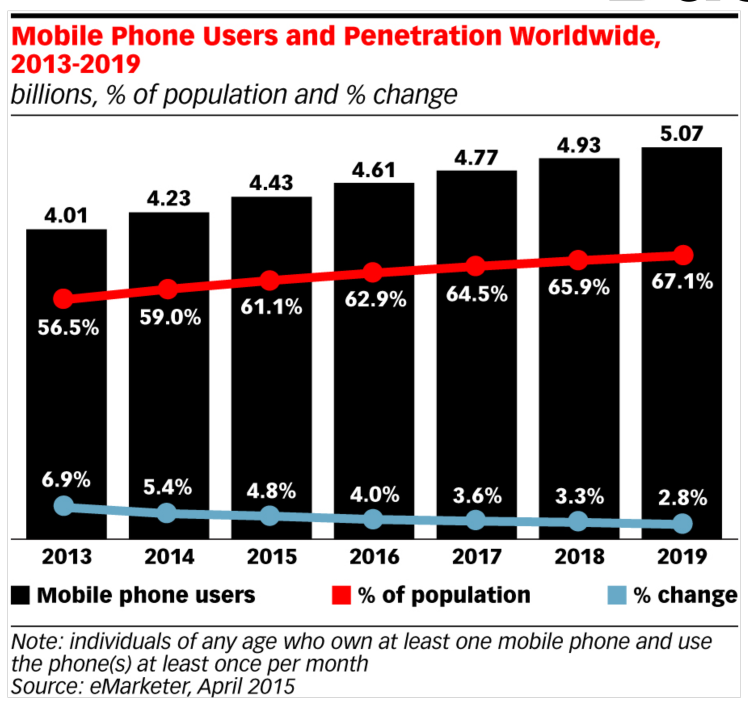
Multimodal Content-Aware Image Thumbnailing

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

Background



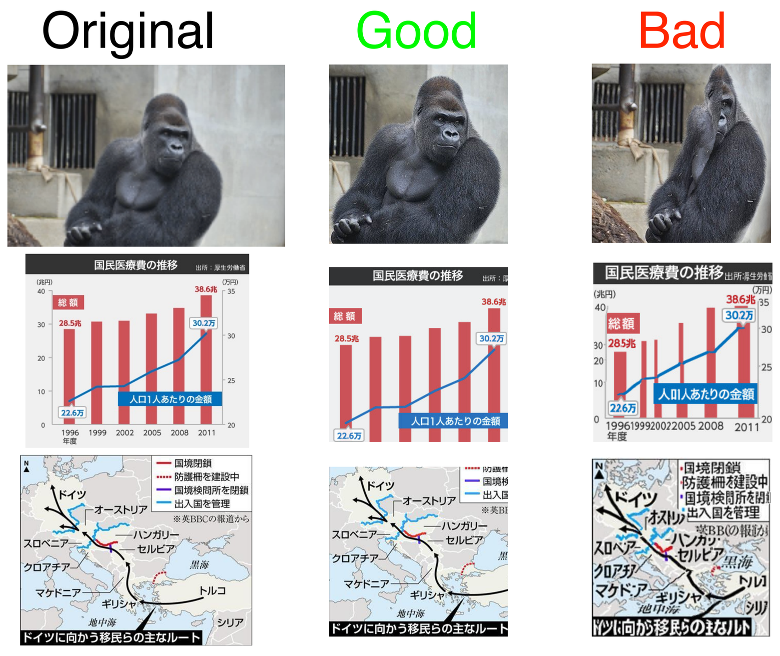
- Smartphones spread rapidly
- Mobile-friendly techniques are important
- Image thumbnailing is important to enhance UX of smartphone applications
- Good image thumbnailing needs specialized skills

