KNOW2LOOK: COMMONSENSE KNOWLEDGE FOR VISUAL SEARCH

Sreyasi Nag Chowdhury, Niket Tandon, Gerhard Weikum

Max Planck Institute for Informatics, Saarbrücken, Germany

IMAGE RETRIEVAL

User Query

Concrete

Q: *bicycle in street*

Abstract

Q: *environment friendly traffic*

IMAGE RETRIEVAL

User Query

Concrete

Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Abstract

Q: *environment friendly traffic*

IMAGE RETRIEVAL

User Query

Concrete

Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Abstract

Q: *environment friendly traffic*

Text-only ✗

IMAGE RETRIEVAL

User Query

Concrete

Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Visual objects: bicycle, bus, car

Text-only ✗

Text + visual ✓

Abstract

Q: *environment friendly traffic*

IMAGE RETRIEVAL

User Query

Concrete

Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Visual objects: bicycle, bus, car

"Biking by the river"

Visual objects: train, piano

Abstract

Q: *environment friendly traffic*

Text-only ✗

Text + visual ✓

IMAGE RETRIEVAL

User Query

Concrete

Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Visual objects: bicycle, bus, car

Text-only ✗

Text + visual ✓

"Biking by the river"

Visual objects: train, piano

Text-only ✗

Text + visual ✗

Abstract

Q: *environment friendly traffic*

IMAGE RETRIEVAL

User Query

Concrete

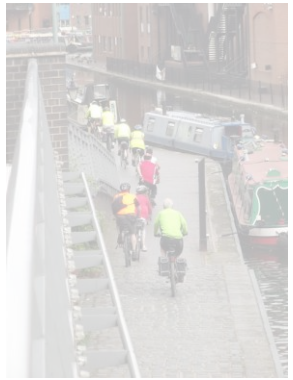
Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Visual objects: bicycle, bus, car

Text-only ✗

Text + visual ✓



"Biking by the river"

Visual objects: train, piano

Text-only ✗

Text + visual ✗

Abstract

Q: *environment friendly traffic*

"Riding for a cause."

Visual objects: person, bicycle

IMAGE RETRIEVAL

User Query

Concrete

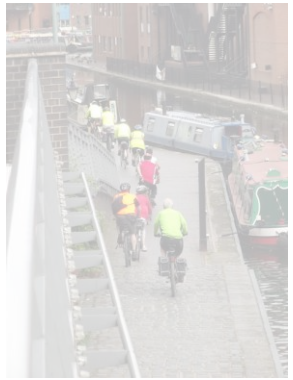
Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Visual objects: bicycle, bus, car

Text-only ✗

Text + visual ✓



"Biking by the river"

Visual objects: train, piano

Text-only ✗

Text + visual ✗

Abstract

Q: *environment friendly traffic*

"Riding for a cause."

Visual objects: person, bicycle

Text/visual ✗

IMAGE RETRIEVAL

User Query

Concrete

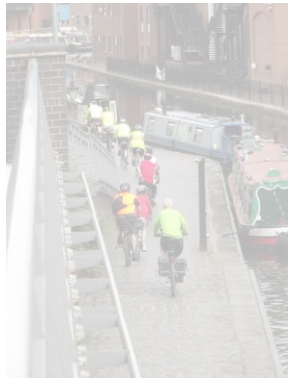
Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Visual objects: bicycle, bus, car

Text-only ✗

Text + visual ✓



"Biking by the river"

Visual objects: train, piano

Text-only ✗

Text + visual ✗

Abstract

Q: *environment friendly traffic*

"Riding for a cause."

Visual objects: person, bicycle

CSK: (riding bicycle, be, environment friendly)

Text/visual ✗

Text + visual + CSK ✓

IMAGE RETRIEVAL

User Query

Concrete

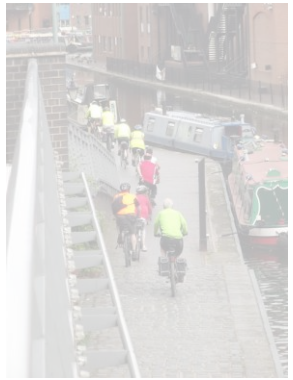
Q: *bicycle in street*

"Wow! Double-decker buses still run!"

Visual objects: bicycle, bus, car

Text-only ✗

Text + visual ✓



"Biking by the river"

Visual objects: train, piano

Text-only ✗

Text + visual ✗

Abstract

Q: *environment friendly traffic*

"Riding for a cause."

