

Cross-Modal Retrieval: A Pairwise Classification Approach

Aditya Krishna Menon^{1,2}, Didi Surian^{1,3} and Sanjay Chawla^{3,4}

¹National ICT Australia, Australia

²Australian National University, Australia

³The University of Sydney, Australia

⁴Qatar Computing Research Institute, Qatar

SIAM International Conference on Data Mining 2015
April 30th, 2015

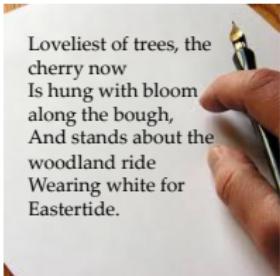


Overview of talk

- 1 Overview
- 2 Cross-modal retrieval: unsupervised and supervised
- 3 The similarity function perspective
- 4 Learning without annotations
- 5 Learning with annotations
- 6 Experiments
- 7 Conclusion

Background

Content is increasingly available in multiple **modalities** (images, text, videos, ...)



Inner Revenue: 'I have it and I have no
lower resources, because I am heavily taxed.
People, however, who
like to have fun, especially great Headache.
Many, but not all, big people & figures
are bad for us.

Who loves people & likes to talk, which seems nice
but is very tiring, both to give, but also a drag
and burden for him.
has been caught in a hot, uncomfortable sunny
late summer day or sun or very, reading
behavior may, may,

—John Berryman



3. THE BURIAL OF THE DEAD

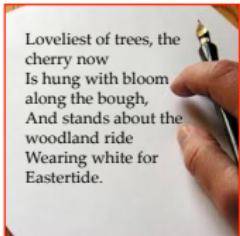
A PRIL is the cruel month, leading
A little dead man to his last load, making
Horses and drivers, etc., etc.,
Dull more with spring rains
Winter long as weevs, covering
Earth in hopeless snows, feeding
A little life with cold tales.
Summer surprised us, coming over the
Starchepaean
With clouds of roses: we stayed in the
valleians.
And went on in midnight, into the Hollow
gates, —

Cross-modal retrieval: informally

Given the representation of an entity in one modality, find its best representation in all other modalities

Cross-modal retrieval: informally

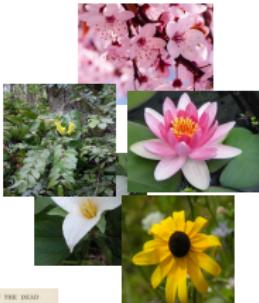
e.g. Given a piece of text, what is the image that best accompanies it?



Dear Friends, "I, too, hate and I, love, or
hate, respectively, because I am Many-hued.
People have said, "I am especially great elsewhere;
Many have me with his gloom, & others
as bad as insects."

Who loves people in a world of cold, hard hearts,
And who thought kindly in a world like this?—
And loves me? A boy
has been born to me, he is exceedingly noisy
We understand, or we do not, exactly
Lambeth says, who?

—William Blake

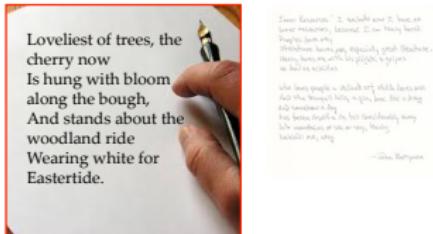


I. THE RITUAL OF THE DEAD

A
TILL is the crooked work, bending
Liberates out of the dead land, setting
Honey and down, stirring
Gold mists with spring rains.
Mists, mists, mists, mists, mists,
Earth in fragrant snows, bending
A little life with closed valances,
Bent over us, coming over the
longing ground.
With a shower of rain, we stepped in the
columns,
And were on in sunlight, into the Hall
gates. — 12

