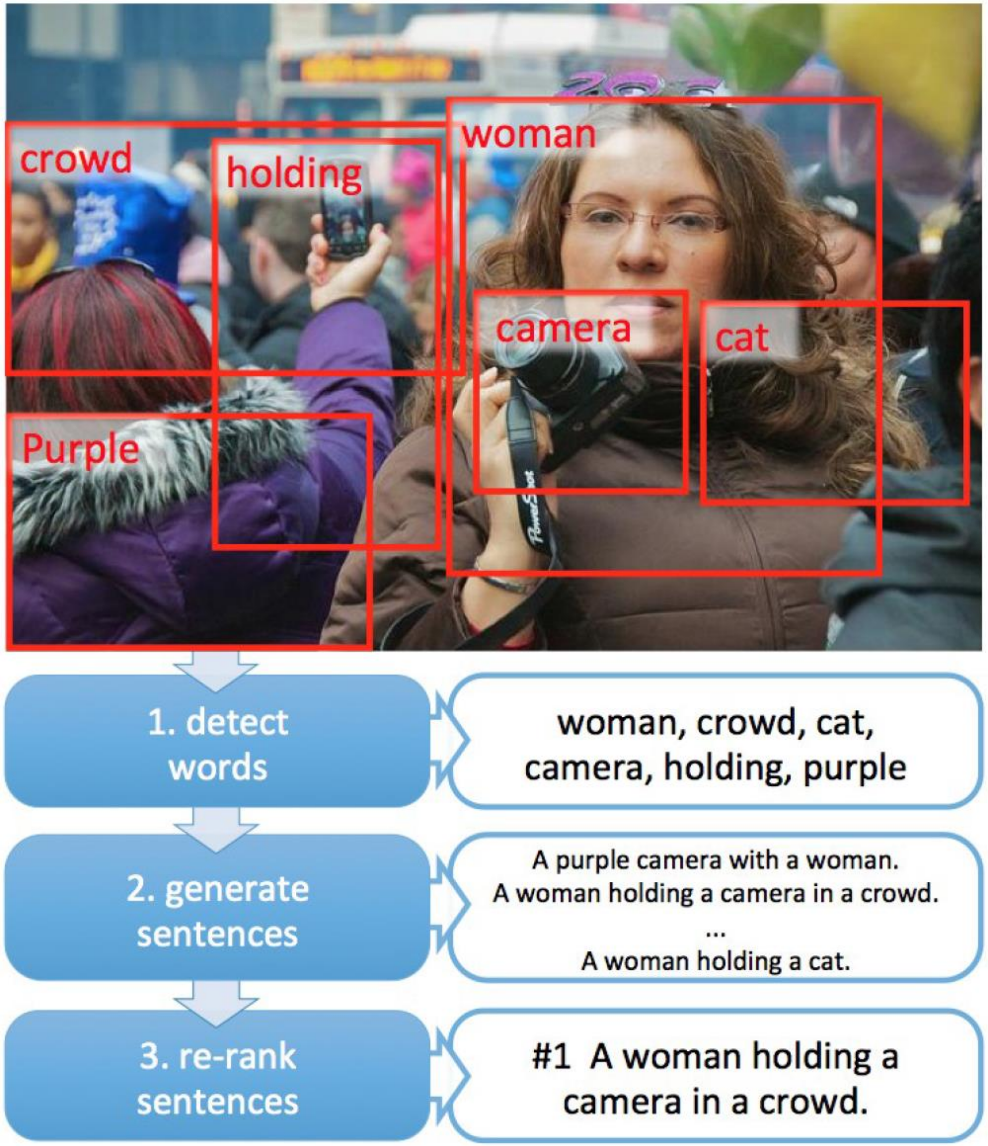


Figure 1. An illustrative example of our pipeline.

[Fang, Gupta, Iandola, Srivastava, Deng, Dollar, Gao, He, Mitchell, Platt, Zitnick, Zweig, "From Captions to Visual Concepts and Back," CVPR, 2015]

Detecting Tags: Attend on key visual concepts

- Treat training caption as bag of image labels
- Train one binary classifier per label on all images
- "Noisy-Or" classifier
  - Image divided into 12x12 overlapping regions
  - fc7 vector used for image features

$$p_i^w = 1 - \prod_{j \in r_i} \left(1 - \overbrace{\sigma(f_{ij} \cdot v_w)}^{p(w\ in\ r_j\ of\ image\ i)}\right)$$

i = image id
r_i = regions
f_{ij} = fc7 vector
v_w = learned classifier weights
$\sigma(x)$ = sigmoid

e.g., the visual "attention" of word **sitting**.

sitting

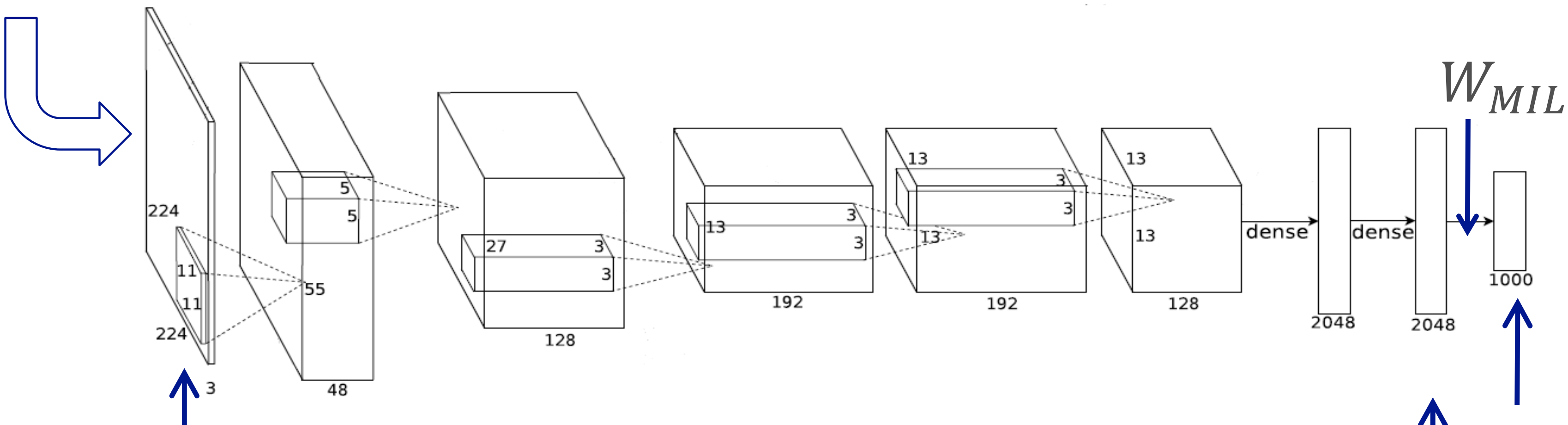
sitting (0.83)

Map of Attention: h$(x, y) = \sum_{r_i, s.t., (x,y) \in r_i} \sigma(f_{ij} \cdot v_{sitting})$

Multiple Instance Learning

a man sitting on a chair with a dog on his lap

$$\vec{P}(w\ in\ region) = 1/(1 + e^{W_{MIL} \times v_{fc7}})$$



Raw pixels from input box

fc7 MIL layer

Tuned image features from AlexNet (Krizhevsky et al., 2012) or other CNNs.

Generating Caption Candidates

a kitchen with wooden

Language model

cabinets

MaxEnt LM

$p(cabinets|with\ wooden)$

a kitchen with wooden cabinets

room
~~wooden~~
~~kitchen~~
sink stove ~~cabinets~~
floor

Image

Beam search to generate 500 candidates

1. wooden cabinets in a kitchen
2. a sink and cabinets

...

500. a room with stove on the floor

DSSM: Bridge the gap between image and language!

The multimodal deep structured semantic model projects images and captions to a common semantic space.

Q = image, D = caption, R = relevance

Relevance: $R(Q,D) = \text{cosine}(y_Q, y_D) = \frac{y_Q{}^T y_D}{\|y_Q\| \|y_D\|}$

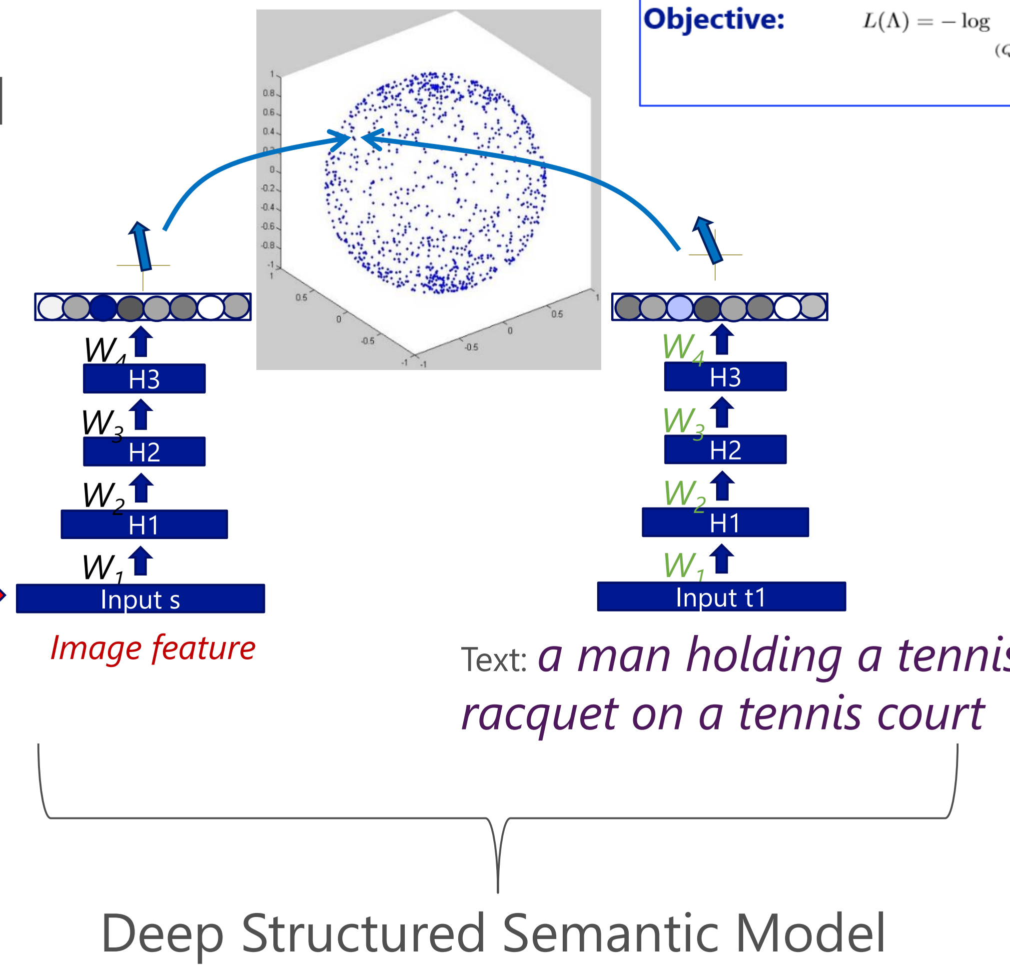
Caption probability: $P(D|Q) = \frac{\exp(\gamma R(Q,D))}{\Sigma_{D' \in \mathbf{D}} \exp(\gamma R(Q,D'))}$

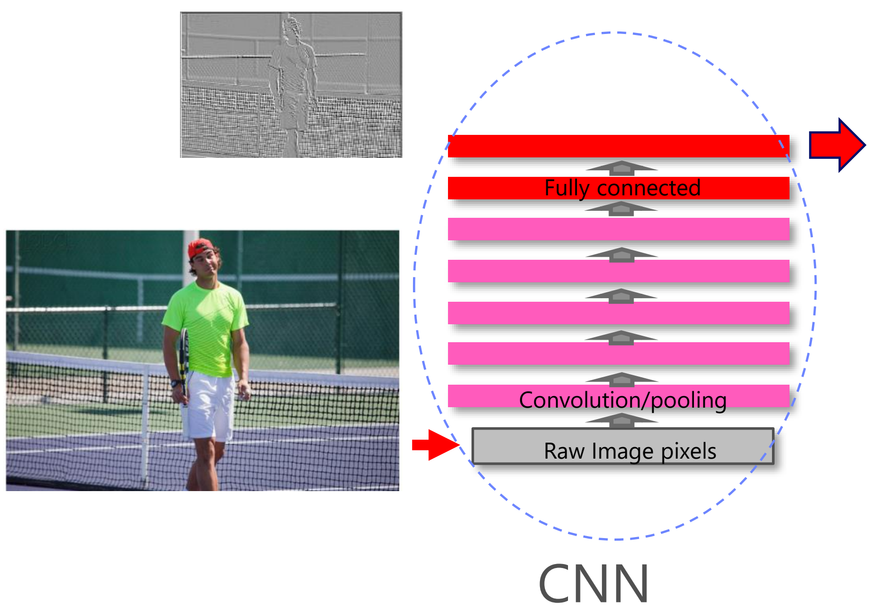
Candidate captions

Smoothing factor

Objective: $L(\Lambda) = -\log \prod_{(Q,D^+)} P(D^+|Q)$

Correct caption

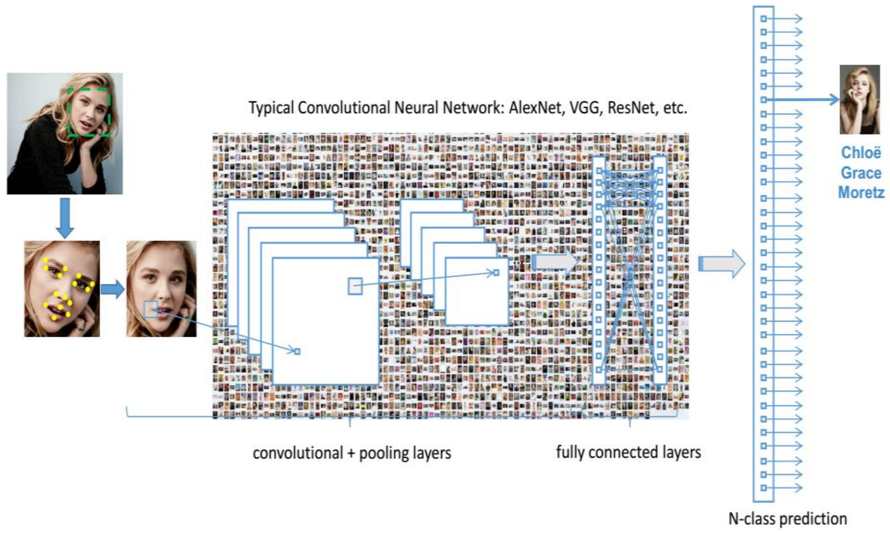




[He, Gao, Deng et al., 2013, 2014, 2015]

Know the Entities

- Recognize entities in images (celebrities, landmarks)



World largest set of celebrities

Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background.