Visual objects: person, bicycle

CSK: (riding bicycle, be, environment friendly)

Text/visual ✗

Text + visual + CSK ✓

Our contribution

OUTLINE

- CSK: Where do we get it from?
- CSK: How do we use it?
- CSK: How to combine noisy signals?
- CSK: Does it help?

CSK: WHERE DO WE GET IT FROM? FROM OPEN-IE…

- Existing CSK knowledge bases: WordNet, ConceptNet, WebChild, Knowlywood

CSK: WHERE DO WE GET IT FROM? FROM OPEN-IE…

- Existing CSK knowledge bases: WordNet, ConceptNet, WebChild, Knowlywood
- Our corpus: Wiki articles from domain 'tourism'

CSK: WHERE DO WE GET IT FROM? FROM OPEN-IE…

- Existing CSK knowledge bases: WordNet, ConceptNet, WebChild, Knowlywood
- Our corpus: Wiki articles from domain 'tourism'
- Pruned by Jaccard Similarity

$$JaccardDistance = 1 - WeightedJaccardSimilarity$$

where,

$$WeightedJaccardSimilarity = \frac{\Sigma_n min[f(d_i, w_n), f(D, w_n)]}{\Sigma_n max[f(d_i, w_n), f(D, w_n)]}$$

CSK: WHERE DO WE GET IT FROM? FROM OPEN-IE…

- Existing CSK knowledge bases: WordNet, ConceptNet, WebChild, Knowlywood
- Our corpus: Wiki articles from domain 'tourism'
- Pruned by Jaccard Similarity

$$JaccardDistance = 1 - WeightedJaccardSimilarity$$

where,

$$WeightedJaccardSimilarity = \frac{\Sigma_n min[f(d_i, w_n), f(D, w_n)]}{\Sigma_n max[f(d_i, w_n), f(D, w_n)]}$$

"tourism" "be travel for" "recreation, leisure, family, business purposes" ✓
"people" "fall in" "love"

"the bloody hell" "be" "you" ✗

Domain-specific ReVerb triples

- ~22,000 CSK triples

CSK: HOW DO WE USE IT? FOR QUERY EXPANSION…

- Query string: *travel with backpack*

- CSK to expand query
  - t1: (tourists, use, travel maps)
  - t2: (tourists, carry, backpack)
  - t3: (backpack, is a type of, bag)

CSK: HOW DO WE USE IT? FOR QUERY EXPANSION…

- Query string: *travel with backpack*

- CSK to expand query
  - t1: (tourists, use, travel maps)
  - t2: (tourists, carry, backpack)
  - t3: (backpack, is a type of, bag)

- Document *x* with features
  - Textual: "A tourist reading a map by the road"
  - Visual: person, bag, bottle, bus

Text-only systems ✗

Text + visual + CSK systems ✓

CSK: HOW DO WE USE IT? FOR QUERY EXPANSION…

- Query string: *travel with backpack*

- CSK to expand query
  - t1: (tourists, use, travel maps)
  - t2: (tourists, carry, backpack)
  - t3: (backpack, is a type of, bag)

- Document *x* with features
  - Textual: "A tourist reading a map by the road"
  - Visual: person, bag, bottle, bus

Text-only systems ✗

Text + visual + CSK systems ✓

✓ CSK bridge vocabulary gap between query and document

✓ CSK establish relations between concepts

✓ CSK diminish noise from modalities – ensemble effect

CSK: HOW DO WE USE IT? EXAMPLE…

A tour group is standing on the grass with ruins in the background.
Group of people standing in front of a stone structure.

Document x

CSK: HOW DO WE USE IT? EXAMPLE…

A tour group is standing on the grass with ruins in the background.
Group of people standing in front of a stone structure.

Textual features x_x

Document x

CSK: HOW DO WE USE IT? EXAMPLE…

A tour group is standing on the grass with ruins in the background.
Group of people standing in front of a stone structure.

Textual features x_x

Visual features x_v : backpack, person

Document x

CSK: HOW DO WE USE IT? EXAMPLE…

A tour group is standing on the grass with ruins in the background.
Group of people standing in front of a stone structure.

Textual features x_x

Visual features x_v : backpack, bag, container
person, casual agent, organism

Document x

CSK: HOW DO WE USE IT? EXAMPLE…

A tour group is standing on the grass with ruins in the background.
Group of people standing in front of a stone structure.

