AMECON: Abstract Meta-Concept Features for Text Illustration

Ines Chami¹,*, Youssef Tamaazousti²,*) and Hervé Le Borgne²
1: Stanford University, USA – 2: CEA LIST, FRANCE

*) Both authors contributed equally
Text-illustration System

Textual query: a man cycling on a mountain

Most appropriate images:
Cross-Modal Retrieval Task

• Cross-Modal Retrieval task
  • Given a document in one modality, find (from database) the most relevant documents in another modality

• Text-illustration
  • Query: sentences
  • Collection: images

• Hard problem: semantic gap
Cross-Modal Retrieval Approach 1

- Canonical Correlation Analysis
  - Hardoon et al. Neural Computation 2004
  - Hwang and Grauman, IJCV 2012
  - Costa Pereira et al. TPAMI 2014
  - Tran et al., CVPR 2016
  - etc.
**Cross-Modal Retrieval Approach 2**

- **Neural Network (NN)**
  - Karpathy and Fei-Fei, NIPS 2014
  - Yan and Mikolajczyk, CVPR 2015
  - Karpathy and Fei-Fei, CVPR 2015
  - Mao et al., ICLR 2015
  - Kiros et al., TACL 2015
  - Wang et al., CVPR 2016
  - etc.
Main Principle of NN Approach

a man is kite-surfing on the water

Text2Common

Image2Common
Main Principle of NN Approach

a man is kite-surfing on the water

Text2Common

1 Symmetry

Image2Common
Main Principle of NN Approach

1. Symmetry
2. Joint Training

- a man is kite-surfing on the water

Text2Common

Image2Common
Main Principle of NN Approach

- 1. Symmetry
- 2. Text2Common
- 3. Image2Common
- 4. Joint Training
- 5. Common Space

a man is kite-surfing on the water
This work: New Approach

a man is kite-surfing on the water

→

Text2Amecon

→

Image2Amecon

→
This work: New Approach

Supervised

Text2Amecon

Unsupervised

Image2Amecon

Asymmetry

1

a man is kite-surfing on the water
This work: New Approach

1. Asymmetry
   - Supervised
     - Text2Amecon
   - Unsupervised
     - Image2Amecon

2. Independent Training

a man is kite-surfing on the water
This work: New Approach

1. Asymmetry
2. Independent Training
3. AMECON Space

Supervised

Text2Amecon

Unsupervised

Image2Amecon

a man is kite-surfing on the water
AMECON principle

- AMECON: Abstract Meta-CONcept
  - Abstract-concept + Meta-concept
AMECON principle

- AMECON: Abstract Meta-CONcept
  - Abstract-concept + Meta-concept
AMECON principle

• AMECON: Abstract Meta-CONcept

• Abstract-concept + Meta-concept
AMECON principle

- **AMECON: Abstract Meta-CONcept**
- **Abstract-concept + Meta-concept**
AMECON principle

• AMECON: Abstract Meta-CONcept
  • Abstract-concept + Meta-concept
Overview of Our Approach

a man is kite-surfing on the water

→

Text2Amecon

AMECON Space

→

Image2Amecon
Learning Text2Amecon block

- **Learning Textual Features**
  a. Select all different words from training-data
  b. Remove stop-words (``is'', ``of'', ``for'', etc.)
  c. Compute word2vec features for each word
  d. Cluster (k-means) the whole set of features
Learning Text2Amecon block

- **Learning Textual Features**
  a. Select all different words from training-data
  b. Remove stop-words (``is'', ``of'', ``for'', etc.)
  c. Compute word2vec features for each word
  d. Cluster (k-means) the whole set of features

![Diagram showing steps a, b, c, and d]