[Guo, Zhang, Hu, He, Gao, MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition, ECCV 2016]

[Tran, He, Zhang, Sun, Carapcea, Thrasher, Buehler, Sienkiewicz, Rich Image Captioning in the Wild, DeepVision, CVPR 2016]

COCO Image Captioning Challenge

Human judgment is the ultimate metric

Turing Test etc. at the MS COCO Image Captioning Challenge 2015

MSR won the 1st prize, tied with Google.

Microsoft COCO
Common Objects in Context

	Official Rank	% of captions that pass the Turing Test	% of captions that are better or equal to human's
MSR	1st	32.2%	26.8%
Google	1st	31.7%	27.3%
MSR Captivator	3rd	30.1%	25.0%
Montreal/Toronto	3rd	27.2%	26.2%
Berkeley LRCN	5th	26.8%	24.6%
Other groups: Baidu/UCLA, Stanford, Tsinghua, etc.			
Human	--	67.5%	63.8%

human generated caption

Visualize the Dynamic Attention Map

a baseball player throwing a ball

baseball (1.00)

a **baseball**

Attention heatmap provides grounded evidence for interpretation

Visualize the Dynamic Attention Map

a baseball player throwing a ball

player (1.00)

a baseball **player**

Attention heatmap provides grounded evidence for interpretation

Visualize the Dynamic Attention Map

a baseball player throwing a ball

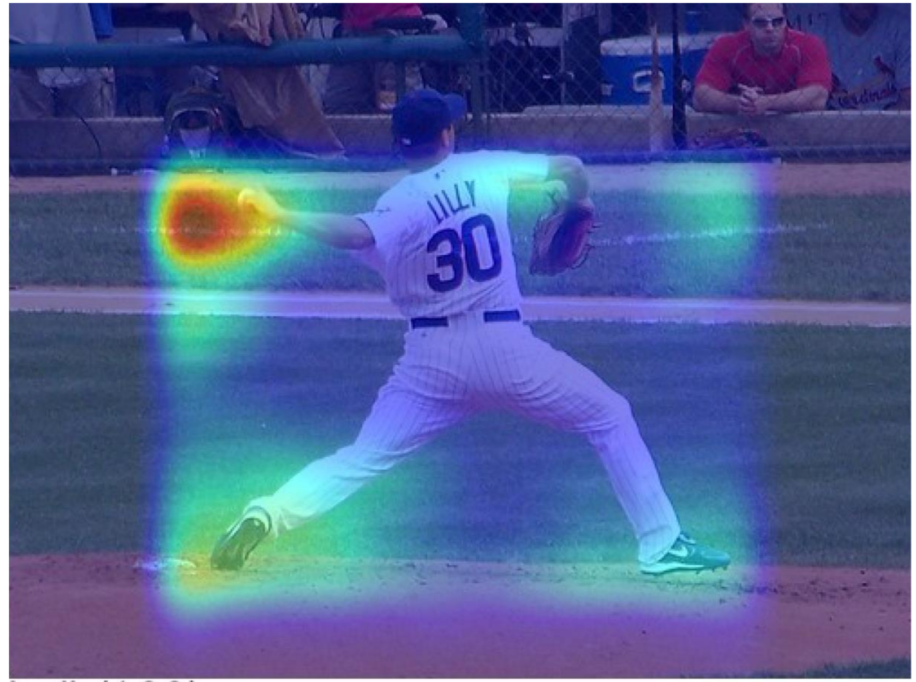
throwing (0.86)

a baseball player **throwing**

Attention heatmap provides grounded evidence for interpretation

Visualize the Dynamic Attention Map

a baseball player throwing a ball



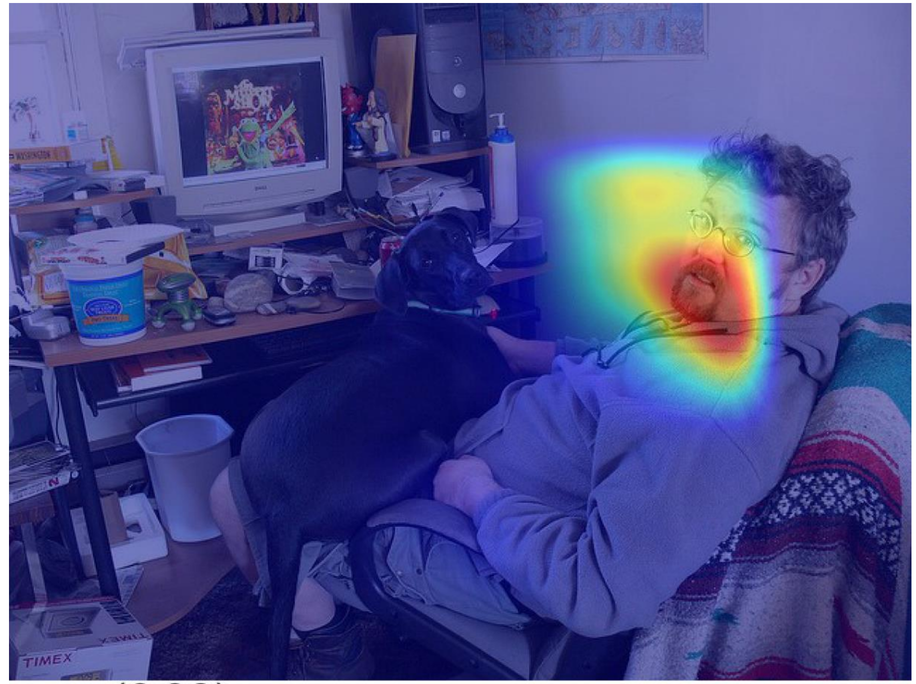
ball (1.00)

a baseball player throwing a **ball**

Attention heatmap provides grounded evidence for interpretation

Visualize the Dynamic Attention Map

a man sitting in a couch with a dog



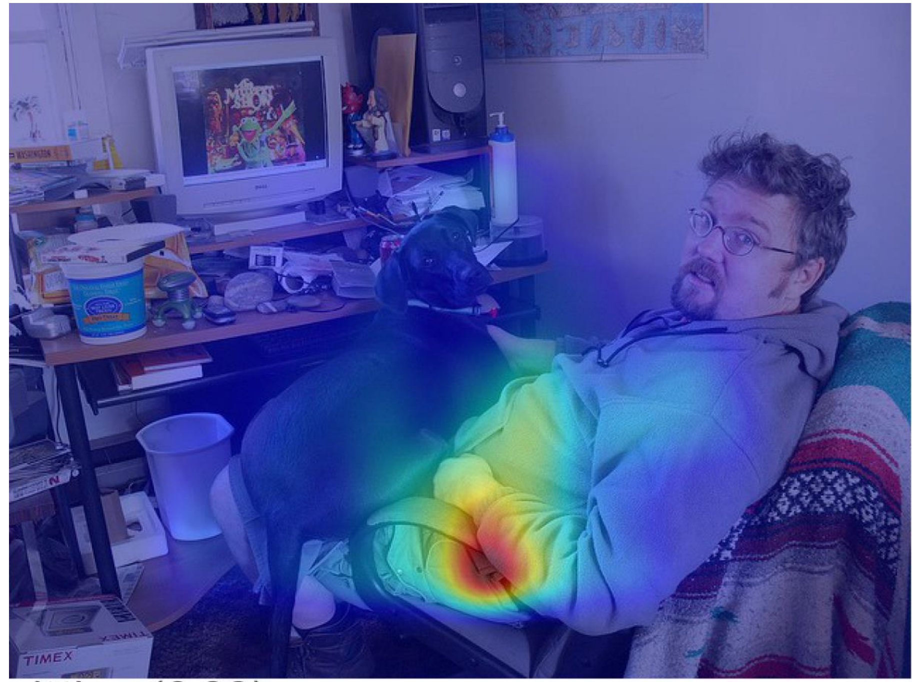
man (0.93)

a **man**

Attention heatmap provides grounded evidence for interpretation

Visualize the Dynamic Attention Map

a man sitting in a couch with a dog



sitting (0.83)

a man **sitting**

Attention heatmap provides grounded evidence for interpretation

Visualize the Dynamic Attention Map

a man sitting in a couch with a dog

couch (0.66)

a man sitting in a **couch**

Attention heatmap provides grounded evidence for interpretation

Visualize the Dynamic Attention Map

a man sitting in a couch with a dog

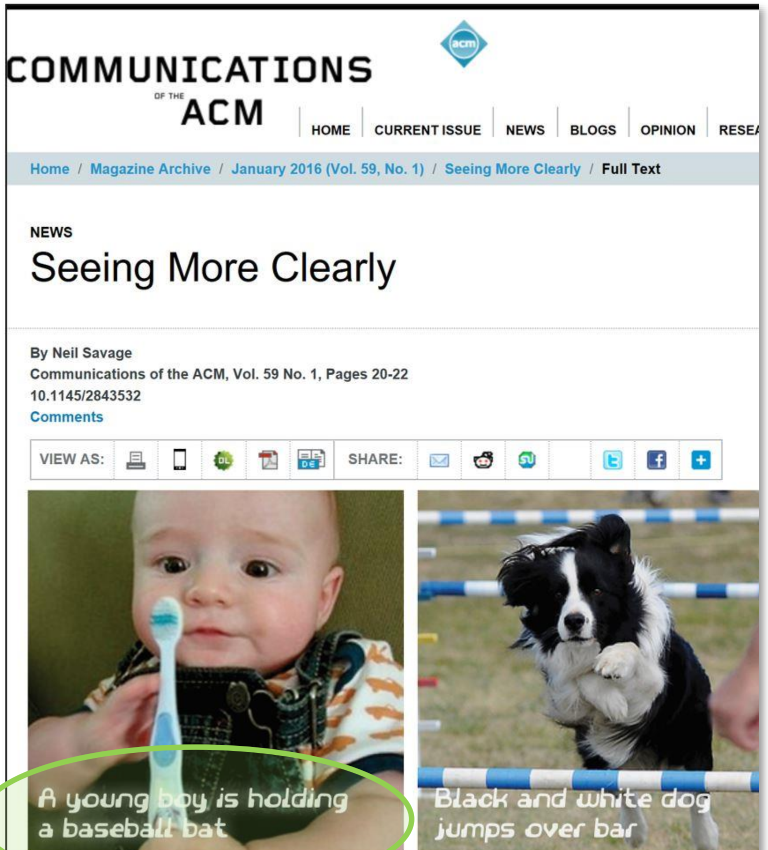
dog (1.00)

a man sitting in a couch with a **dog**

Attention heatmap provides grounded evidence for interpretation

Recent Trend: Interpretability & Semantic Control

One year ago



Now

Detected semantic concepts:
person (0.998), baby (0.983), holding (0.952), small (0.697), sitting (0.638), toothbrush (0.538), child (0.502), mouth (0.438)

Semantic composition:
1. Only using "**baby**": *a baby in a*
2. Only using "**holding**": *a person holding a hand*
3. Only using "**toothbrush**": *a pair of toothbrush*
4. Only using "**mouth**": *a man with a toothbrush*
5. Using "**baby**" and "**mouth**": *a baby brushing its teeth*