Textual features x_x

Document x

Visual features x_v : backpack, bag, container
person, casual agent, organism

CSK features

Query: "group excursion"
Query expansion: (an excursion, be trip by, a group of people)
(organized excursions, book through, a tour company)

CSK: HOW DO WE USE IT? EXAMPLE…

A tour group is standing on the grass with ruins in the background.
Group of people standing in front of a stone structure.

Visual features x_v : backpack, bag, container
person, casual agent, organism

Query: "group excursion"
Query expansion: (an excursion, be trip by, a group of people)
(organized excursions, book through, a tour company)

CSK: HOW TO COMBINE NOISY SIGNALS? RETRIEVAL RANKING MODEL…

- Mixture LM: $$P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x]$$
- Commonsense-aware LM: $$P_{CS}[q|x] = \prod_i \left[\frac{\sum_k P[q_i|y_k] P[y_k|x]}{|k|} \right]$$
- Smoothed LM: $$P_{smoothed}[q|x] = \alpha P_{basic}[q|x] + (1-\alpha) P[q|B], \text{ where}$$
$$P[q|B] = \prod_i P[q_i|B]$$
- Basic LM: $$P_{basic}[q|x] = \prod_i P[q_i|x], \text{ where}$$
$$P[q_i|x] = \alpha_x P[q_i|xx_j] P[xx_j|x] + \alpha_v P[q_i|xv_j] P[xv_j|x]$$

CSK: HOW TO COMBINE NOISY SIGNALS? RETRIEVAL RANKING MODEL…

- Mixture LM: $P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x]$
- Commonsense-aware LM: $P_{CS}[q|x] = \prod_i \left[\frac{\sum_k P[q_i|y_k] P[y_k|x]}{|k|} \right]$
- Smoothed LM: $P_{smoothed}[q|x] = \alpha P_{basic}[q|x] + (1-\alpha) P[q|B]$, where

  $P[q|B] = \prod_i P[q_i|B]$
- Basic LM: $P_{basic}[q|x] = \prod_i P[q_i|x]$, where

  $P[q_i|x] = \alpha_x P[q_i|xx_j] P[xx_j|x] + \alpha_v P[q_i|xv_j] P[xv_j|x]$

CSK: HOW TO COMBINE NOISY SIGNALS? RETRIEVAL RANKING MODEL…

- Mixture LM: $$P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x]$$
- Commonsense-aware LM: $$P_{CS}[q|x] = \prod_i \left[\frac{\sum_k P[q_i|y_k] P[y_k|x]}{|k|} \right]$$ CSK triple
- Smoothed LM: $$P_{smoothed}[q|x] = \alpha P_{basic}[q|x] + (1-\alpha) P[q|B],$$ where $$P[q|B] = \prod_i P[q_i|B]$$
- Basic LM: $$P_{basic}[q|x] = \prod_i P[q_i|x],$$ where $$P[q_i|x] = \alpha_x P[q_i|xx_j] P[xx_j|x] + \alpha_v P[q_i|xv_j] P[xv_j|x]$$

CSK: HOW TO COMBINE NOISY SIGNALS? RETRIEVAL RANKING MODEL…

- Mixture LM: $$P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x]$$
- Commonsense-aware LM: $$P_{CS}[q|x] = \prod_i \left[\frac{\sum_k P[q_i|y_k] P[y_k|x]}{|k|} \right]$$

Probabilities based on word-wise overlaps

- Smoothed LM: $$P_{smoothed}[q|x] = \alpha P_{basic}[q|x] + (1-\alpha) P[q|B],$$ where
$$P[q|B] = \prod_i P[q_i|B]$$
- Basic LM: $$P_{basic}[q|x] = \prod_i P[q_i|x],$$ where
$$P[q_i|x] = \alpha_x P[q_i|xx_j] P[xx_j|x] + \alpha_v P[q_i|xv_j] P[xv_j|x]$$

CSK: HOW TO COMBINE NOISY SIGNALS? RETRIEVAL RANKING MODEL…

- Mixture LM: $$P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x]$$
- Commonsense-aware LM: $$P_{CS}[q|x] = \prod_i \left[\frac{\sum_k P[q_i|y_k] P[y_k|x]}{|k|} \right]$$
- Smoothed LM: $$P_{smoothed}[q|x] = \alpha P_{basic}[q|x] + (1-\alpha) P[q|B], \text{ where}$$
$$P[q|B] = \prod_i P[q_i|B]$$