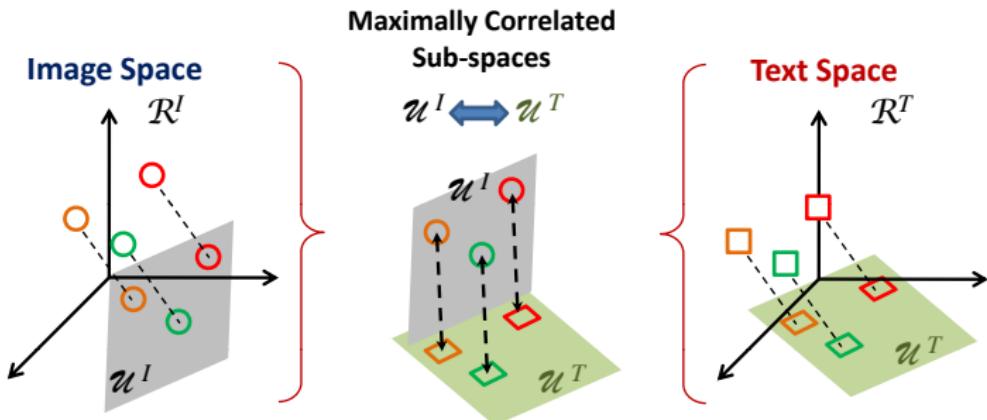
Cross-modal retrieval: informally

e.g. Given a piece of text, what is the image that best accompanies it?



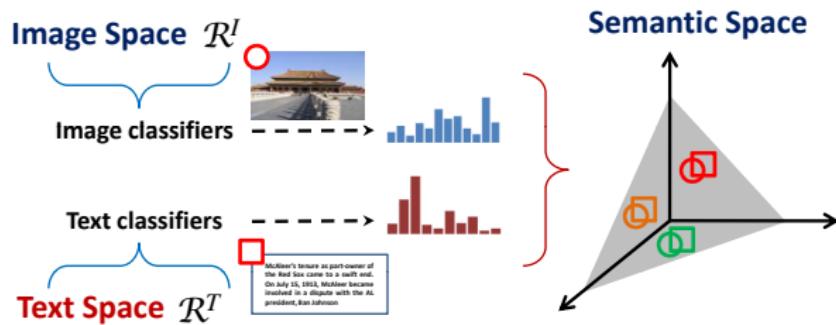
Prior work: CCA

Canonical Correlation Analysis (CCA): project both modalities onto a **latent** subspace where they exhibit maximum correlation



Prior work: SM

Semantic Matching (SM): project both modalities into a rich **supervised** subspace



This paper

A **pairwise classification** perspective on cross-modal retrieval

- Goal is to **score pairs of instances** based on affinity

Seamlessly study scenarios with and without **ground-truth annotation**

- Essentially, reduction to logistic regression on pairs of instances
- **Unification** of the CCA and SM approaches

Good empirical performance on real-world retrieval tasks

Overview of talk

- 1 Overview
- 2 Cross-modal retrieval: unsupervised and supervised
- 3 The similarity function perspective
- 4 Learning without annotations
- 5 Learning with annotations
- 6 Experiments
- 7 Conclusion

(Unsupervised) Cross-modal retrieval: training set

Training examples: $\mathcal{D} = \{(a^{(i)}, b^{(i)})\}_{i=1}^n$



(Unsupervised) Cross-modal retrieval: training set

Training examples: $\mathcal{D} = \{(a^{(i)}, b^{(i)})\}_{i=1}^n$

- Set of **suitable pairings** of entities across two modalities
- $(a^{(i)}, b^{(i)}) \in \mathcal{A} \times \mathcal{B}$, $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ and $\mathcal{B} \subseteq \mathbb{R}^{d_b}$ the modality representations



Cross-modal retrieval: goal

Goal: Learn **retrieval mappings** $f : \mathcal{A} \rightarrow \mathcal{B}$ and $g : \mathcal{B} \rightarrow \mathcal{A}$

Cross-modal retrieval: goal

Goal: Learn retrieval mappings $f : \mathcal{A} \rightarrow \mathcal{B}$ and $g : \mathcal{B} \rightarrow \mathcal{A}$

- e.g. Given a piece of text, what is the best accompanying image?