Overall caption generated by the SCN:
a baby holding a toothbrush in its mouth

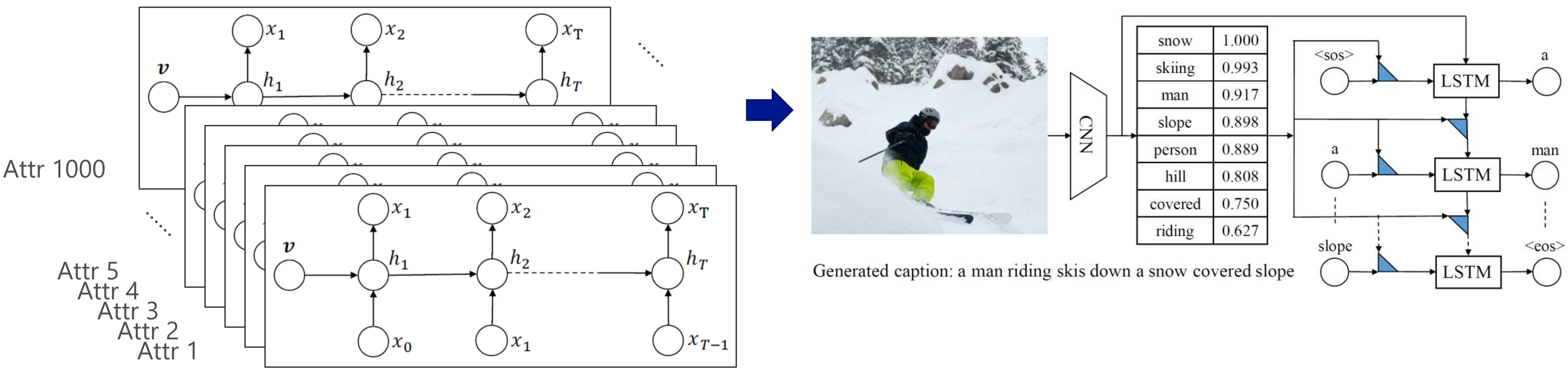
Influence the caption by changing the tag:
6. Replace "**baby**" with "**girl**": *a little girl holding a toothbrush in her mouth*
7. Replace "**toothbrush**" with "**baseball**": *a baby holding a baseball bat in his hand*
8. Replace "**toothbrush**" with "**pizza**": *a baby holding a piece of pizza in his mouth*

[Gan, et al., ***Semantic Compositional Net***, CVPR17]

Semantic Compositional Networks

A very wide Model (as wide as 1000 LSTM slices)

- Conceptually, learn 1000 LSTMs, one for each semantic attributes.
- Combine these 1000 LSTMs, weighted by attributes' likelihood.
- Run tensor decomposition to reduce # parameters to fit in GPU



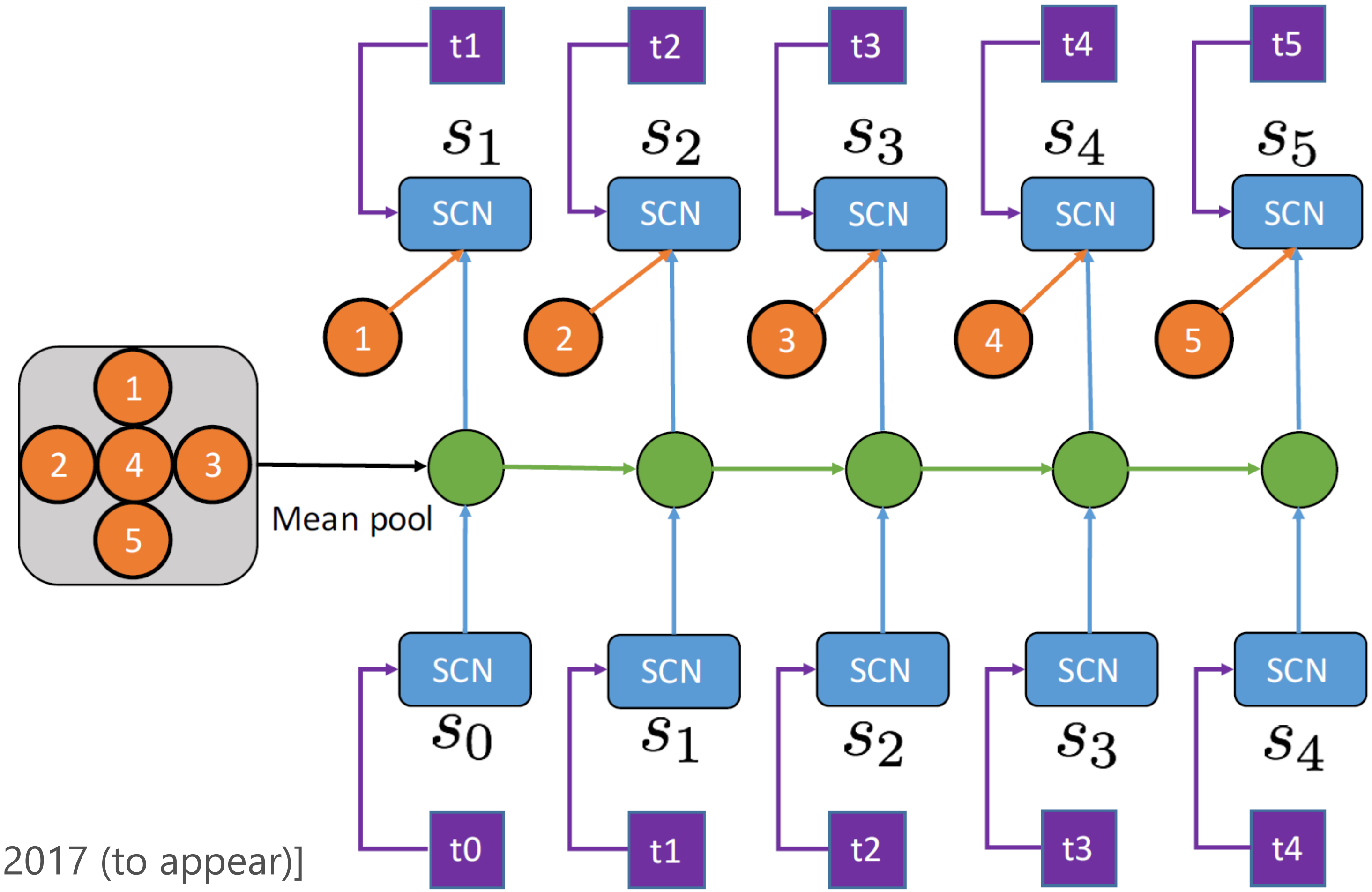
Generated caption: a man riding skis down a snow covered slope

Recent Trend: storytelling with a consistent theme

RL: We went to the carnival today. There were many different kinds of cool things. Some kids were playing in the air. There was a giant dragon. At the end of the day, the kids played a lot of fun.

RL: The bride and groom were very happy. They were all happy to be married. The family was happy for the ceremony. At the end of the day they all posed for a picture. At night , the couple had a great time dancing.

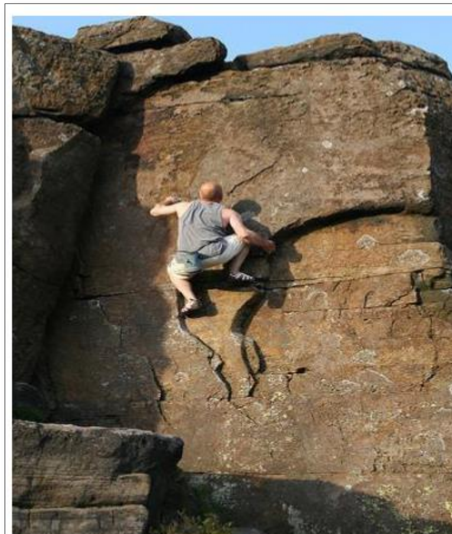
Hierarchical Reinforcement Learning for Visual Storytelling



[Gan, et al., 2017 (to appear)]

Recent Trend: Style Control in Captioning

StyleNet:
Control the style of captions



CaptionBot: A man on a rocky hillside next to a stone wall.

Romantic: A man uses rock climbing to overcome the obstacle in the life.

Humorous: A man is climbing the rock like a lizard.

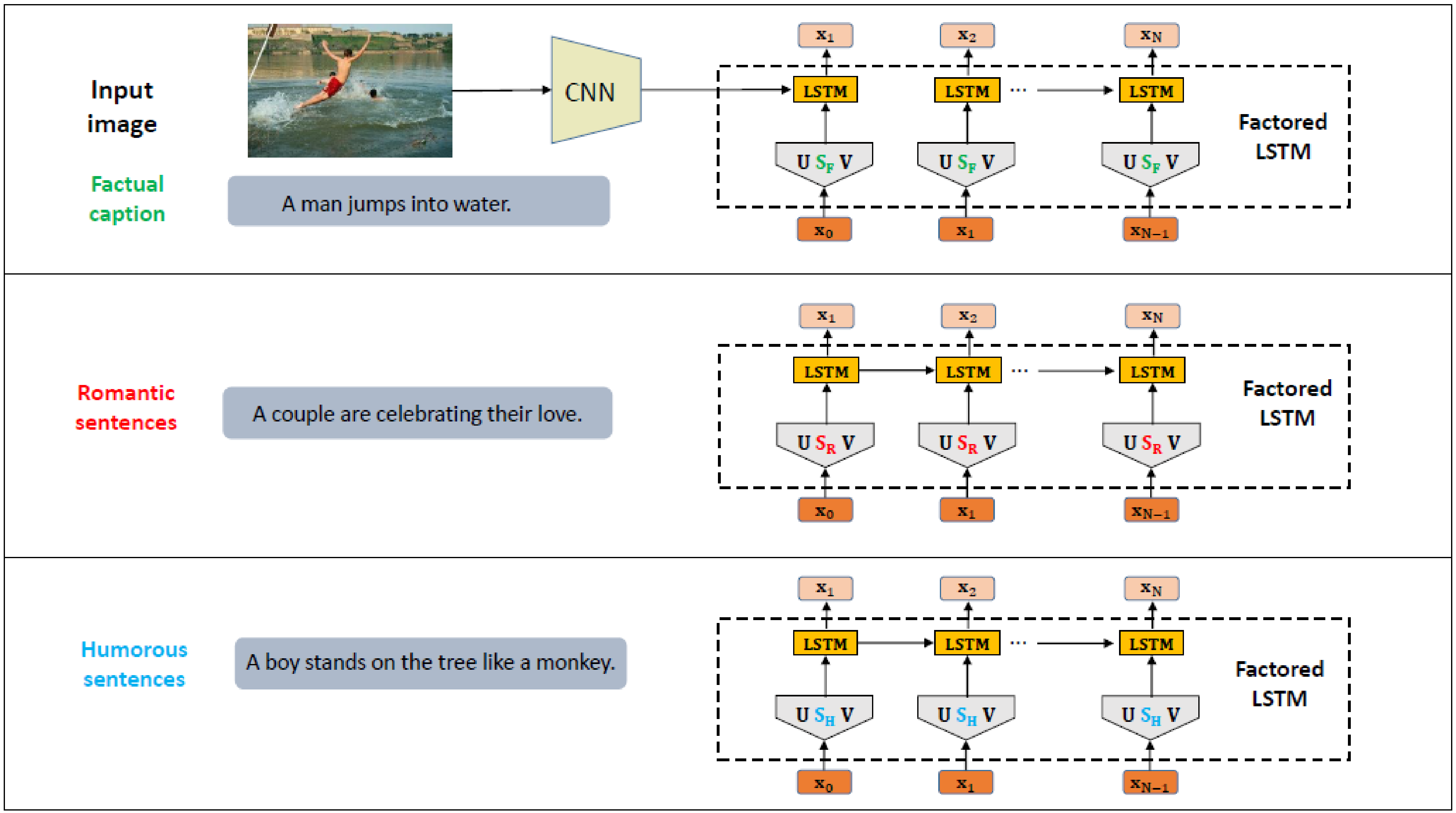
CaptionBot: A dog runs in the grass.

Romantic: A dog runs through the grass to meet his lover.

Humorous: A dog runs through the grass in search of the missing bones.

[Gan, et al., ***StyleNet***, CVPR17]

Disentangle specific styles from generic linguistic structure



Add emotion in language expression

Recognizing: *outdoor, woman, grass*

Captioning: a woman wearing a blue shirt.

Commenting: gorgeous and beautiful as an angel！

Recognizing: *indoor, dog, woman*

Captioning: a woman and a dog posing for the camera.

Commenting: awww so cute, I mean the dog 😁

Recognizing: *tattoo, foot*

Captioning: a tattoo on display.

