Aligning Images and Text in a Digital Library

Jack Hessel and David Mimno
Aligning Images and Text in a Digital Library

Jack Hessel and David Mimno
Aligning text and images

- Text and images
- The British Library corpus
- Why is this hard?
- Neural Network alignment
- Alignment examples
Aligning text and images

- Text and images
- The British Library corpus
- Why is this hard?
- Neural Network alignment
- Alignment examples
Then, when he saw dawn breaking over beach and sea, he yoked his horses to his chariot, and bound the body of Hektor behind it that he might drag it about. Thrice did he drag it round the tomb [sêma] of the son of Menoitios, and then went back into his tent, leaving the body on the ground full length and with its face downwards. But Apollo would not suffer it to be disfigured, for he pitied the man, dead though he now was; therefore he shielded him with his golden aegis continually, that he might take no hurt while Achilles was dragging him. (Samuel Butler trans.)
The Boston Evening Transcript

BY T. S. ELIOT

The readers of the Boston Evening Transcript
Sway in the wind like a field of ripe corn.

When evening quickens faintly in the street,
Wakening the appetites of life in some
And to others bringing the Boston Evening Transcript,
I mount the steps and ring the bell, turning
Wearily, as one would turn to nod good-bye to Rochefoucauld,
If the street were time and he at the end of the street,
And I say, "Cousin Harriet, here is the Boston Evening Transcript."

(Poetry Foundation)
Goals

• Scholars need tools to trace semantic and aesthetic qualities through time and space, tracking influence and allusion that cannot be easily measured otherwise.

• Use text to provide context for the images we see in digital libraries, and as a noisy "label" for computer vision tasks.

• Use images to provide grounding for text.
Are these related?
What color are sheep?
What color are sheep?
Previous multi-modal work

- **Datasets**
  - Microsoft Common Objects in Context (COCO): Images labeled by anonymous workers on Amazon Mechanical Turk.
  - Flickr: Images with user-provided tags.

- **Tasks**
  - Cross-modal information retrieval (ImageCLEF)
  - Caption search / generation
Earlier work


Recent work


Aligning text and images

- Text and images
- The British Library corpus
- Why is this hard?
- Neural Network alignment
- Alignment examples
Source: Books from the British Library

- Released by British Library https://data.bl.uk/digbks/: "The datasets in this collection are comprised and derived from 49,455 digitised books (65,227 volumes) largely from the 19th Century."

- We use images from the "plates" and "medium sized" categories, labels provided.

- 405K images associated with a mean of 2.3K tokens. The average token-to-type ratio is 2.5.
FIRST APPROACH TO TIMBUKTU.
His Royal Highness Prince Frederick George Duke of Gloucester born Janu. 20. 1708.
The Arms of the Family of Issachar.

Behold how low the Couchant Beast does kneel.
NANCY ETTICOA
A musical rhymeland phantasy

Written and composed by E. Lucia Turnbull

Pictured by Captain Barry-Briant

Book of Words
Aligning text and images

- Text and images
- The British Library corpus
- Why is this hard?
- Neural Network alignment
- Alignment examples
Text/image links

- Variation in relationships between text and images.

- COCO / Flickr are strongly aligned, Wikipedia / BL are weakly aligned.

- Both are available for cultural heritage materials, but weak alignment is more abundant.
Here’s me at the British Library!
Here’s me at the British Library!
Here’s me at the British Library!
Edward Geary Lansdale (February 6, 1908 – February 23, 1987)[1] was a United States Air Force officer who served in the Office of Strategic Services and the Central Intelligence Agency (CIA). He rose to the rank of Major General and was awarded the Distinguished Service Medal in 1963. He was an early proponent of more aggressive U.S. actions in the Cold War. Lansdale was born in Detroit, Michigan and died in McLean, Virginia.