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,

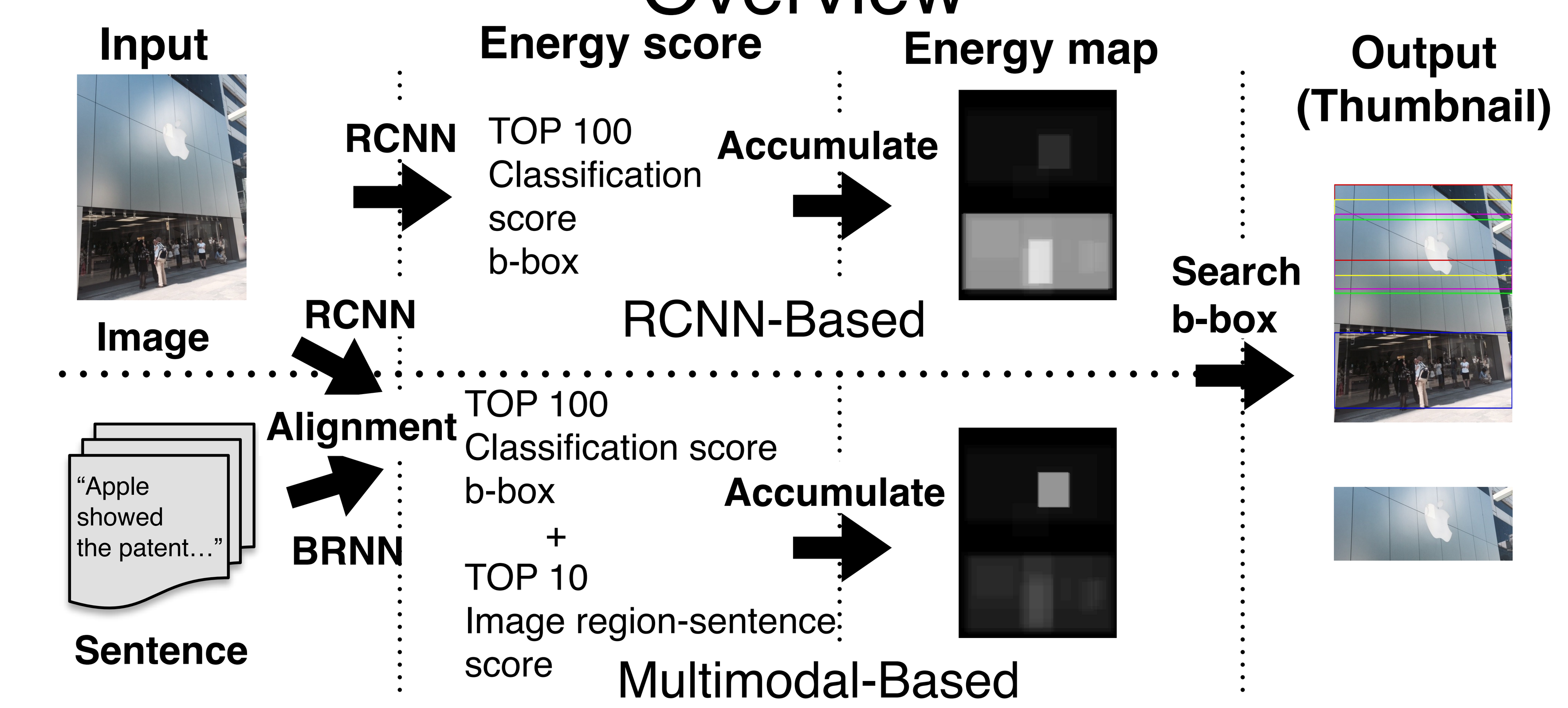
News, AD… **Scope**

If foreground object, If corresponding text,

- **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

Overview



Multimodal alignment[1]

Embed images: RCNN[2]

Detect objects & their regions in an image using region-based CNN $v_i = W_m[\mathrm{CNN}(I_i)]$

Embed sentences: BRNN[3]

Transform words in a sentence into vectors using bidirectional RNN $s_t = \mathrm{BRNN}(\mathbb{I}_t)$