Commenting: fabulous 👍

Deploy the **CaptionBot** in the real world

http://CaptionBot.ai & Cloud Computer Vision API Service released to Public

Microsoft
Cognitive Services

BUSINESS INSIDER **Microsoft's newest bot offered a spot-on caption to this photo of Satya Nadella**

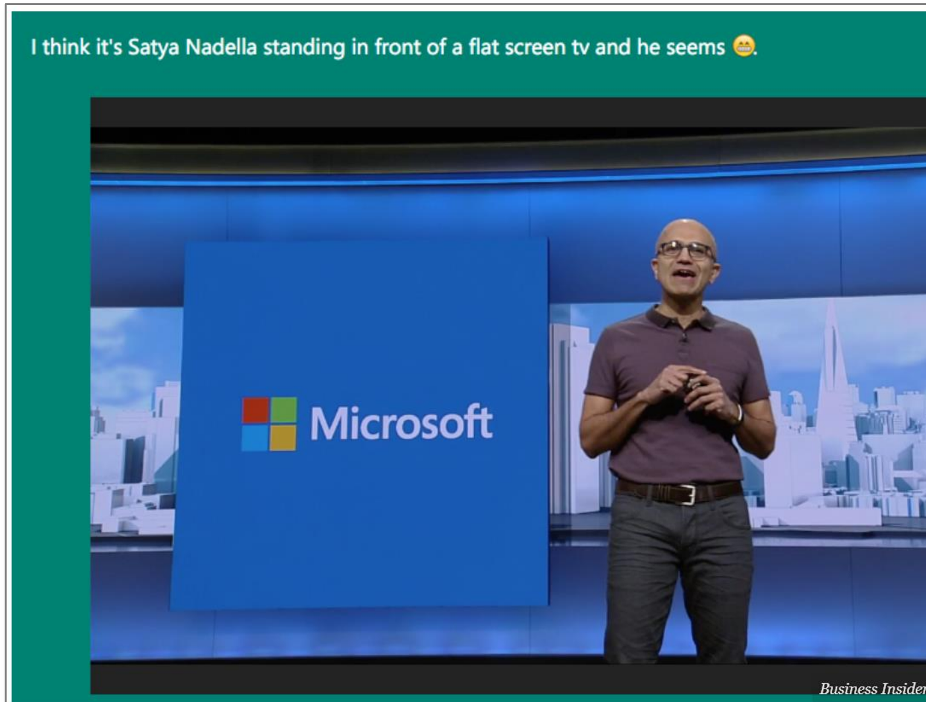


Photo of Satya Nadella caption by Microsoft's new Caption Bot

CaptionBot says: *I think it's Satya Nadella standing in front of a flat screen tv and he seems "happy".*

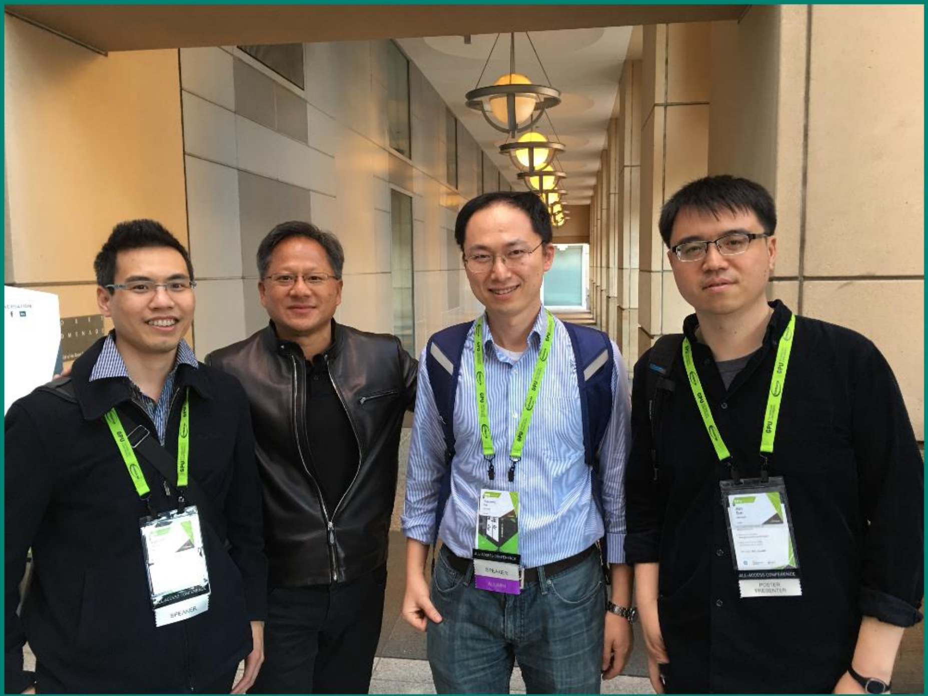
More Examples from CaptionBot

I think it's a group of football pla
field.

I think it's a man riding a
over an obstacle.

CaptionBot

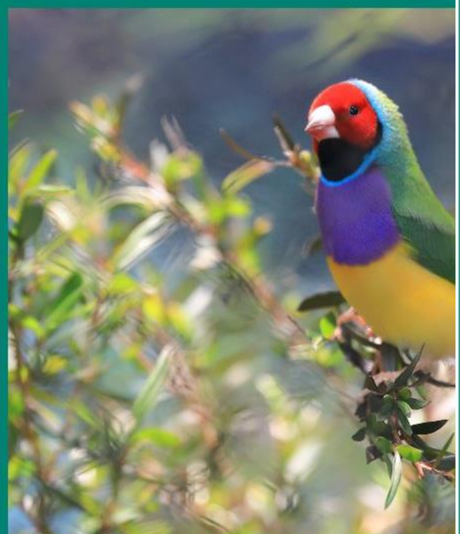
I think it's Jen-Hsun Huang, Xiaodong He, Jian Sun et al. that are posing for a picture and they seem 😐😁😁😁.



More Examples

I think it's a mar
on the beach.

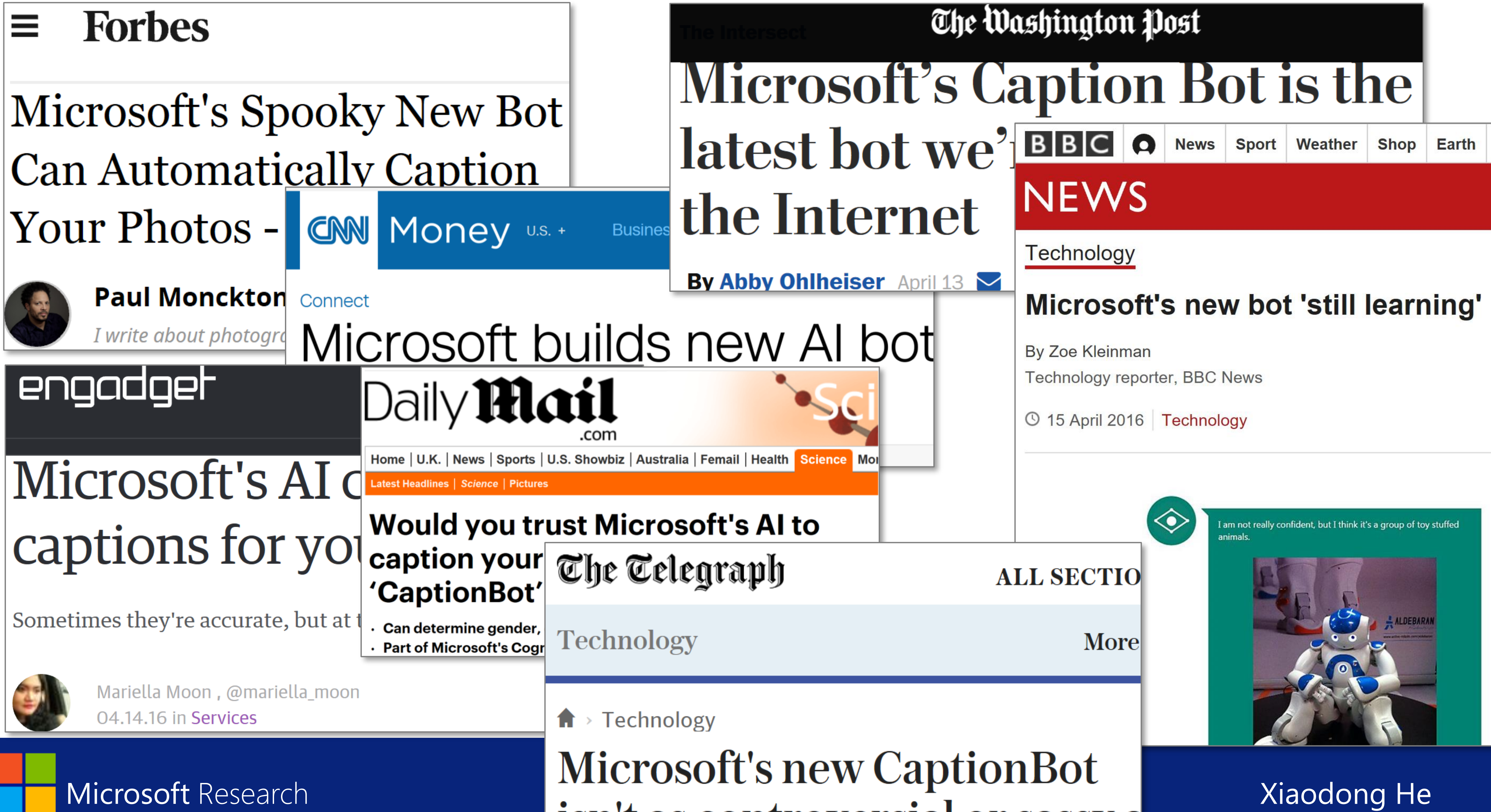
I think it's a colorful bird perched on a
tree branch.



I think it's a boat that is l
in a city.

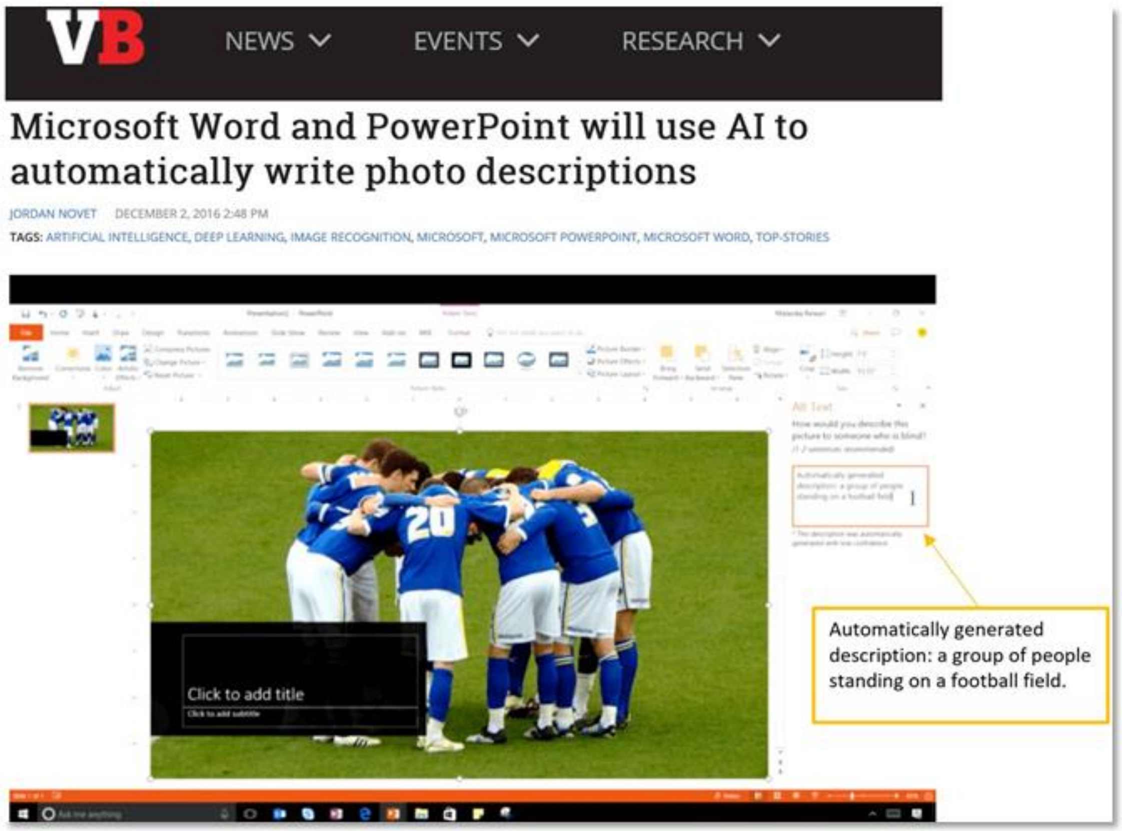
I think it's a little boy sitting in front of a birthday cake and he seems 😁.

Millions of Data Collected World-wide



Help people in the real world

Released in Office, serve millions requests daily. Also shipped Seeing AI app



http://www.microsoft.com/en-us/seeing-ai/

From Captioning to Question Answering

- Answer natural language questions according to the content of a reference image.

Question:
What are sitting in the basket on a bicycle?

Image Question Answering (IQA)

Answer:
dogs

Caption vs. QA: need reasoning

To answer a question about a image:

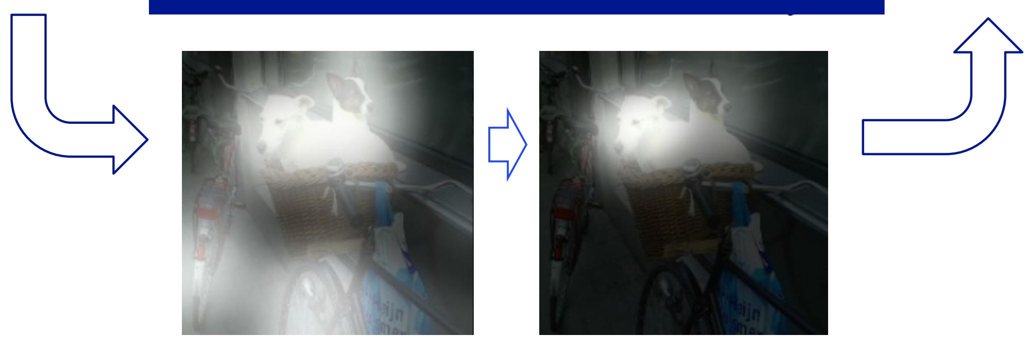
Need to understand subtle relationships among multiple objects

Need to focus on the specific regions that are relevant to the answer.