$f :$

*They call me 'The Wild Rose'
But my name was Elisa Day
Why they call me it, I do not know
For my name was Elisa Day*



Assumption: ground-truth annotations

We assume there is a set of **ground-truth annotations** for each entity

- For our purposes, a label from some set $\mathcal{Y} = \{1, 2, \dots, K\}$



Arts



Nature



Nature



History



Politics



Science

These annotations may be **unobserved**

- But are still used for **evaluation**

(Supervised) Cross-modal retrieval: training set

Training examples: $\mathcal{D} = \{((a^{(i)}, b^{(i)}),)\}_{i=1}^n$

(Supervised) Cross-modal retrieval: training set

Training examples: $\mathcal{D} = \{((a^{(i)}, b^{(i)}), y^{(i)})\}_{i=1}^n$

- $y^{(i)} \in \mathcal{Y}$ are the **annotations**



Goal is as before

Overview of talk

- 1 Overview
- 2 Cross-modal retrieval: unsupervised and supervised
- 3 The similarity function perspective
- 4 Learning without annotations
- 5 Learning with annotations
- 6 Experiments
- 7 Conclusion

Latent map approach

Popular approach: learn **latent maps** $\psi_{\mathcal{A}} : \mathcal{A} \rightarrow \mathbb{R}^k$, $\psi_{\mathcal{B}} : \mathcal{B} \rightarrow \mathbb{R}^k$

- $k \in \mathbb{N}_+$ is the dimensionality of some **latent subspace**

Compute retrieval maps

$$f : a \mapsto \operatorname{argmin}_{b \in \mathcal{B}} d(\psi_{\mathcal{A}}(a), \psi_{\mathcal{B}}(b))$$

$$g : b \mapsto \operatorname{argmin}_{a \in \mathcal{A}} d(\psi_{\mathcal{A}}(a), \psi_{\mathcal{B}}(b)),$$

where $d(\cdot, \cdot)$ is some suitable distance function

Latent map approach

Popular approach: learn **latent maps** $\psi_{\mathcal{A}} : \mathcal{A} \rightarrow \mathbb{R}^k$, $\psi_{\mathcal{B}} : \mathcal{B} \rightarrow \mathbb{R}^k$

- $k \in \mathbb{N}_+$ is the dimensionality of some **latent subspace**

Compute retrieval maps

$$f : a \mapsto \operatorname{argmin}_{b \in \mathcal{B}} d(\psi_{\mathcal{A}}(a), \psi_{\mathcal{B}}(b))$$

$$g : b \mapsto \operatorname{argmin}_{a \in \mathcal{A}} d(\psi_{\mathcal{A}}(a), \psi_{\mathcal{B}}(b)),$$

where $d(\cdot, \cdot)$ is some suitable distance function

e.g.: In CCA, we learn **linear projections**

$$\psi_{\mathcal{A}} : a \mapsto Ua$$

$$\psi_{\mathcal{B}} : b \mapsto Vb,$$

and use **Euclidean distance** for retrieval

Similarity function approach

Our approach: learn a **similarity function**, $s : \mathcal{A} \times \mathcal{B} \rightarrow \mathbb{R}$

- Assigns a high score when the given pair are related to each other

Compute retrieval maps

$$f : a \mapsto \operatorname{argmax}_{b \in \mathcal{B}} s(a, b)$$

$$g : b \mapsto \operatorname{argmax}_{a \in \mathcal{A}} s(a, b).$$

Captures the latent map approach as a special case

Similarity function approach

Our approach: learn a **similarity function**, $s : \mathcal{A} \times \mathcal{B} \rightarrow \mathbb{R}$

- Assigns a high score when the given pair are related to each other

Compute retrieval maps

$$f : a \mapsto \operatorname{argmax}_{b \in \mathcal{B}} s(a, b)$$

$$g : b \mapsto \operatorname{argmax}_{a \in \mathcal{A}} s(a, b).$$

Captures the latent map approach as a special case

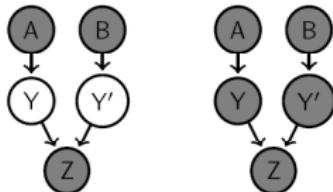
- But what is a good similarity?

What is a good similarity?

$s(a, b)$ should be high iff a, b have the same ground-truth annotation

What is a good similarity?

$s(a, b)$ should be high iff a, b have the same ground-truth annotation



A: random variable over \mathcal{A} (1^{st} modality)

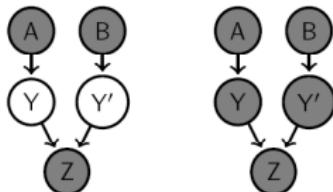
B: random variable over \mathcal{B} (2^{nd} modality)

Y: random variable over \mathcal{Y} (annotation for A)

Y': random variable over \mathcal{Y}' (annotation for B)

What is a good similarity?