Micronesia (from Greek: μικρός, mikrós, "small" + Greek: νῆσος, nêsos, "island") is a subregion of Oceania, comprising thousands of small islands in the western Pacific Ocean. It has a shared cultural history with two other island regions, Polynesia to the east and Melanesia to the south. The region has a tropical marine climate, and is part of the Oceania ecozone. There are four main archipelagos along with numerous outlying islands.
Aligning text and images

- Text and images
- The British Library corpus
- Why is this hard?
  - Neural Network alignment
- Alignment examples
Features

RESNET 50

0.143
0.353
0.823
0.064
0.546
0.745
0.319
Russia has invariably been friendly to us. Far back, when we were just entering the society of nations, it was an ordeal to go as Minister representing not only a new and unknown Power, but one in the dreaded form of a Republic. Franklin had the only good post, that to Paris, where the local and growing political feeling made a welcoming party for him.
Russia has invariably been friendly to us. Far back, when we were just entering the society of nations, it was an ordeal to go as Minister representing not only a new and unknown Power, but one in the dreaded form of a Republic. Franklin had the only good post, that to Paris, where the local and growing political feeling made a welcoming party for him.
Russia has invariably been friendly to us. Far back, when we were just entering the society of nations, it was an ordeal to go as Minister representing not only a new and unknown Power, but one in the dreaded form of a Republic. Franklin had the only good post, that to Paris, where the local and growing political feeling made a welcoming party for him.
Volume-level holdout

- At test time, we hold out 5K full volumes, and sample only 1 image from each, to ensure that no training images are in the testing data, and no testing images are sampled from the same volumes that the training images are.

- We only sample at most 10 images from each volume during training to ensure that the sample isn't skewed (a few volumes contain a large proportion of the images).
<table>
<thead>
<tr>
<th></th>
<th>NN</th>
<th>IM2TX</th>
<th>DCCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1.0/1.0</td>
<td>1.0/1.0</td>
<td>1.1/1.0</td>
</tr>
<tr>
<td>Unigram</td>
<td>10.9/11.0</td>
<td>8.8/14.4</td>
<td>15.8/15.3</td>
</tr>
<tr>
<td>PV</td>
<td>13.5/11.8</td>
<td>17.1/18.5</td>
<td>17.8/17.8</td>
</tr>
<tr>
<td>LDA</td>
<td>12.0/8.9</td>
<td>17.1/18.0</td>
<td>17.0/16.4</td>
</tr>
</tbody>
</table>

Table 1: Recall-at-1% for image-search-text/text-search-image, averaged over ten cross-validation splits; higher is better. The parametric approaches IM2TX and DCCA improve over the NN baseline, though all outperform random.
Aligning text and images

- Text and images
- The British Library corpus
- Why is this hard?
- Neural Network alignment

- Alignment examples
Russia has invariably been friendly to us. Far back, when we were just entering the society of nations, it was an ordeal to go as Minister representing not only a new and unknown Power, but one in the dreaded form of a Republic.

Franklin had the only good post, that to Paris, where the local and growing political feeling made a welcoming party for him.
Russia has invariably been friendly to us. Far back, when we were just entering the society of nations, it was an ordeal to go as Minister representing not only a new and unknown Power, but one in the dreaded form of a Republic. Franklin had the only good post, that to Paris, where the local and growing political feeling made a welcoming party for him.
RESNET 50 → X → LDA
Good job!
Original

- iron furnace acid gas fig process copper air heat fire 0.60
- made end side long iron wood hand work round cut 0.098
- boat boats water shore canoe landing oars river canoes men 0.042
- steam fig cylinder engine shaft water pressure pump valve inch 0.036
- great time present part form generally large found number small 0.023

Predicted

- iron furnace acid gas fig process copper air heat fire 0.051
- made end side long iron wood hand work round cut 0.038
- gold ore mill water stamp solution battery silver tailings ores 0.034
- steam fig cylinder engine shaft water pressure pump valve inch 0.032
- time made work make great means place found con purpose 0.025
These furnaces should be deeper than the preceding one.

- Iron furnace acid gas, fig. process copper air, heat fire 0.49
- Made end side long, iron wood hand work, round cut 0.122
- Gold ore mill, water stamp solution battery, silver tailings ores 0.115
- Made point line plan, case general found, work position time 0.076
- Steam fig cylinder, engine shaft water pressure pump valve, inch 0.040
- Great time present part form generally large, found number small 0.035
• view beautiful scenery
place picturesque fine
village town situated
beauty 0.256