Question: What are sitting in the basket on a bicycle?

Multiple-steps of reasoning over the image to infer the answer

Answer: dogs

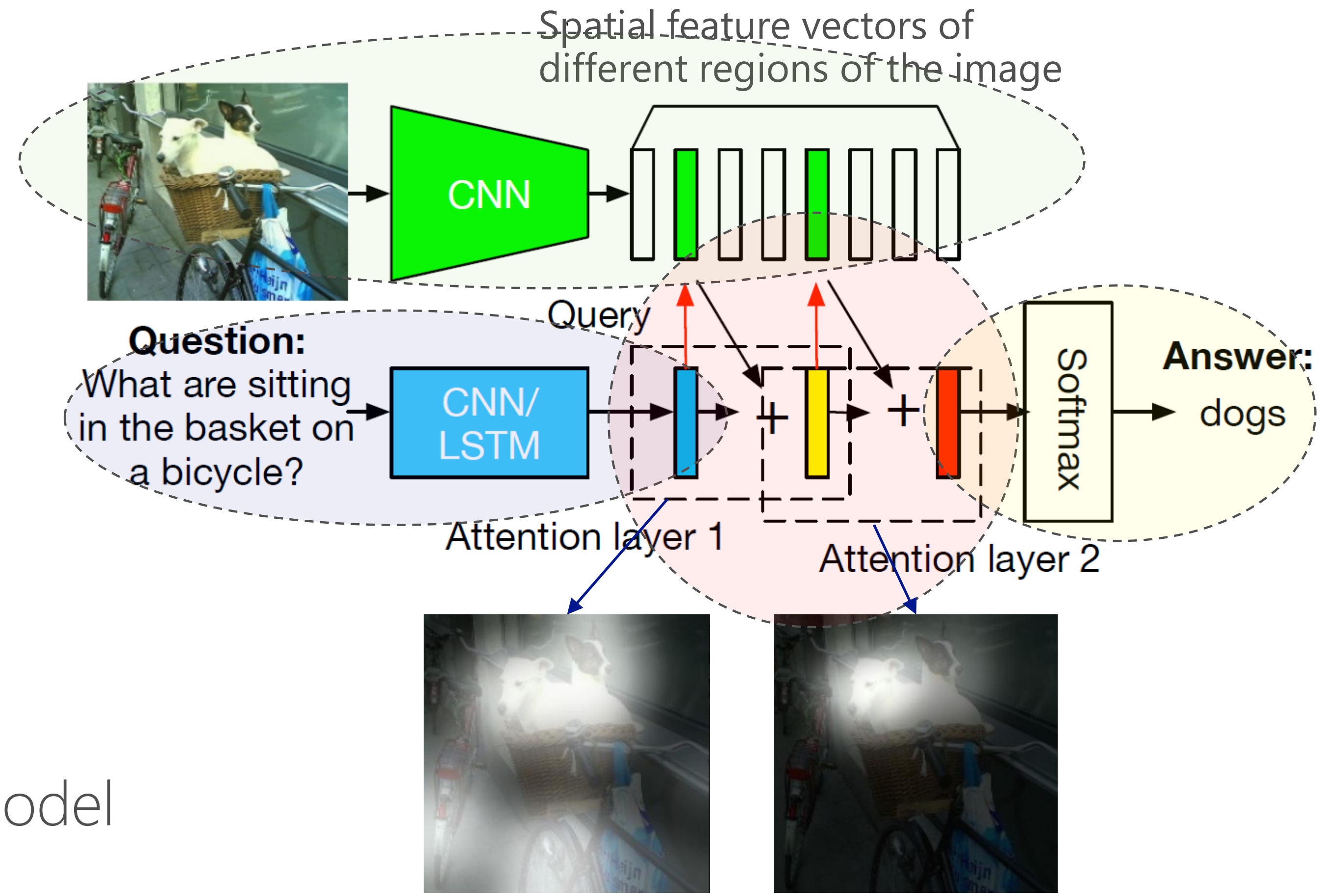


Next: Multimodal Reasoning / QA

[Stacked Attention Networks, Yang, He, Gao, Deng, Smola, CVPR 2016]