$s(a, b)$ should be high iff a, b have the same ground-truth annotation



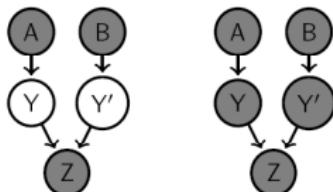
A: random variable over \mathcal{A} (1^{st} modality)
B: random variable over \mathcal{B} (2^{nd} modality)
Y: random variable over \mathcal{Y} (annotation for A)
Y': random variable over \mathcal{Y}' (annotation for B)

Here, $Z := 2[\mathbb{Y} = \mathbb{Y}'] - 1 \in \{\pm 1\}$

- whether or not the two annotations agree

What is a good similarity?

$s(a, b)$ should be high iff a, b have the same ground-truth annotation



A: random variable over \mathcal{A} (1^{st} modality)
B: random variable over \mathcal{B} (2^{nd} modality)
Y: random variable over \mathcal{Y} (annotation for A)
Y': random variable over \mathcal{Y}' (annotation for B)

Here, $Z := 2[\![Y = Y']\!] - 1 \in \{\pm 1\}$

- whether or not the two annotations agree

Judge s according to probability of incorrectly assessing compatible pairs:

$$\mathbb{L}(s) = \mathbb{P}_{A,B,Z}(s(A, B) \cdot Z < 0).$$

- sign of s should indicate whether the ground truths align

Reduction to pairwise classification

We can view Z as a **binary label** on the instance pair (A, B)

Reduction to pairwise classification

We can view Z as a **binary label** on the instance pair (A, B)

Learning a similarity function \equiv binary classification on the **paired instance space** $\mathcal{A} \times \mathcal{B}$

Reduction to pairwise classification

We can view Z as a **binary label** on the instance pair (A, B)

Learning a similarity function \equiv binary classification on the **paired instance space $\mathcal{A} \times \mathcal{B}$**

Ideally, treat $\mathcal{D} = \{((a^{(i)}, b^{(i)}), z^{(i)})\}_{i=1}^n$ as input to a standard binary classifier

Reduction to pairwise classification

We can view Z as a **binary label** on the instance pair (A, B)

Learning a similarity function \equiv binary classification on the **paired instance space $\mathcal{A} \times \mathcal{B}$**

Ideally, treat $\mathcal{D} = \{((a^{(i)}, b^{(i)}), z^{(i)})\}_{i=1}^n$ as input to a standard binary classifier

More effort is needed depending on whether or not we have ground-truth annotations

Overview of talk

- 1 Overview
- 2 Cross-modal retrieval: unsupervised and supervised
- 3 The similarity function perspective
- 4 Learning without annotations**
- 5 Learning with annotations
- 6 Experiments
- 7 Conclusion

Learning without annotations

Without ground-truth annotations, we **only observe pairs that are linked**

- We do not observe pairs that **do not** link

Training set: $\mathcal{D}' = \{((a^{(i)}, b^{(i)}), z^{(i)})\}$ where all labels $z^{(i)} = +1$



Learning without annotations

Without ground-truth annotations, we **only observe pairs that are linked**

- We do not observe pairs that **do not** link

Training set: $\mathcal{D}' = \{((a^{(i)}, b^{(i)}), z^{(i)})\}$ where all labels $z^{(i)} = +1$

- Learning from **positive only** data
- Cannot apply standard binary classifier!



Reduction to positive and unlabelled learning

Basic idea: Construct **all pairs** of instances from each modality

$$\mathcal{D}'' = \{((a^{(i)}, b^{(j)}), z^{(i,j)})\}_{i,j=1}^n$$

We know $z^{(i,i)} = 1$, but $z^{(i,j)}$ **unknown** for $i \neq j$



Reduction to positive and unlabelled learning

Basic idea: Construct **all pairs** of instances from each modality

$$\mathcal{D}'' = \{((a^{(i)}, b^{(j)}), z^{(i,j)})\}_{i,j=1}^n$$

We know $z^{(i,i)} = 1$, but $z^{(i,j)}$ **unknown** for $i \neq j$

- Learning from **positive and unlabelled** data
- Aims to **better discriminate** linked and unlinked pairs



Reduction to binary classification

Simple recipe for positive and unlabelled learning [Elkan and Noto, 2008]:

- Treat unlabelled instances as negatives
- Learn a **class-probability estimator** (e.g. logistic regression)
- Optimal for ranking purposes!