Background corpus – Co-occurring Flickr tags

- Basic LM: $$P_{basic}[q|x] = \prod_i P[q_i|x], \text{ where}$$
$$P[q_i|x] = \alpha_x P[q_i|xx_j] P[xx_j|x] + \alpha_v P[q_i|xv_j] P[xv_j|x]$$

CSK: HOW TO COMBINE NOISY SIGNALS? RETRIEVAL RANKING MODEL…

- Mixture LM: $$P[q|x] = \beta_{CS} P_{CS}[q|x] + (1 - \beta_{CS}) P_{smoothed}[q|x]$$
- Commonsense-aware LM: $$P_{CS}[q|x] = \prod_i \left[\frac{\sum_k P[q_i|y_k] P[y_k|x]}{|k|} \right]$$
- Smoothed LM: $$P_{smoothed}[q|x] = \alpha P_{basic}[q|x] + (1-\alpha) P[q|B], \text{ where}$$
$$P[q|B] = \prod_i P[q_i|B]$$
- Basic LM: $$P_{basic}[q|x] = \prod_i P[q_i|x], \text{ where}$$
$$P[q_i|x] = \alpha_x P[q_i|xx_j] P[xx_j|x] + \alpha_v P[q_i|xv_j] P[xv_j|x]$$

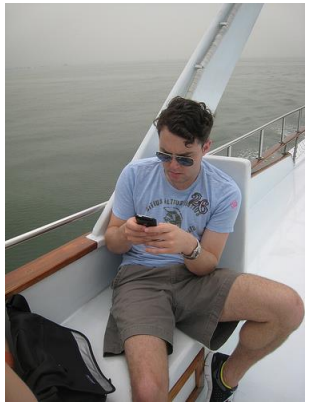
Textual and visual features

CSK: DOES IT HELP? EVALUATION…

- Image Dataset
  - Flickr30k
  - MS COCO captioned dataset
  - Pascal Sentence Dataset
  - SBU captioned dataset

CSK: DOES IT HELP? EVALUATION…

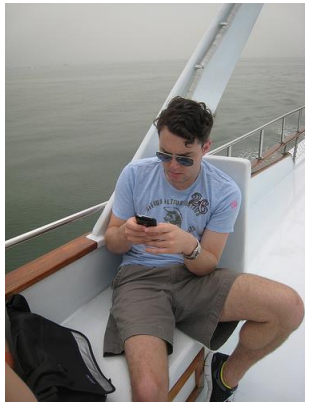
- Image Dataset
  - Flickr30k
  - MS COCO captioned dataset
  - Pascal Sentence Dataset
  - SBU captioned dataset



Boat trip to see the mythical pink dolphins…
this is John checking in with the office for that day.

CSK: DOES IT HELP? EVALUATION…

- Image Dataset
  - Flickr30k
  - MS COCO captioned dataset
  - Pascal Sentence Dataset
  - SBU captioned dataset

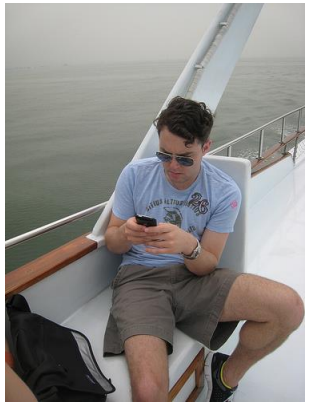


Boat trip to see the mythical pink dolphins…
this is John checking in with the office for that day.

A group of tourists is crossing a bridge that connects a walking path to a trail of nature.
Many people cross a very tall footbridge with a tree-covered hill in the background.
This shows a group of people walking over an arched red bridge.
People cross a large bridge to get over the body of water.
People walking over a white and red bridge over a pond.

CSK: DOES IT HELP? EVALUATION…

- Image Dataset
  - Flickr30k
  - MS COCO captioned dataset
  - Pascal Sentence Dataset
  - SBU captioned dataset



Boat trip to see the mythical pink dolphins…
this is John checking in with the office for that day.