SANs perform multi-step reasoning

1. Question model
2. Image model
3. Multi-level attention model
4. Answer predictor
5. End-to-end learning using SGD



1. The image model in the SAN

- Image Model

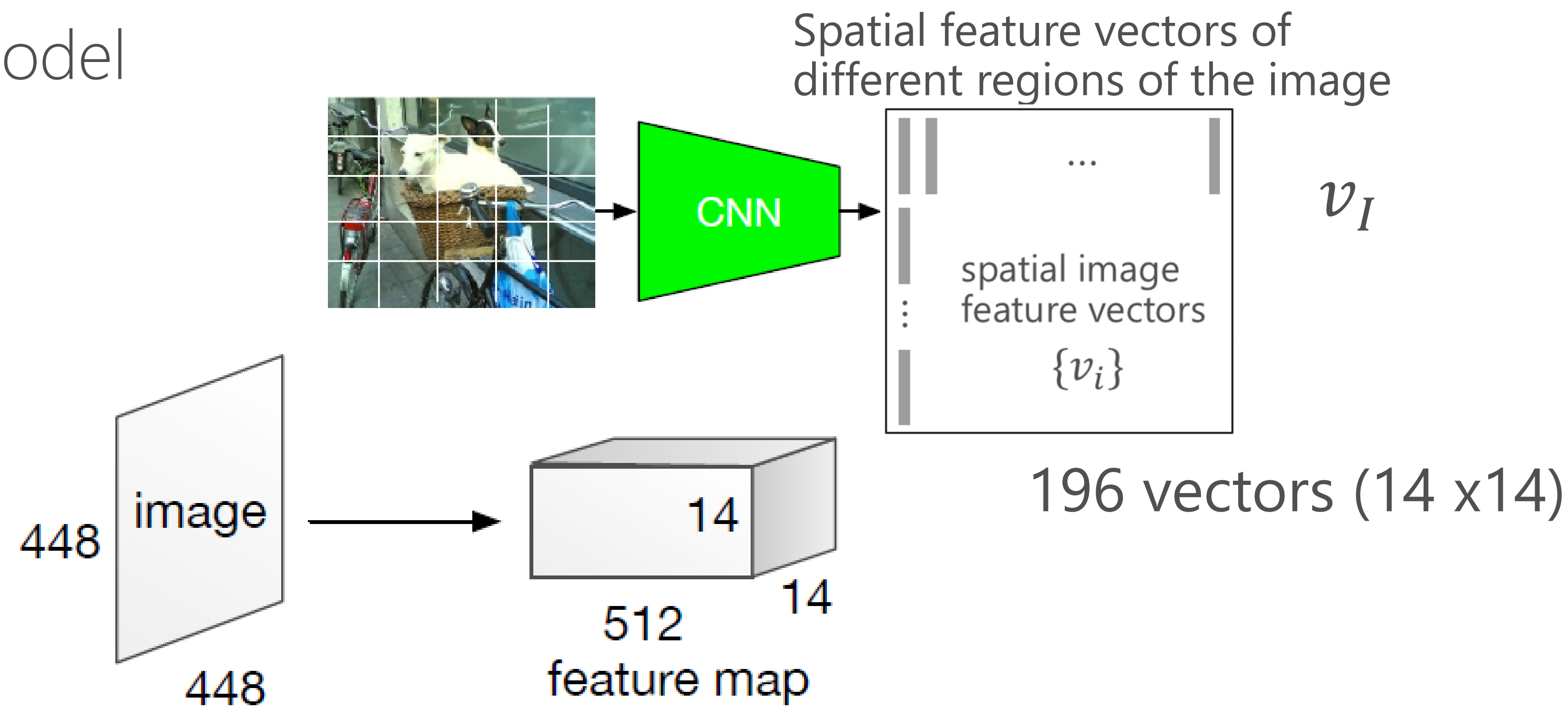


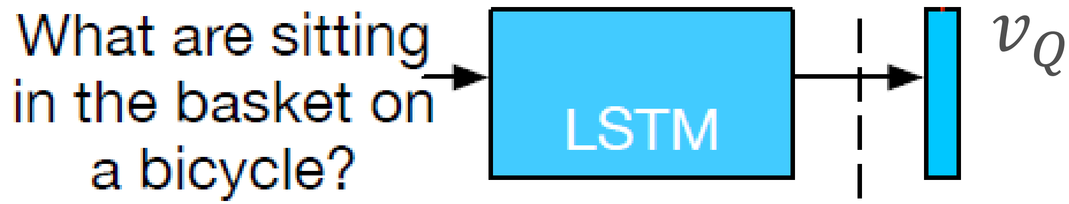
Figure 2: CNN based image model

$$f_I = \text{CNN}_{vgg}(I). \quad v_I = \tanh(W_I f_I + b_I)$$

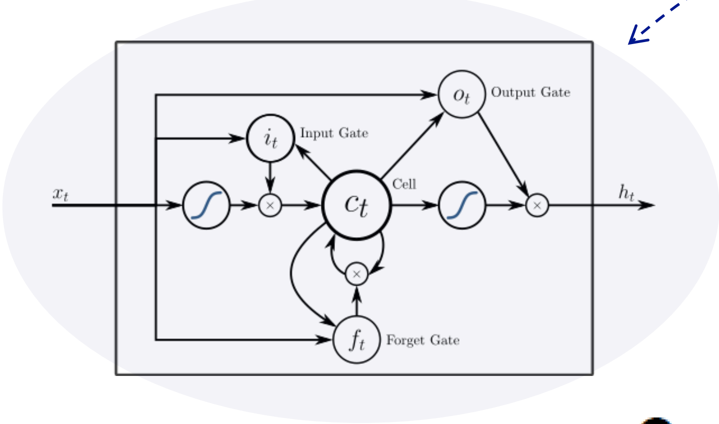
2. The question model in the SAN

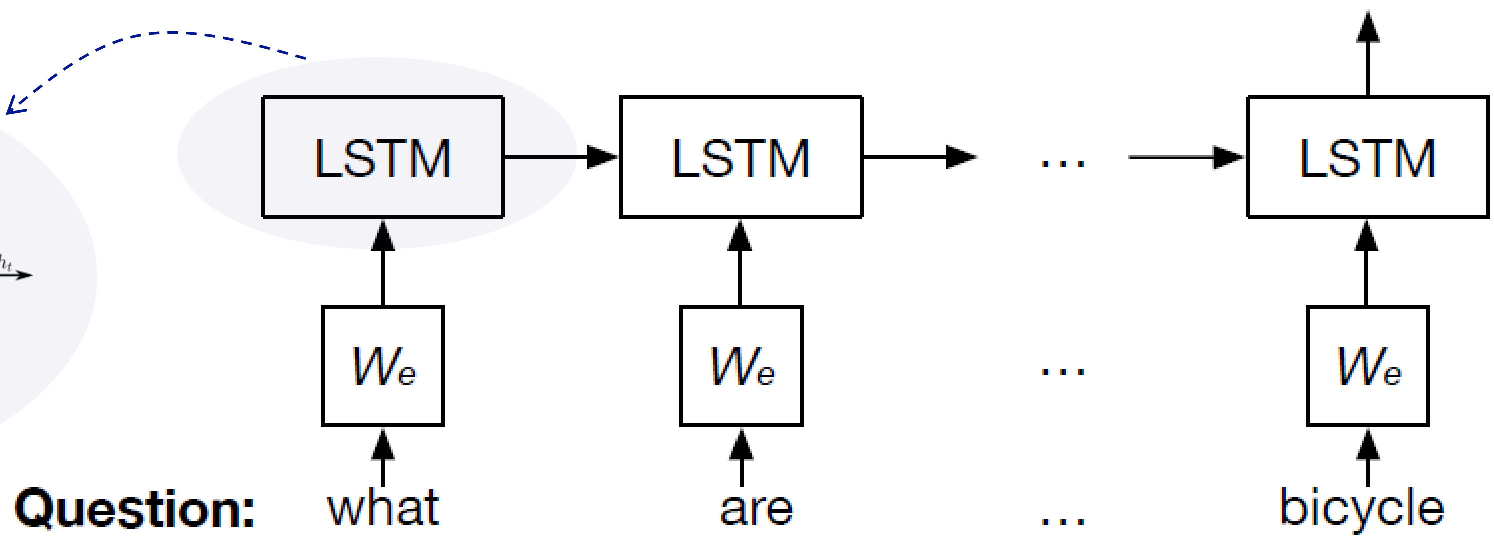
- Question Model

  Code the question into a vector using a LSTM

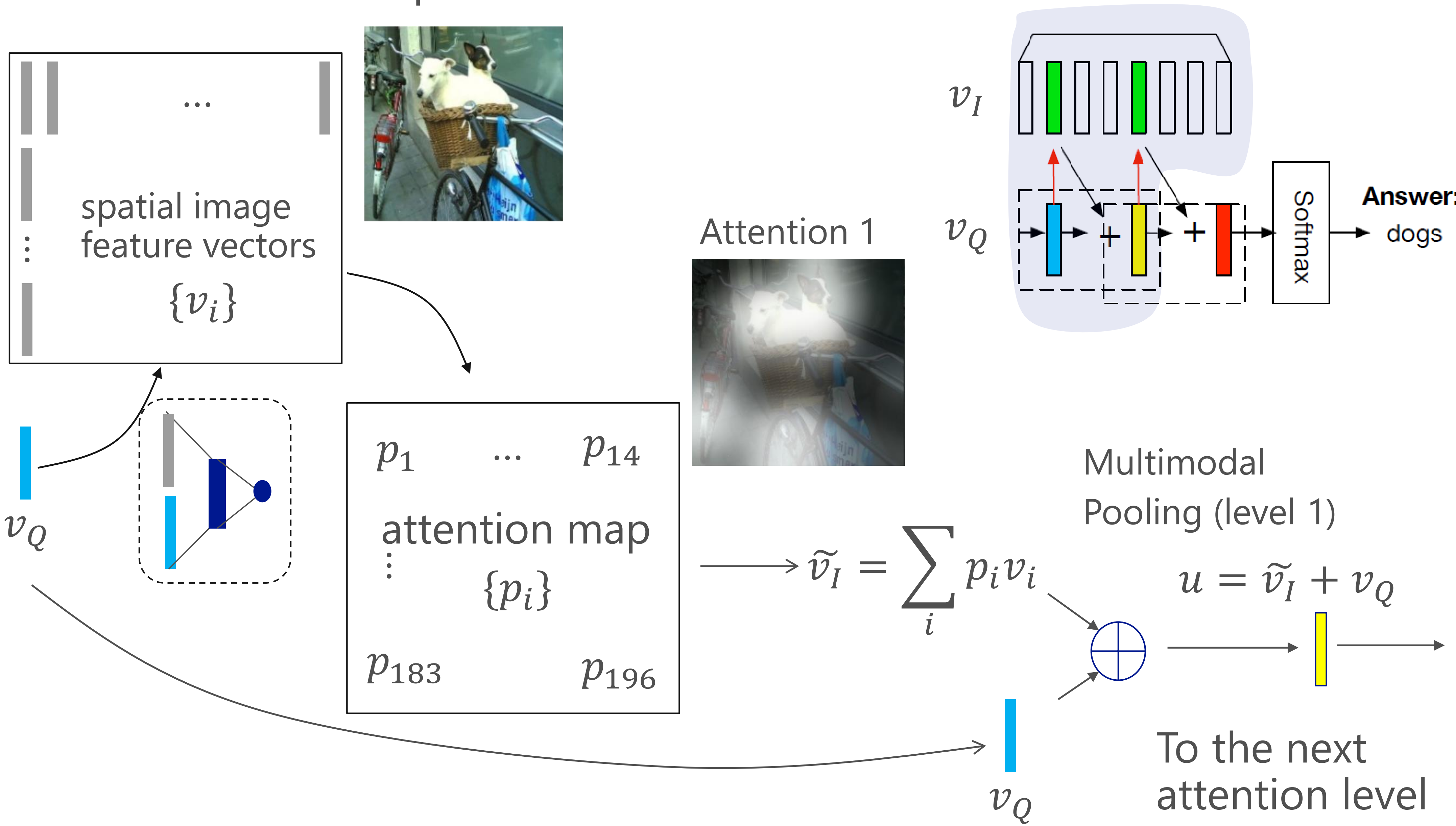


A LSTM cell

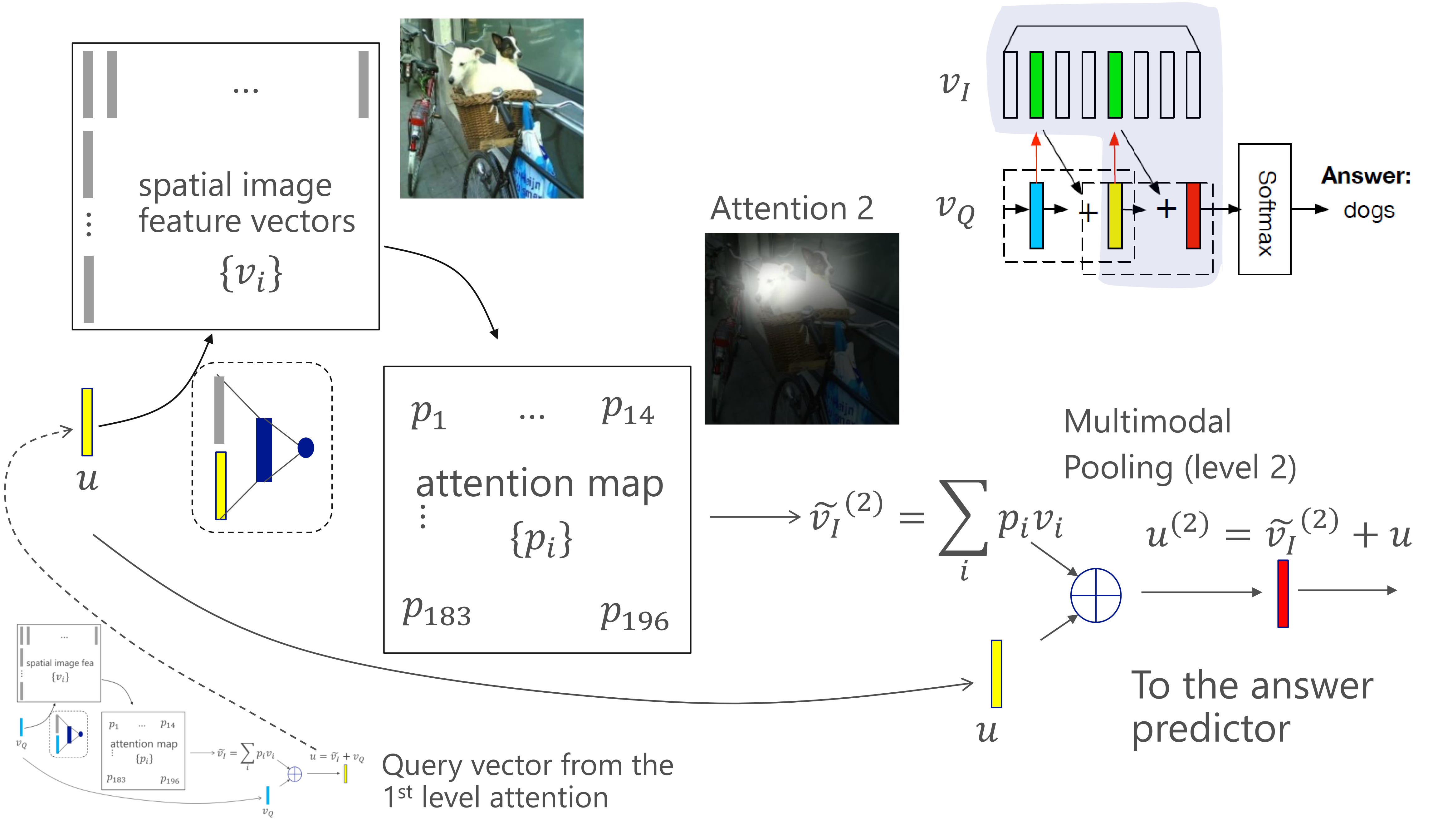




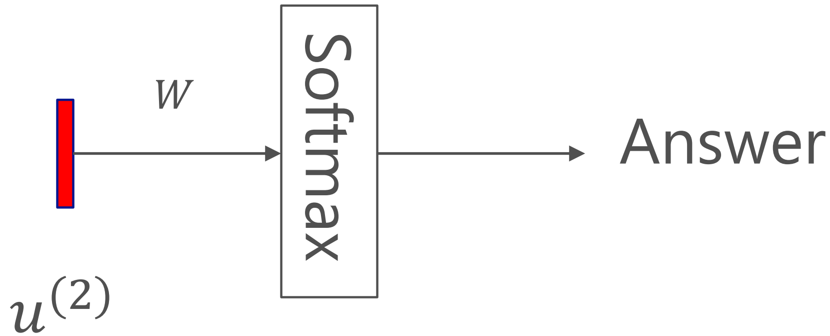
3. SAN: Compute the 1st level attention

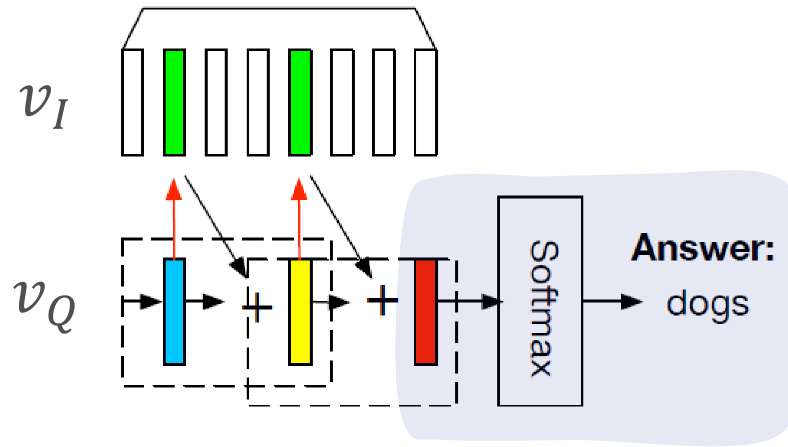


3. SAN: Compute the 2nd level attention



4. Answer prediction





$$p_{ans} = softmax(W\, u^{(2)} + b)$$

$$ans^* = \underset{\{ans\}}{\operatorname{argmax}}\{p_{ans}\}$$

Results

	test-dev				test-std
Methods	All	Yes/No	Number	Other	All
VQA: [1]					
Question	48.1	75.7	36.7	27.1	-
Image	28.1	64.0	0.4	3.8	-
Q+I	52.6	75.6	33.7	37.4	-
LSTM Q	48.8	78.2	35.7	26.6	-
LSTM Q+I	53.7	78.9	35.2	36.4	54.1
SAN(2, CNN)	**58.7**	**79.3**	**36.6**	**46.1**	**58.9**

Table 5: VQA results on the official server, in percentage

Other:
Object
Color
Location
...

Big improvement on the VQA benchmark (and COCO-QA, DAQUAR).

Q: what stands between two blue lounge chairs on an empty beach?

1st attention layer   2nd attention layer

Answer: **umbrella**

Bottom-Up Attention

A new view to the attention mechanism in deep learning

In human visual system, there are two kinds of attentions:

Top-down attention:

proactively initiated by the current task (e.g., look for something)

Bottom-up attention:

spontaneously emerge from visual salient stimuli

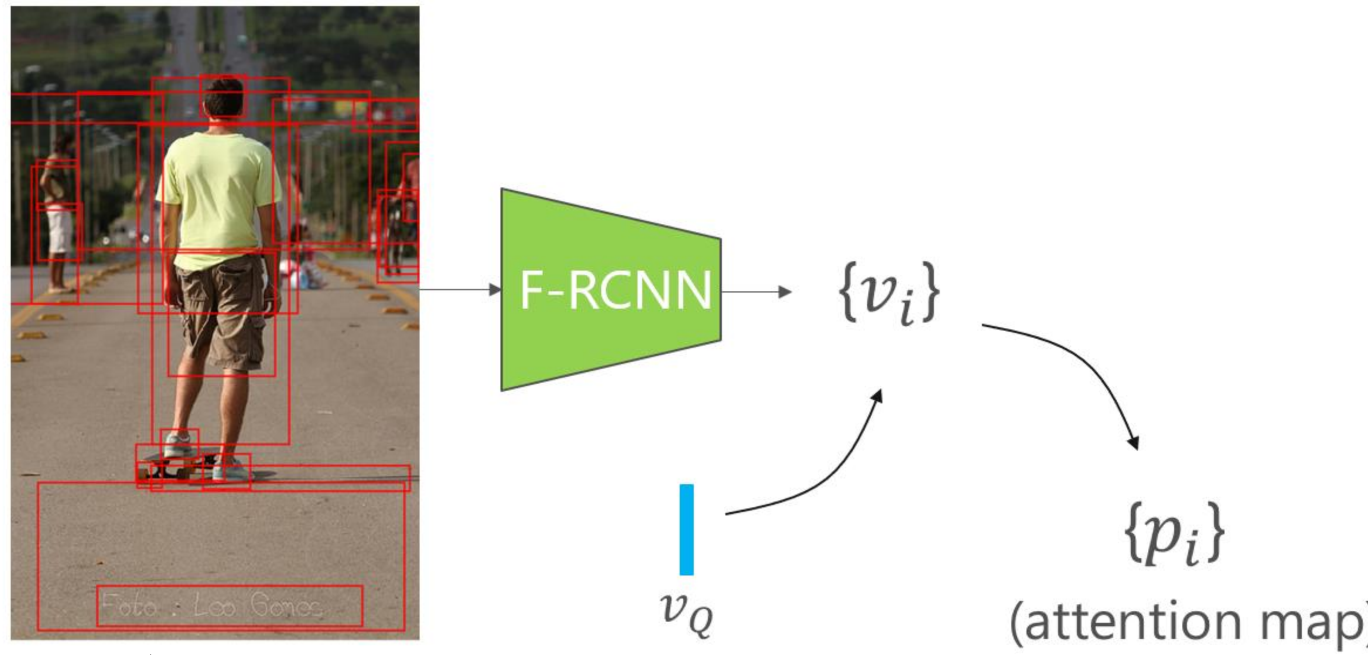
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson[1,*], Xiaodong He[2], Chris Buehler[2], Damien Teney[3]
Mark Johnson[4], Stephen Gould[1], Lei Zhang[2]
[1]Australian National University [2]Microsoft Research
[3]University of Adelaide [4]Macquarie University

Bottom-Up attention mechanism (new)

Bottom-Up attention:

- Use F-RCNN to detect key objects
- Compute spatial feature vector for each object
- Keep complete visual information for each object



Attend on actual objects, rather than on uniform grid regions like conventional top-down attention

Combine Bottom-Up & Top-Down Attention

Adopt similar terminology to humans' attention system:

- attention mechanisms driven by non visual or task-specific context as 'top-down'
- purely visual feed-forward attention mechanisms as 'bottom-up'.