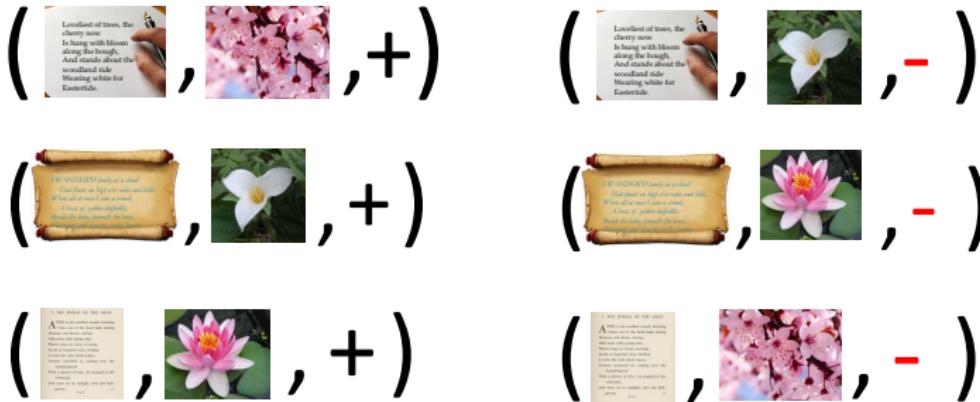
Reduction to binary classification

Simple recipe for positive and unlabelled learning [Elkan and Noto, 2008]:

- Treat unlabelled instances as negatives
- Learn a **class-probability estimator** (e.g. logistic regression)
- Optimal for ranking purposes!

Thus, we can perform **logistic regression** on the **labelled instances**

$$\mathcal{D}'' = \{((a^{(i)}, b^{(j)}), (2\llbracket i=j \rrbracket - 1))\}_{i,j=1}^n$$



Choice of feature mapping

We consider ℓ_2 regularised logistic regression with some **linear scorer**:

$$\min_w \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \log(1 + e^{-z^{(i,j)} \cdot s(a^{(i)}, b^{(j)}; w)}) + \frac{\lambda}{2} \|w\|_2^2$$

where

$$s(a, b; w) = \langle w, \Phi(a, b) \rangle$$

for some **feature mapping** Φ

Simplest Φ is concatenation:

$$\Phi(a, b) = [a \quad b]$$

Choice of feature mapping

We consider ℓ_2 regularised logistic regression with some **linear scorer**:

$$\min_w \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \log(1 + e^{-z^{(i,j)} \cdot s(a^{(i)}, b^{(j)}; w)}) + \frac{\lambda}{2} \|w\|_2^2$$

where

$$s(a, b; w) = \langle w, \Phi(a, b) \rangle$$

for some **feature mapping** Φ

Simplest Φ is concatenation:

$$\Phi(a, b) = [a \quad b]$$

Problem: Induces the **same ranking** over $b \in \mathcal{B}$ for **any** $a \in \mathcal{A}$!

Choice of feature mapping

Better Φ :

$$\Phi(a, b) = [a_d \cdot b_{d'}]_{d,d'}$$

i.e. consider the **cross-product** of features

The similarity function is equivalently:

$$s(a, b; w) = \langle w, \Phi(a, b) \rangle = a^T W b$$

- c.f. bilinear regression
- CCA \approx low rank approximation to W

Choice of feature mapping

Better Φ :

$$\Phi(a, b) = [a_d \cdot b_{d'}]_{d,d'}$$

i.e. consider the **cross-product** of features

The similarity function is equivalently:

$$s(a, b; w) = \langle w, \Phi(a, b) \rangle = a^T W b$$

- c.f. bilinear regression
- CCA \approx low rank approximation to W

Learn, for each **pair of features** across the modalities, their predictive power in determining affinity

Overview of talk

- 1 Overview
- 2 Cross-modal retrieval: unsupervised and supervised
- 3 The similarity function perspective
- 4 Learning without annotations
- 5 Learning with annotations
- 6 Experiments
- 7 Conclusion