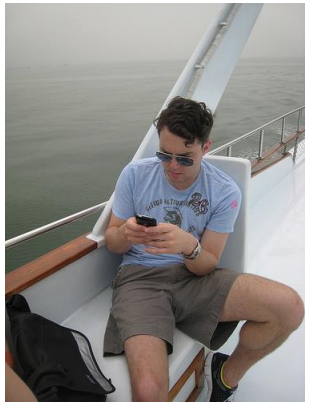
social media post

A group of tourists is crossing a bridge that connects a walking path to a trail of nature.
Many people cross a very tall footbridge with a tree-covered hill in the background.
This shows a group of people walking over an arched red bridge.
People cross a large bridge to get over the body of water.
People walking over a white and red bridge over a pond.

blog post

CSK: DOES IT HELP? EVALUATION…

- Image Dataset
  - Flickr30k
  - MS COCO captioned dataset
  - Pascal Sentence Dataset
  - SBU captioned dataset



Boat trip to see the mythical pink dolphins…
this is John checking in with the office for that day.

social media post

A group of tourists is crossing a bridge that connects a walking path to a trail of nature.
Many people cross a very tall footbridge with a tree-covered hill in the background.
This shows a group of people walking over an arched red bridge.
People cross a large bridge to get over the body of water.
People walking over a white and red bridge over a pond.

blog post

- ~ 50,000 images with captions

CSK: DOES IT HELP? EVALUATION…

- Baselines: Text-only and Text + Visual search approaches
- Evaluation metric: Average Precision @ 10

CSK: DOES IT HELP? EVALUATION…

- Baselines: Text-only and Text + Visual search approaches
- Evaluation metric: Average Precision @ 10

Text-only	Text + Visual	Text + Visual + CSK (Know2Look)
47%	64%	85%

CSK: DOES IT HELP? EVALUATION…

- Baselines: Text-only and Text + Visual search approaches
- Evaluation metric: Average Precision @ 10

Text-only	Text + Visual	Text + Visual + CSK (Know2Look)
47%	64%	85%

- Examples queries:
  - Concrete – *ball park, bridge road, table home, bicycle road*
  - Abstract – *diesel transport, housing town*
  - Mixed – *old clock, backpack travel, boat tour*

CSK: DOES IT HELP? EVALUATION…

- Baselines: Text-only and Text + Visual search approaches
- Evaluation metric: Average Precision @ 10

Text-only	Text + Visual	Text + Visual + CSK (Know2Look)
47%	64%	85%

- Examples queries:
  - Concrete – *ball park, bridge road, table home, bicycle road*
  - Abstract – *diesel transport, housing town*
  - Mixed – *old clock, backpack travel, boat tour*

} Co-occurring Flickr tags

CSK: DOES IT HELP? EVALUATION…EXAMPLE

Query: "group excursion"

Text-only

xx_j: "A small excursion boat anchored on the beach at the resort in Mexico. "

Text + Visual

xv_j: lunar excursion module, conveyance

Text + Visual + CSK (Know2Look)

xx_j: "A group of people riding camels."
y_k: (an excursion, be trip by,
a group of people)

CSK: DOES IT HELP? EVALUATION…EXAMPLE

Query: "group excursion"

Text-only

xx_j: "A small excursion boat anchored on the beach at the resort in Mexico. "

Text + Visual

xv_j: lunar excursion module, conveyance

Text + Visual + CSK (Know2Look)

xx_j: "A group of people riding camels."
y_k: (an excursion, be trip by, a group of people)

OBSERVATIONS

- Noisy OpenIE triples capture commonsense knowledge
- Noisy textual cues + noisy visual object detection + noisy commonsense knowledge → **ensemble effect** → better results for multimodal document retrieval
- CSK act as bridge between text and vision

OBSERVATIONS

- Noisy OpenIE triples capture commonsense knowledge
- Noisy textual cues + noisy visual object detection + noisy commonsense knowledge → **ensemble effect** → better results for multimodal document retrieval
- CSK act as bridge between text and vision

- Do word co-occurrences or word embeddings provide similar results?
- Does structured commonsense knowledge improve retrieval?

OBSERVATIONS

- Noisy OpenIE triples capture commonsense knowledge
- Noisy textual cues + noisy visual object detection + noisy commonsense knowledge → **ensemble effect** → better results for multimodal document retrieval
- CSK act as bridge between text and vision

- Do word co-occurrences or word embeddings provide similar results?
- Does structured commonsense knowledge improve retrieval?

Thank you