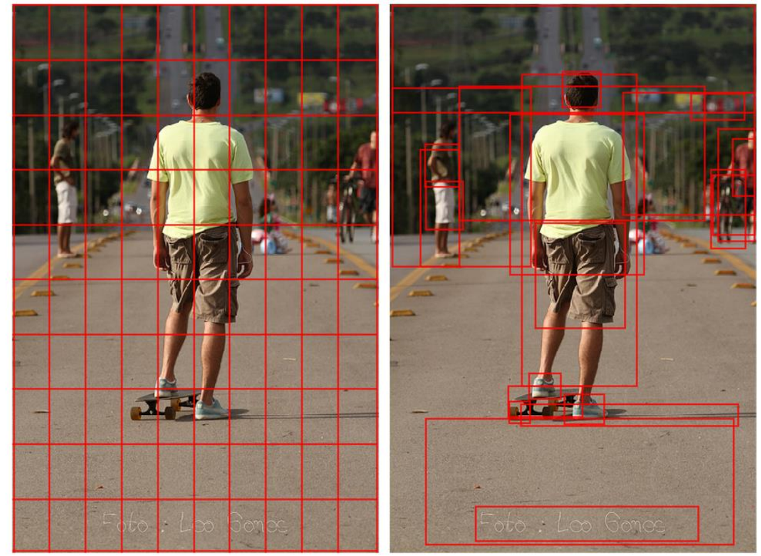
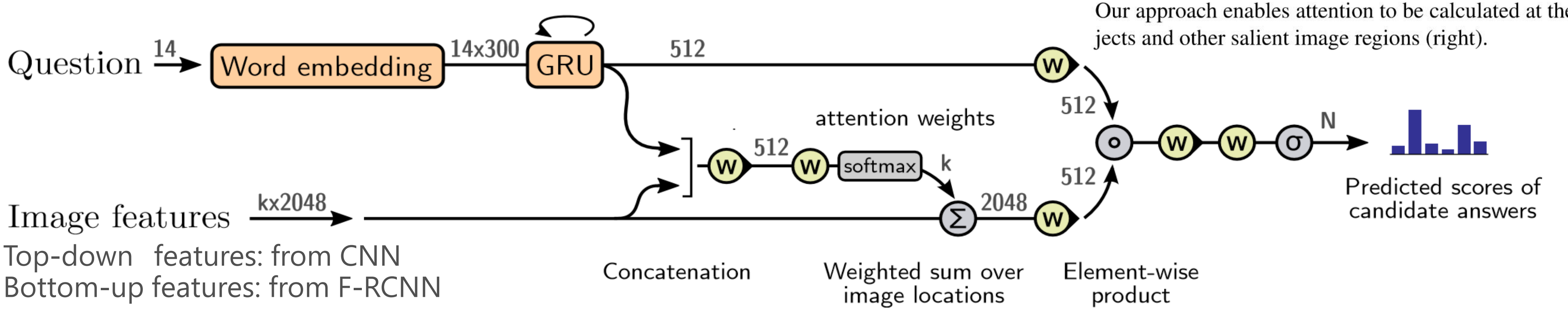


Figure 1. Typically, attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions (left). Our approach enables attention to be calculated at the level of objects and other salient image regions (right).

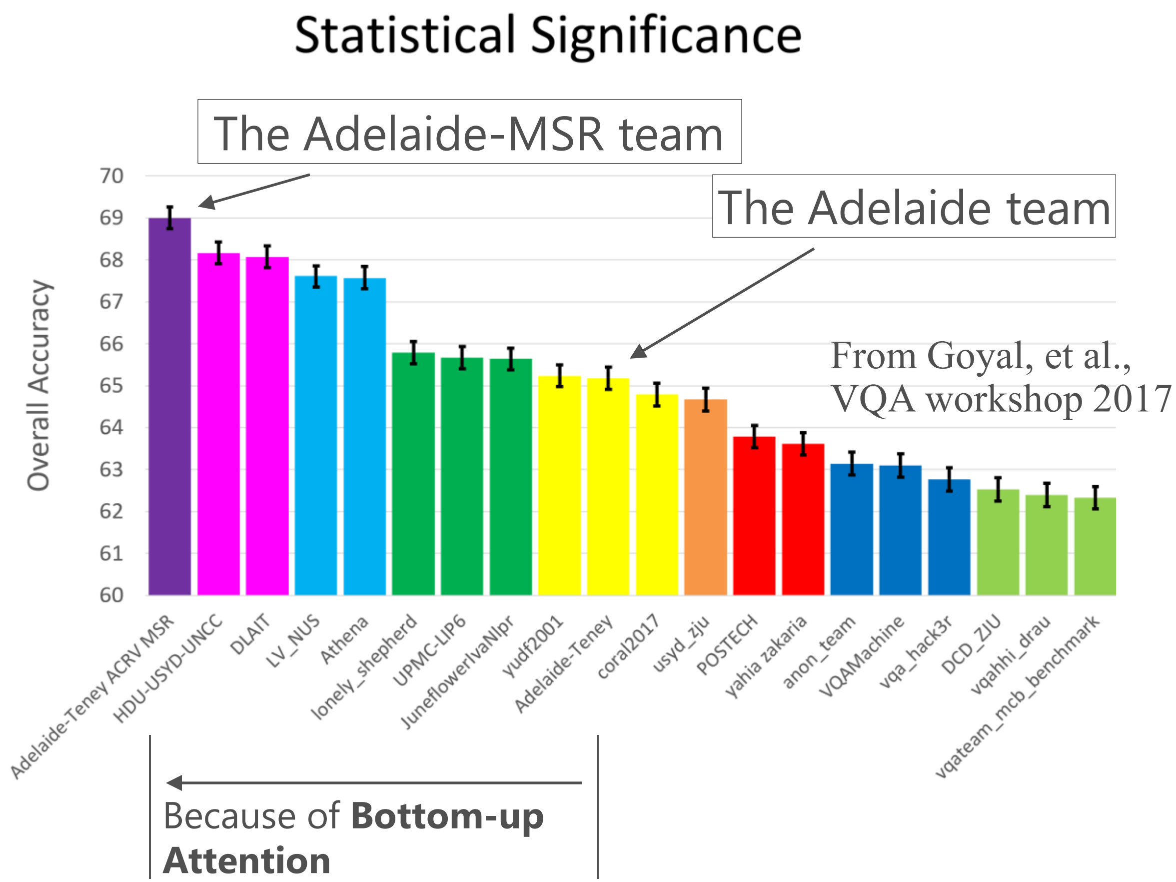
Overall Attention Net for VQA:

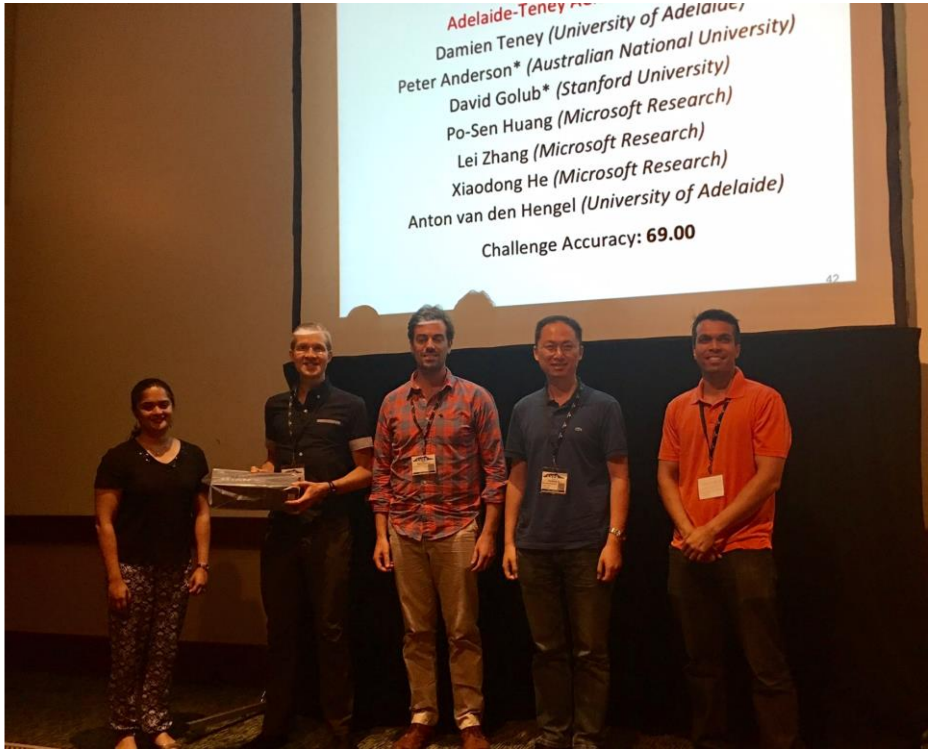


Top-down features: from CNN
Bottom-up features: from F-RCNN

VQA Challenge @ CVPR2017

MSR & Uni. of Adelaide won VQA2017 Challenge





[1] Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge, arXiv:1708.02711
[2] Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, arXiv:1707.07998

Language to Image Generation

- Express the abstract ideas described in natural language by drawing a picture (fill-in lots of details)

I dreamed a colorful bird with a sharp beak and black eye rings

Draw one for you

Image Synthesis with GAN

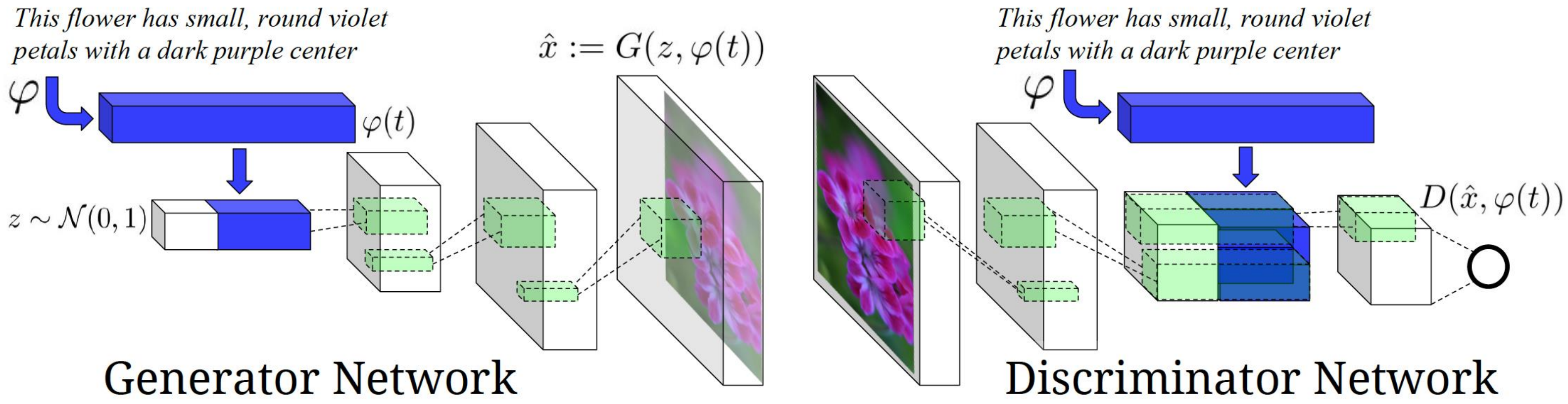


Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

Objective function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

[Reed et al., Generative adversarial text-to-image synthesis, ICML, 2016]

Stacked Attention for GANs

Propose Attention GANs to improve Image Synthesis

- Goals:
  - Improve the quality of generated images
  - Improve the interpretability of GANs
  - Stabilize the training of GANs

- Solution: Attention Generative Adversarial Networks (AttnGANs)
  - Propose a deep attention multimodal similarity model to learn visually-discriminative word features in an semi-supervised manner.
  - Propose the generative networks with stacked attention to generate images from low-to-high resolutions at different stages.