Learning with annotations

With ground-truth annotations, we still **only observe pairs that are linked**

Training set: $\mathcal{D}' = \{((a^{(i)}, b^{(i)}), y^{(i)})\}_{i=1}^n$

- $y^{(i)} \in \mathcal{Y}$ is the category for the i th entity



Learning with annotations

With ground-truth annotations, we still **only observe pairs that are linked**

Training set: $\mathcal{D}' = \{((a^{(i)}, b^{(i)}), y^{(i)})\}_{i=1}^n$

- $y^{(i)} \in \mathcal{Y}$ is the category for the i th entity



However, using annotations, we can **more accurately infer** whether other pairs should be linked or not

Pairwise classification approach

Basic idea: Construct **all pairs** of instances from each modality

$$\mathcal{D}'' = \{((a^{(i)}, b^{(j)}), z^{(i,j)})\}_{i,j=1}^n$$

where $z^{(i,j)} = 2[\![y^{(i)} = y^{(j)}]\!] - 1$

- Link a pair iff ground-truth annotation coincides

Now learn logistic regression model as in the case without annotations

- Only changing the labelling procedure

An alternate approach

There is a simpler approach to learning with annotations

If well-specified, logistic regression will learn a transformation of

An alternate approach

There is a simpler approach to learning with annotations

If well-specified, logistic regression will learn a transformation of

$$\begin{aligned}\mathbb{P}(Z = 1 | A = a, B = b) &= \mathbb{P}(Y = Y' | A = a, B = b) \\ &= \sum_{k=1}^K \mathbb{P}(Y = k | A = a) \cdot \mathbb{P}(Y = k' | B = b) \\ &= \langle \eta_A(a), \eta_B(b) \rangle,\end{aligned}$$

where η_A and η_B are vectors of instance-conditional distributions over labels

An alternate approach

There is a simpler approach to learning with annotations

If well-specified, logistic regression will learn a transformation of

$$\begin{aligned}\mathbb{P}(Z = 1 | A = a, B = b) &= \mathbb{P}(Y = Y' | A = a, B = b) \\ &= \sum_{k=1}^K \mathbb{P}(Y = k | A = a) \cdot \mathbb{P}(Y = k' | B = b) \\ &= \langle \eta_A(a), \eta_B(b) \rangle,\end{aligned}$$

where η_A and η_B are vectors of instance-conditional distributions over labels

Why not just directly learn η_A, η_B ?

Marginal approach

Basic idea: Learn **independent** models for η_A and η_B

- e.g. using multi-class logistic regression on $\{(a^{(i)}, y^{(i)})\}$ and $\{(b^{(i)}, y^{(i)})\}$

Use their **dot-product** as the similarity function

- c.f. use of cosine similarity in SM method

While individual probability models are linear, final similarity function will be **nonlinear** in the inputs

Overview of talk

- 1 Overview
- 2 Cross-modal retrieval: unsupervised and supervised
- 3 The similarity function perspective
- 4 Learning without annotations
- 5 Learning with annotations
- 6 Experiments
- 7 Conclusion

Overview of experiments

Experiments on three datasets:

- WIKIPEDIA: 2,173 training, 693 test pairs; 10 semantic categories.
- PASCAL: 700 training, 300 test pairs; 20 semantic categories.
- TVGRAZ: 1,558 training, 500 testing pairs; 10 semantic categories.

Feature representations:

