Multimodal Content-Aware Image Thumbnailing

Kohei Yamamoto† Hayato Kobayashi‡ Yukihiro Tagami‡ Hideki Nakayama†
† The University of Tokyo ‡Yahoo Japan Corporation
{yamamoto, nakayama}@nlab.ci.i.u-tokyo.ac.jp,  {hakobaya, yutagami}@yahoo-corp.jp

The University of Tokyo  Yahoo Japan Corporation

Kohei Yamamoto  Hayato Kobayashi  Yukihiro Tagami  Hideki Nakayama

RCNN

Embed sentences: BRNN

Transform words in a sentence into vectors using bidirectional RNN

Proposed Method

Overview

Input

Image

RCNN

TOP 100 Classification score

b-box

Energy score

Accumulate

Output

Thumbnail

Search final thumbnail

1. Find set of candidate regions

\[
\mathcal{R}(\lambda) = \{ \mathbf{r} \mid \sum_{(x,y) \in \mathbf{r}} E(x,y) \geq \lambda \} \]

\( E(x,y) \): Energy score of (x,y)

\( \mathbf{P} \): set of all pixels in a given image

\( \mathbf{r} \): set of all pixels in a select region (satisfy required aspect ratio)

2. Determine final thumbnail region

\[
\mathcal{R}_C = \{ \mathbf{r} \mid \arg \max_{\mathbf{r} \in \mathcal{R}(\lambda)} \sum_{(x,y) \in \mathbf{r}} E(x,y) \in \mathbf{A}_r \}
\]

\( \lambda \): threshold

\( \mathbf{A}_r \): area of the region \( r \)

\( \mathcal{R} \): set of all regions that satisfy the required aspect ratio

Dataset

Images & texts & thumbnail images extracted from Yahoo! News

Train: 2,654, Test: 300

Original image

Ground truth

Fixed aspect ratio

Early-fusion

Evaluation

- combine each energy map in early-fusion

- IOU > 0.5 for accuracy

Table 1: Experimental results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saliency Map</td>
<td>0.7697</td>
</tr>
<tr>
<td>RCNN-based Saliency Map</td>
<td>0.7333</td>
</tr>
<tr>
<td>Multimodal Saliency Map</td>
<td>0.7633</td>
</tr>
<tr>
<td>Multimodal + Saliency Map</td>
<td>0.7967</td>
</tr>
</tbody>
</table>

Results

Figure 1: Left: Original image, green rectangle is ground truth, Right: left top is Saliency map, right top is RCNN-based, left bottom is Multimodal, right bottom is Saliency maps + Multimodal, Article: "iPhone6s is now on sale... Apple showed the patent..."

Conclusion & Future work

- Proposed method to generate thumbnails that preserve content of images & texts as much as possible

- Saliency map was the worst

- Multimodal model was better than only visual information models

- Saliency + Multimodal (early-fusion) was the best. Combination ratio of each energy map is important

- In our dataset, if adding an energy map derived from face recognition, each energy map is important

- Our method could reflect the content of texts.

- Create a bigger open dataset (now preparing)

- Consider better approach

- Deep attention model / Submodular optimization

- Summarize both images & texts simultaneously

References


Background

- Mobile phones spread rapidly
- Mobile-friendly techniques are important
- Image thumbnailing is important to enhance UX of smartphone applications
- Good image thumbnailing needs specialized skills

Thumbnails

- Reduced-size images maximize visibility in each display & window
- High visibility: Enable to recognize content of images easily
- Requirements
  1. Preserve important structure in the original image as much as possible
  2. Preserve important content of the original image as much as possible
  3. Support any size & aspect ratio

Important structure

Original

Good

Bad

Important content

Depend on what you assume important content

Original

If face,

News, AD...

Scope

If corresponding text,

Good

Bad

• Challenging problems
- Object detection (Regress coordinates)
  Need Ground Truth data of any size & aspect ratio to train
- Content-aware image resizing
  Carry a risk to lose important structure

• Objective
  Propose a image thumbnailing method satisfies above requirements using multimodal information

Proposed Method

Input

Image

RCNN

TOP 100 Classification score

b-box

Energy score

Accumulate

Output

Thumbnail

Search final thumbnail

1. Find set of candidate regions

\[
\mathcal{R}(\lambda) = \{ \mathbf{r} \mid \sum_{(x,y) \in \mathbf{r}} E(x,y) \geq \lambda \} \]

\( E(x,y) \): Energy score of (x,y)

\( \mathbf{P} \): set of all pixels in a given image

\( \mathbf{r} \): set of all pixels in a select region (satisfy required aspect ratio)

2. Determine final thumbnail region

\[
\mathcal{R}_C = \{ \mathbf{r} \mid \arg \max_{\mathbf{r} \in \mathcal{R}(\lambda)} \sum_{(x,y) \in \mathbf{r}} E(x,y) \in \mathbf{A}_r \}
\]

\( \lambda \): threshold

\( \mathbf{A}_r \): area of the region \( r \)

\( \mathcal{R} \): set of all regions that satisfy the required aspect ratio

Dataset

Images & texts & thumbnail images extracted from Yahoo! News

Train: 2,654, Test: 300

Original image

Ground truth

Fixed aspect ratio

Early-fusion

Evaluation

- combine each energy map in early-fusion

- IOU > 0.5 for accuracy

Table 1: Experimental results.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saliency Map</td>
<td>0.7697</td>
</tr>
<tr>
<td>RCNN-based Saliency Map</td>
<td>0.7333</td>
</tr>
<tr>
<td>Multimodal Saliency Map</td>
<td>0.7633</td>
</tr>
<tr>
<td>Multimodal + Saliency Map</td>
<td>0.7967</td>
</tr>
</tbody>
</table>

Results

Figure 1: Left: Original image, green rectangle is ground truth, Right: left top is Saliency map, right top is RCNN-based, left bottom is Multimodal, right bottom is Saliency maps + Multimodal, Article: "iPhone6s is now on sale... Apple showed the patent..."

Conclusion & Future work

- Proposed method to generate thumbnails that preserve content of images & texts as much as possible

- Saliency map was the worst

- Multimodal model was better than only visual information models

- Saliency + Multimodal (early-fusion) was the best. Combination ratio of each energy map is important

- In our dataset, if adding an energy map derived from face recognition, each energy map is important

- Our method could reflect the content of texts.

- Create a bigger open dataset (now preparing)

- Consider better approach

- Deep attention model / Submodular optimization

- Summarize both images & texts simultaneously

References