AttnGAN (Xu et al., 2017 @MSR)

AttnGAN (Xu et al., 2017 @MSR)

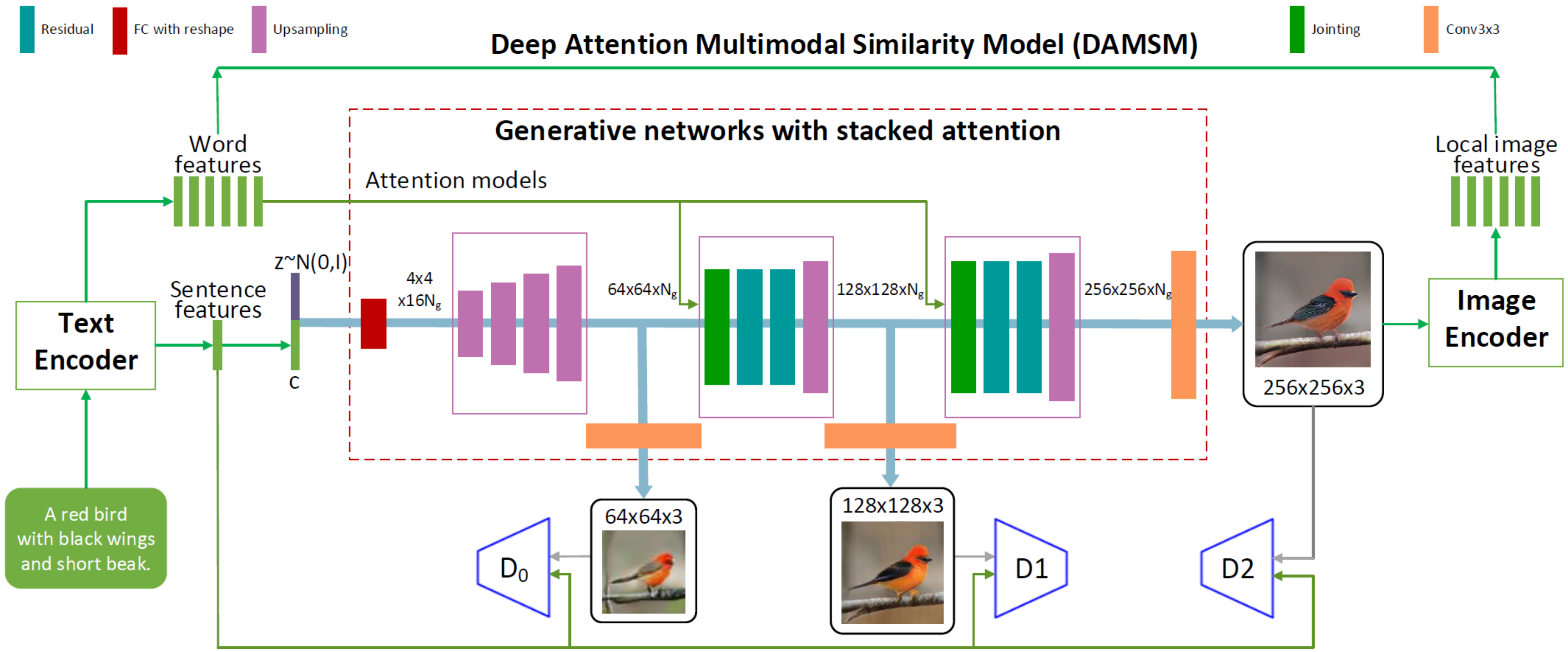


Figure 1: The architecture of the proposed AttnGANs.

- Generative networks with stacked attention → mult-scale images
- Discriminators $D_1, D_2, \ldots, D_m$ at multiple scales → the GAN loss
- Deep Attention Multimodal Similarity Model → the perception loss
  - Text encoder → sentence and word features
  - Image encoder → global and local image features

AttnGAN

$$\mathcal{L} = \underbrace{\mathcal{L}_G}_{\text{Regular GAN loss}} + \lambda \underbrace{\mathcal{L}_{DAMSM}}_{\text{Attention loss}}, \quad \text{where } \mathcal{L}_G = \sum_{i=1}^{m} \mathcal{L}_{G_i}$$

- Results:

Dataset	GAN-INT-CLS 64×64	GAWWN 128×128	StackGAN 256×256	MDAGAN 256×256	our AttnGAN 256×256
CUB	$2.88 \pm .04$	$3.62 \pm .07$	$3.70 \pm .04$	$3.82 \pm .06$	**$4.28 \pm .03$**
COCO	$7.88 \pm .07$	/	$8.45 \pm .03$	/	**$13.56 \pm .05$**

Table 1: Inception scores by state-of-the-art methods and the proposed AttnGAN on CUB and COCO test set. Higher inception scores mean better image quality.

Examples

this bird is red with white and
this bird has a green crown
black primaries and a white belly
1:bird 0:this 2:has 11:belly 10:white
6:black 4:green 10:white 0:this 1:bird

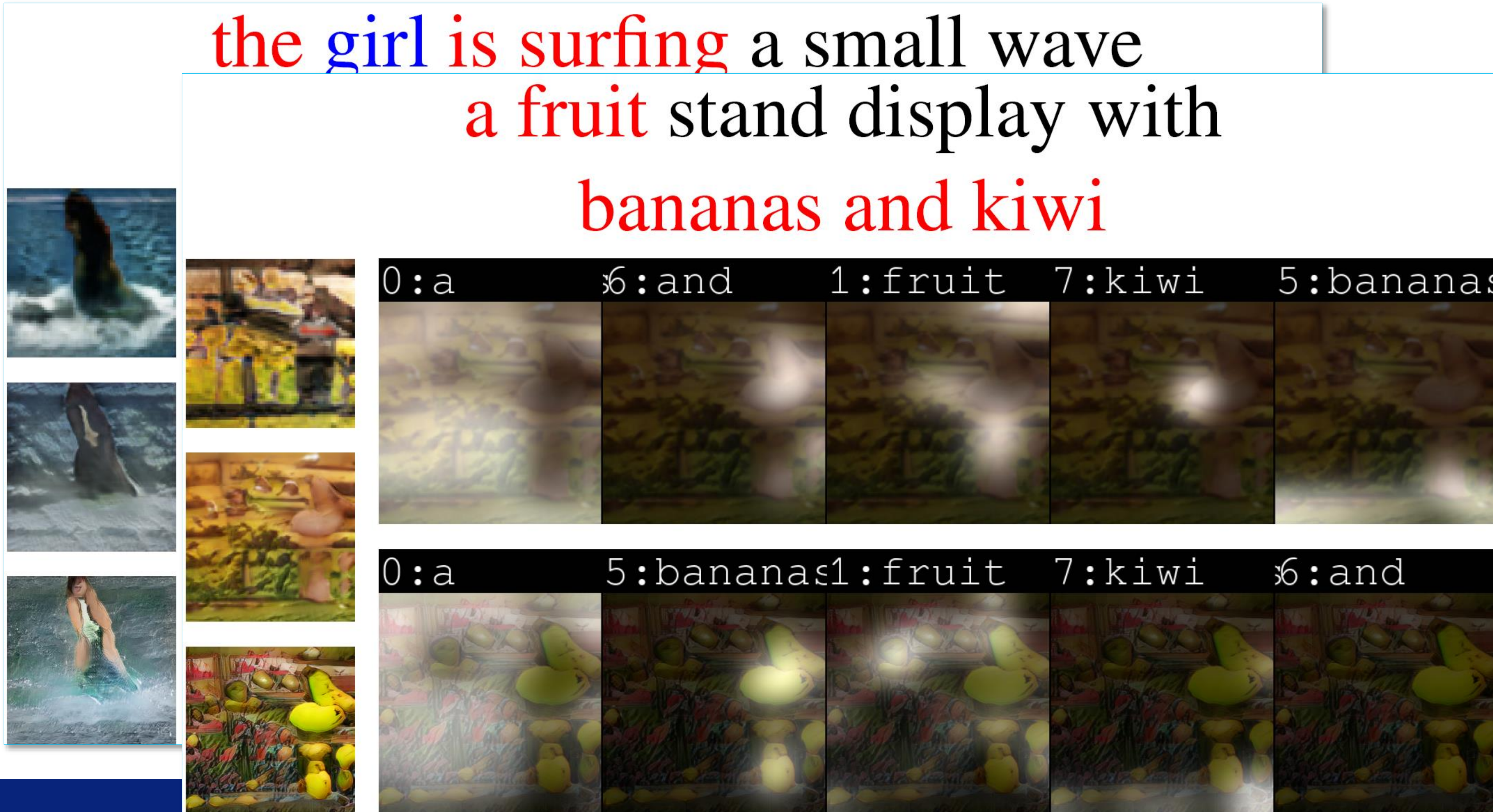
Control the image details by language

This bird has **wings that are black** and has a **white belly**.

This bird has **wings that are red** and has a **yellow belly**

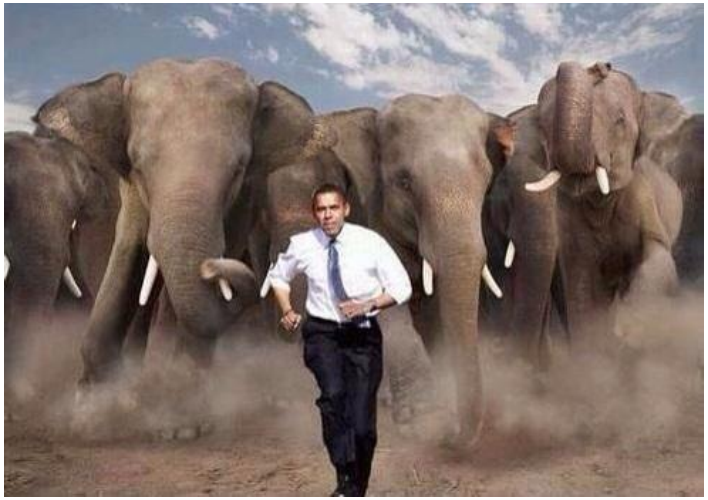
This bird has **wings that are blue** and has a **red belly**

Challenges


the girl is surfing a small wave
a fruit stand display with bananas and kiwi
0:a
6:and
1:fruit
7:kiwi
5:bananas
0:a
5:bananas
1:fruit
7:kiwi
6:and

And how creative or crazy AI can go ☺

Look Forward



In 2015:

➡ a herd of elephants standing next to a man

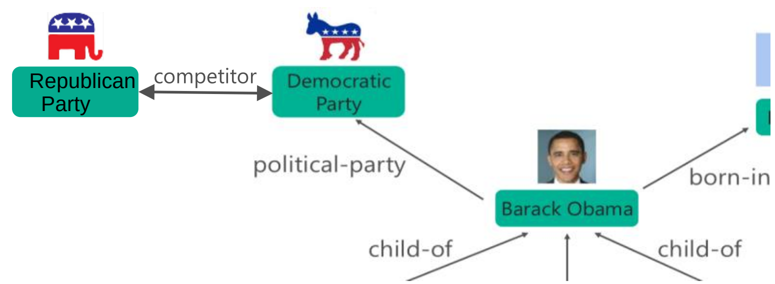
In 2016, + Entity:

➡ a herd of elephants standing next to Obama

Next, + knowledge & reasoning:

Obama is *the president from* the **Democratic party**,
whose *competitor is* the **Republican party**,
whose *mascot is* **Elephant**.

➡ Obama is chased by his Republican competitors ☺



Knowledge Graph

Who is that person?

Why these elephants are chasing him?

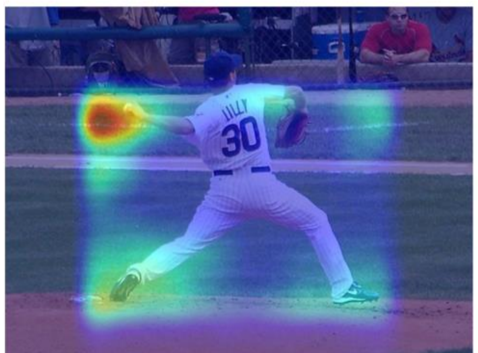
Draw me a picture showing Obama are fighting with Republican competitors ☺

Summary

Universal Chatbot, Digital Assistant, Mixed Reality, …

Multimodal Intelligence: perception, reasoning, and expression across language & vision

Image-to-language



ball (1.00)

a baseball player throwing a **ball**

Visual QA/Dialog

Q: what are sitting in the basket on a bicycle?
A: dogs.

Language-to-image

This bird is red with white and has a very short beak

Deep Learning / Deep Attention Mechanisms

Collaborators:

Hao Fang
Saurabh Gupta
Forrest Iandola
Rupesh Srivastava
Li Deng
Piotr Dollár
Jianfeng Gao
Xiaodong He
Margaret Mitchell
John Platt
Lawrence Zitnick
Geoffrey Zweig
Jacob Devlin

Kenneth Tran
Lei Zhang
Jian Sun
Chris Buehler
Chris Thrasher
Chris Sienkiewicz
Cornelia Carapcea
Yuxiao Hu
Yandong Guo
Zichao Yang
Alex Smola
Tao Xu
Chuang Gan

Zhe Gan

MSR Deep Learning Tech Center

People

Asli Celikyilmaz
Researcher

Jianshu Chen
Researcher

Roland Fernandez
Senior Researcher

Xiaodong He
Principal Researcher

Po-Sen Huang
Researcher

Qiuyuan Huang
Postdoc Researcher, Associate Researcher II

Ricky Loynd
Senior RSDE

James McCaffrey
Research Advanced Development

Hamid Palangi
Associate Researcher II

Paul Smolensky
Partner Researcher

Adith Swaminathan
Researcher

Kenneth Tran
Principal Research Engineer

Pengchuan Zhang
Associate Researcher II

Q & A

Thanks!