- **Text:** LDA topic assignment distributions
- **Images:** SIFT features

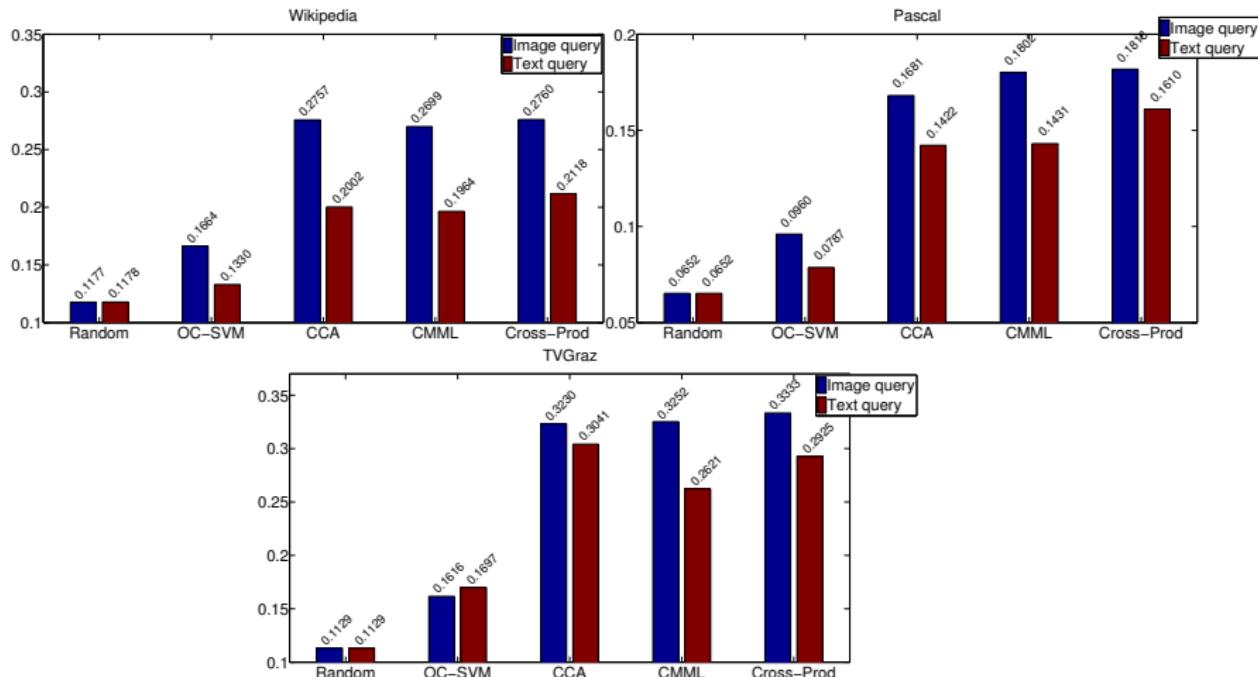
Methods compared

Baselines:

- Random predictor
- OC-SVM
- CCA
- SM, SCM
- Cross-modal metric learning (CMLL)

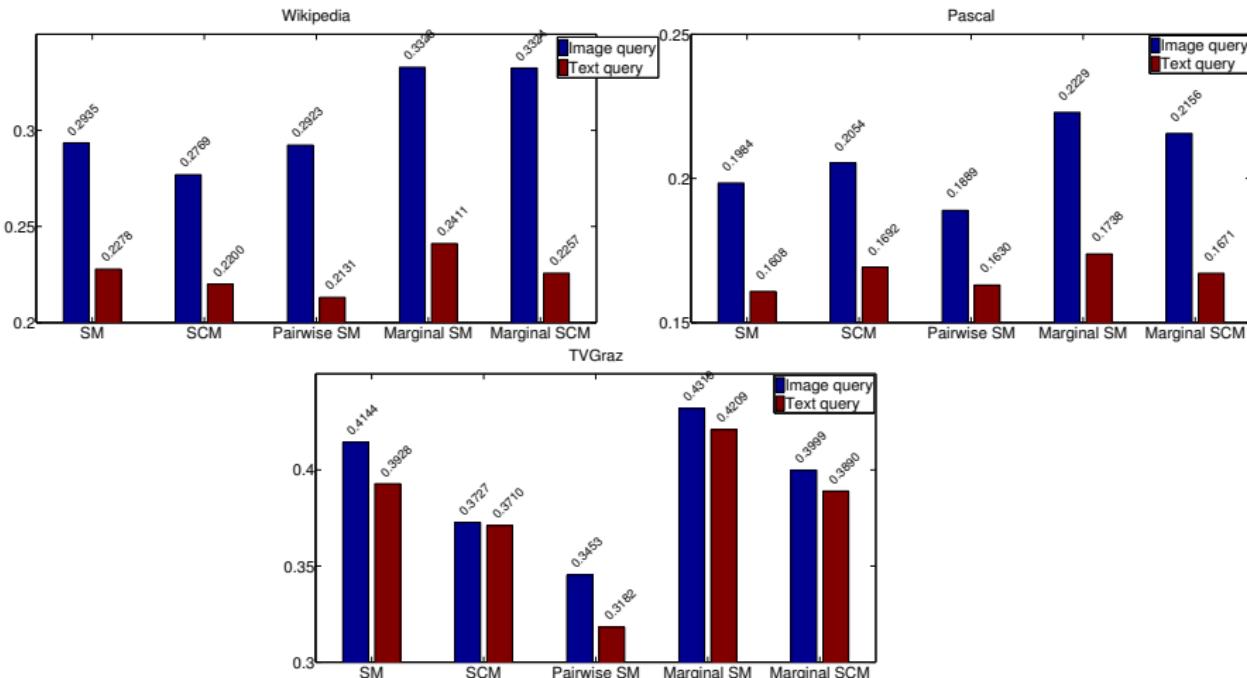
Measure performance based on **mean average precision** (MAP)