**AMECONs**
Computing Textual AMECON Features

Test phase

Textual AMECON Features

a man is cycling on a mountain

1 0 1 0 0 0 1 0
Overview of Our Approach

a man is kite-surfing on the water

→

Text2Amecon

→

AMECON Space

→

Image2Amecon
Learning Image2Amecon block
Learning Image2Amecon block
Learning Image2Amecon block

\[
\begin{align*}
\text{a man is cycling on a mountain} & \quad \Rightarrow \quad \text{Word2Vec} \\
\end{align*}
\]
Learning Image2Amecon block
Learning Image2Amecon block

Word2Vec

ConvNet

MLP

Pre-trained and fixed weights

Randomly initialized weights

Ground-Truth

a man is cycling on a mountain
Learning Image2Amecon block

Unsupervised

Supervised

Word2Vec

ConvNet

Pre-trained and fixed weights

Randomly initialized weights

Ground-Truth

MLP

a man is cycling on a mountain
Computing Visual AMECON Features

Test phase

Visual AMECON Features
Overview of Our Approach

a man is kite-surfing on the water

\[ \rightarrow \text{Text2Amecon} \]

\[ \rightarrow \text{Image2Amecon} \]

\[ \rightarrow \text{AMECON Space} \]
Matching Multi-Modal Data in AMECON Space

• Matching texts & images in the same AMECON Space

• Text and Images directly comparable

• Perform ANY multi-modal task
Text-Illustration in AMECON Space

Two **dogs** are **playing** on the **grass**

QUERY
Text-Illustration in AMECON Space

COLLECTION

Two dogs are playing on the grass

QUERY
Text-Illustration in AMECON Space

COLLECTION

Two dogs are playing on the grass

QUERY

Cosine similarity
Amecon Features of data query
Amecon Features of data collection
Text-Illustration in AMECON Space

COLLECTION

Two dogs are playing on the grass

QUERY

Cosine similarity

Amecon Features of data query
Amecon Features of data collection
Experimental Protocol

• Training data
  • 6,000/30,000 images in Flickr-8k/Flickr-30k
  • Each image associated to 5 captions

• Testing data (same for Flickr-8k & 30k)
  • 1000 images and 5000 captions
  • All captions as data-queries
  • All images as data-collection
  • Evaluation metric: Recall@K (K = 1, 5, 10)
## Text Illustration Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Denotation</th>
<th>Flicker-8k R@1</th>
<th>Flicker-8k R@5</th>
<th>Flicker-8k R@10</th>
<th>Flicker-30k R@1</th>
<th>Flicker-30k R@5</th>
<th>Flicker-30k R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karpathy <em>et al.</em> [17]</td>
<td>DeFrag</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kiros <em>et al.</em> [18]</td>
<td>MNLM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mao <em>et al.</em> [21]</td>
<td>m-RNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Karpathy <em>et al.</em> [16]</td>
<td>BRNN*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yan <em>et al.</em> [36]</td>
<td>DCCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tran <em>et al.</em> [32]</td>
<td>MACC†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Our Approach</strong></td>
<td>AMECON</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Neural Network-based Approach**

**CCA-based Approach**

**Our Approach**
## Text Illustration Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Denotation</th>
<th>Flickr-8k</th>
<th>Flickr-30k</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>Karpathy et al. [17]</td>
<td>DeFrag</td>
<td>9.7</td>
<td>10.3</td>
</tr>
<tr>
<td>Kiros et al. [18]</td>
<td>MNLM</td>
<td>10.4</td>
<td>11.8</td>
</tr>
<tr>
<td>Mao et al. [21]</td>
<td>m-RNN</td>
<td>11.5</td>
<td>12.6</td>
</tr>
<tr>
<td>Karpathy et al. [16]</td>
<td>BRNN*</td>
<td>11.8</td>
<td>15.2</td>
</tr>
<tr>
<td>Yan et al. [36]</td>
<td>DCCA</td>
<td>12.7</td>
<td>12.6</td>
</tr>
<tr>
<td>Tran et al. [32]</td>
<td>MACC†</td>
<td>10.2</td>
<td>12.1</td>
</tr>
<tr>
<td>Our Approach</td>
<td>AMECON</td>
<td>15.9</td>
<td>18.3</td>
</tr>
</tbody>
</table>