• rock rocks mountain
feet wild scene deep
water waters
mountains 0.107

• trees village green hill
country long wood
hills road forest 0.071

• great time present part
form generally large
found number small
0.057

• ofthe part con account
ion country pro
present general state
0.052

• thy eye nature till ring
mind oft vain tis pride
0.039

• london street author
printed vol edition
illustrations john
volumes reserved 0.037

• ofthe part con account
ion country pro
present general state
0.037
thy eye nature till ring
till mind oft vain tis pride

wine table tea bread
dinner drink good
water meat eat

sweet love song day
summer flowers heart
fair bright green

man world nature
human earth life great
mind power things

life men man great

good character day
world fact nature
Less successful
- London street author
  printed vol edition illustrations john volumes reserved 0.179

- Island bay sea coast miles point shore water land harbour 0.111

- Great time present part form generally large found number small 0.065

- Wales Welsh Conway castle Chester Dee Llandudno north Wye David 0.064

- Rio America Brazil Spanish San Peru Santa Cuba Island south 0.063

- Don good man sir make thought time give thing asked 0.055

- Eyes face hand looked voice moment door stood head back 0.054

- Time found made day left party return received returned visit 0.042

- Life men man great good character day world fact nature 0.042

- Back time round moment head man side feet men began 0.041
Original

- york hull yorkshire ripon bolton abbey scarborough hall north leeds 0.30
- valley mountain mountains hills miles feet range road plain great 0.097
- view beautiful scenery place picturesque fine village town situated beauty 0.066
- town city houses inhabitants miles place streets towns built large 0.028
- power state conduct act manner death fate mind length evil 0.028

Predicted

- view beautiful scenery place picturesque fine village town situated beauty 0.052
- rock rocks mountain feet wild scene deep water waters mountains 0.028
- time made work make great means place found con purpose 0.021
- great time present part form generally large found number small 0.021
- life men man great good character day world fact nature 0.020
Original

- history work published
  book author edition
  account present volume
  historical o.378

- london street house lane
  westminster square
  great inn paul thames
  o.121

- time made work make
  great means place found
  con purpose o.093

- oxford court windsor
  hampton thames
  college richmond queen
  house surrey o.071

- great time present part
  form generally large
  found number small
  o.062

Predicted

- church south nave
  north tower window
  side windows chancel
  east o.048

- sweet love song day
  summer flowers heart
  fair bright green o.034

- building marble great
  columns palace front
  beautiful architecture
  feet centre o.033

- london street author
  printed vol edition
  illustrations john
  volumes reserved o.028

- church rev chapel
  minister congregation
  pastor sunday meeting
  worship school o.026
• chapter chap iii page contents vii viii xii xiii xiv 0.283

• les paris des france french est qui par louis sur 0.122

• london street author printed vol edition illustrations john volumes reserved 0.106

• sword battle fight hand men blood war man arms brave 0.050

• church abbey bishop cathedral chapel monastery monks archbishop abbot saint 0.027

• london street author printed vol edition illustrations john volumes reserved 0.045

• life men man great good character day world fact nature 0.041

• time found made day left party return received returned visit 0.036

• time made work make great means place found con purpose 0.032

• power state conduct act manner death fate mind length evil 0.024
Original

- water sand waters feet spring stream springs great surface salt 0.389
- rocks beds limestone strata sandstone clay rock geological geology formation 0.112
- iron furnace acid gas fig process copper air heat fire 0.079
- great time present part form generally large found number small 0.073
- cave rock caves cavern entrance caverns roof grotto floor rocks 0.067

Predicted

- life men man great good character day world fact nature 0.024
- thy eye nature till ring mind oft vain tis pride 0.023
- time made work make great means place found con purpose 0.022
- rocks beds limestone strata sandstone clay rock geological geology formation 0.022
- rock rocks mountain feet wild scene deep water waters mountains 0.021
What next?

- Why are some concepts easier to match across modalities?
- How can we give users the ability to search for themes across modalities?
- Can multimodality separate meaning from appearance?
- How limiting are tiny, square images?
• Images and text complement one another
• Existing algorithms do a pretty good job!
• Cultural data is more complicated than purpose-built CV datasets
• “Broken” models and “errors” can be more interesting than “success”
• More useful analysis will come from better understanding of datasets, not necessarily from better algorithms

Jack Hessel (jhessel@cs.cornell.edu)
David Mimno (mimno@cornell.edu)