Results without annotation



MAP score improvements to CCA: ~2% on WIKIPEDIA, ~10% on PASCAL, and ~3% on TVGRAZ.

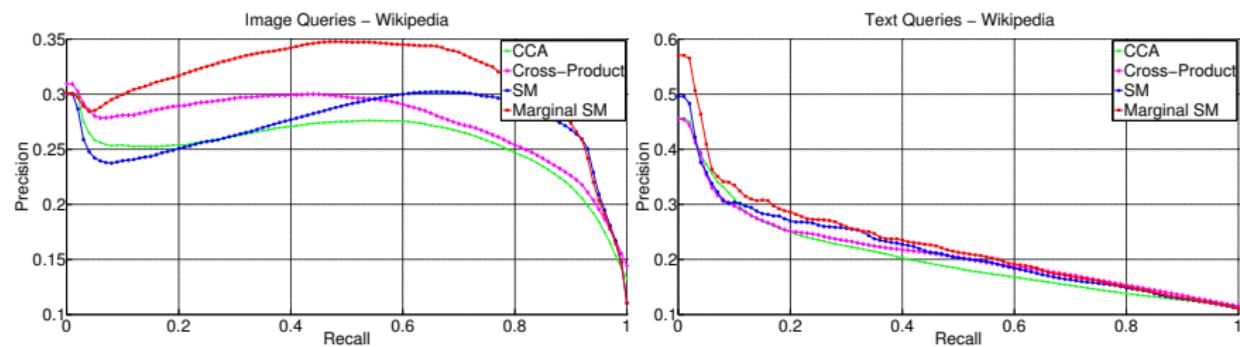
Results with annotation



MAP score improvements to SM: ~10% on WIKIPEDIA, ~10% on PASCAL, and ~5% on TVGRAZ.

Precision-recall curves on WIKIPEDIA

Consistent dominance of cross-product method over CCA, marginal over SM



Case study: text query

Given a text, find images

TEXT QUERY

Cobby was a member of the RAAF Reserve (also known as the Citizen Air Force) during his time with the Civil Aviation Board, and rejoined the Permanent Air Force following the outbreak of World War II ...



Image Results

cross-product



Warfare



Warfare



Warfare



Warfare

CCA



Warfare



Warfare



Geography



Music

CCA suffers from not considering negative links

Case study: image query

Given an image (top), find texts

IMAGE QUERY



Geography

Text Results

Cross-product

Most of the Columbia's drainage basin (which, at, is about the size of France) lies roughly between the ...

Until the 20th century, the slough and its side channels and associated ponds and lakes were part of the active...

The Columbia begins its journey in the southern Rocky Mountain Trench in British Columbia (BC) ...

There are over 100 km of hiking trails at Worlds End State Park. Most of the trails are rocky and steep, so hikers are ...

The history of Kaziranga as a protected area can be traced back to 1904, when Mary Victoria Leiter Curzon ...

As of the census of 2006, there were 382,872 people, 165,743 households, and 99,114 families residing in ...

Various public sector organisations have an important role in the stewardship of the country's fauna ...

Only humans are recognized as persons and protected in law by the United Nations Universal Declaration of ...



Overview of talk

- 1 Overview
- 2 Cross-modal retrieval: unsupervised and supervised
- 3 The similarity function perspective
- 4 Learning without annotations
- 5 Learning with annotations
- 6 Experiments
- 7 Conclusion

Conclusion

Pairwise classification perspective on cross-modal retrieval

- Unified perspective in the presence and absence of ground truth annotations

Logistic regression approach for learning without and with annotations

- **Without:** reduction to positive and unlabelled learning
- **With:** learn probability distributions for each modality

Good empirical performance

Questions?