### Neural Network-based Approach
### CCA-based Approach
### Our Approach
## Text Illustration Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Denotation</th>
<th>Flickr-8k</th>
<th></th>
<th>Flickr-30k</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
</tr>
<tr>
<td>Karpathy et al. [17]</td>
<td>DeFrag</td>
<td>9.7</td>
<td>29.6</td>
<td></td>
<td>10.3</td>
</tr>
<tr>
<td>Kiros et al. [18]</td>
<td>MNLM</td>
<td>10.4</td>
<td>31.0</td>
<td></td>
<td>11.8</td>
</tr>
<tr>
<td>Mao et al. [21]</td>
<td>m-RNN</td>
<td>11.5</td>
<td>31.0</td>
<td></td>
<td>12.6</td>
</tr>
<tr>
<td>Karpathy et al. [16]</td>
<td>BRNN*</td>
<td>11.8</td>
<td>32.1</td>
<td></td>
<td>15.2</td>
</tr>
<tr>
<td>Yan et al. [36]</td>
<td>DCCA</td>
<td>12.7</td>
<td>31.2</td>
<td></td>
<td>12.6</td>
</tr>
<tr>
<td>Tran et al. [32]</td>
<td>MACC†</td>
<td>10.2</td>
<td>29.3</td>
<td></td>
<td>12.1</td>
</tr>
<tr>
<td><strong>Our Approach</strong></td>
<td>AMECON</td>
<td>15.9</td>
<td>37.9</td>
<td></td>
<td>18.3</td>
</tr>
</tbody>
</table>

**Neural Network-based Approach**

**CCA-based Approach**

**Our Approach**
Text Illustration Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Denotation</th>
<th>Flickr-8k</th>
<th></th>
<th>Flickr-30k</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R@1</td>
<td>R@5</td>
<td>R@10</td>
<td>R@1</td>
</tr>
<tr>
<td>Karpathy et al. [17]</td>
<td>DeFrag</td>
<td>9.7</td>
<td>29.6</td>
<td>42.5</td>
<td>10.3</td>
</tr>
<tr>
<td>Kiros et al. [18]</td>
<td>MNLM</td>
<td>10.4</td>
<td>31.0</td>
<td>43.7</td>
<td>11.8</td>
</tr>
<tr>
<td>Mao et al. [21]</td>
<td>m-RNN</td>
<td>11.5</td>
<td>31.0</td>
<td>42.4</td>
<td>12.6</td>
</tr>
<tr>
<td>Karpathy et al. [16]</td>
<td>BRNN*</td>
<td>11.8</td>
<td>32.1</td>
<td>44.7</td>
<td>15.2</td>
</tr>
<tr>
<td>Yan et al. [36]</td>
<td>DCCA</td>
<td>12.7</td>
<td>31.2</td>
<td>44.1</td>
<td>12.6</td>
</tr>
<tr>
<td>Tran et al. [32]</td>
<td>MACC†</td>
<td>10.2</td>
<td>29.3</td>
<td>41.1</td>
<td>12.1</td>
</tr>
<tr>
<td>Our Approach</td>
<td>AMECON</td>
<td>15.9</td>
<td>37.9</td>
<td>49.5</td>
<td>18.3</td>
</tr>
</tbody>
</table>

Neural Network-based Approach
CCA-based Approach
Our Approach
Analysis of Parameters

• Quite robust to the parameters
  • Robust to #selected neighbours
  • Sensitive to #clusters (C) but stable when for a large range of C values
Conclusion

• Novelty:
  • Principle of AMECONs
    • Abstract MEta-CONcepts
  • Mixing supervised and unsupervised learning to build a multi-modal space

• Results on Text-illustration:
  • +4 points of R@K (avg.) compared to best methods of the literature

• Future Work:
  • Image captioning with AMECON-features
Code will be released at:
http://perso.ecp.fr/~tamaazouy/

Thank you (questions ?)