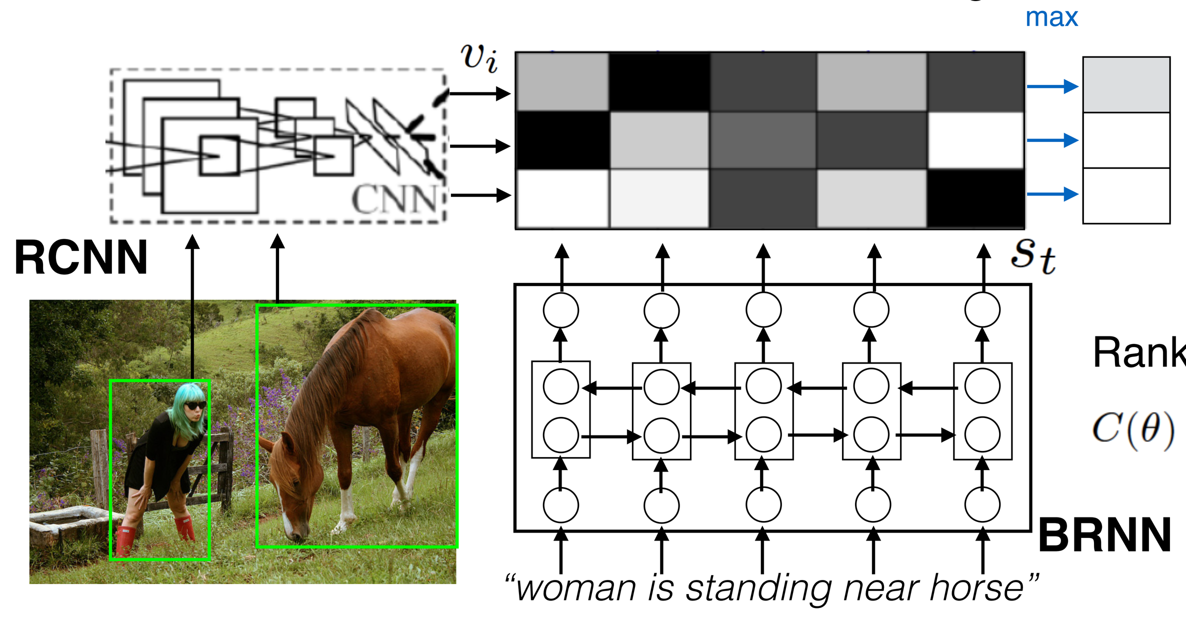


Image-sentence score S_{kl} (image k, sentence l)

$$S_{kl} = \sum_{i \in g_k} \max_{t \in h_l} v_i^T s_t$$

Ranking loss

$$C(\theta) = \sum_k \left[\sum_l \max(0, S_{kl} - S_{kk} + 1) + \sum_l \max(0, S_{lk} - S_{kk} + 1) \right]$$

Search final thumbnail

1. Find set of candidate regions

Preservation ratio of energy

$$\mathfrak{R}(\lambda) = \left\{ r \mid \boxed{\frac{\sum_{(x,y)\in r} E(x,y)}{\sum_{(x,y)\in P} E(x,y)}} > \lambda \right\}$$

$E(x,y)$: Energy score of (x,y)
P : set of all pixels in a given image
r : set of all pixels in a select region (satisfy required aspect ratio)
λ : threshold
$\mathfrak{R}(\lambda)$: set of candidate regions

2. Determine final final thumbnail region

$$R_C = \begin{cases} \underset{r \in \mathcal{R}}{\arg\max} \sum_{(x,y)\in r} E(x,y) & (\mathfrak{R}(\lambda) = \emptyset) \\ \underset{r \in \mathfrak{R}(\lambda)}{\arg\min} A_r & (otherwise) \end{cases}$$

A_r : area of the region r
$\mathcal{R}$: set of all regions that satisfy the required aspect ratio

Experiments

Dataset

Images & texts & thumbnail images extracted from Yahoo! News
Train: 2,654, Test: 300

Original image

2013年に県内で初確認された毒グモ「セアカゴケグモ」（オーストラリア原産）や、最近では、生息域拡大が懸念されるスズメバ

Texts

Ground truth
- Fixed aspect ratio
- Made by professional editor

Set up

- RCNN-based energy map
  - Pre-trained CNN (VGG-16)
  - Use TOP100 classification scores & boxes
- Multimodal energy map
  - Use word embedding matrix initialized 300-dim Word2Vec weights in BRNN
  - Use TOP10 classification score regions
  - Use image region-sentence scores
- Search final region
  - Threshold λ = 1
- Early-fusion
  - combine each energy map in early-fusion
- Evaluation
  - IOU > 0.5 for accuracy

Results

Table 1: Experimental results.

	Accuracy
Saliency Map	0.7067
RCNN-based	0.7533
Multimodal	0.7633
Saliency Map + Multimodal	**0.7967**

Saliency Map[4] wrongly recognized an irrelevant region as the important content since it extract color, edges & brightness changes.

RCNN-based model tended to focus on a region that has many objects.

In this case, **Multimodal** model appropriately cropped the Apple logo.

Our method could reflect the content of texts.

Multimodal
Multimodal + Saliency Map
Saliency map

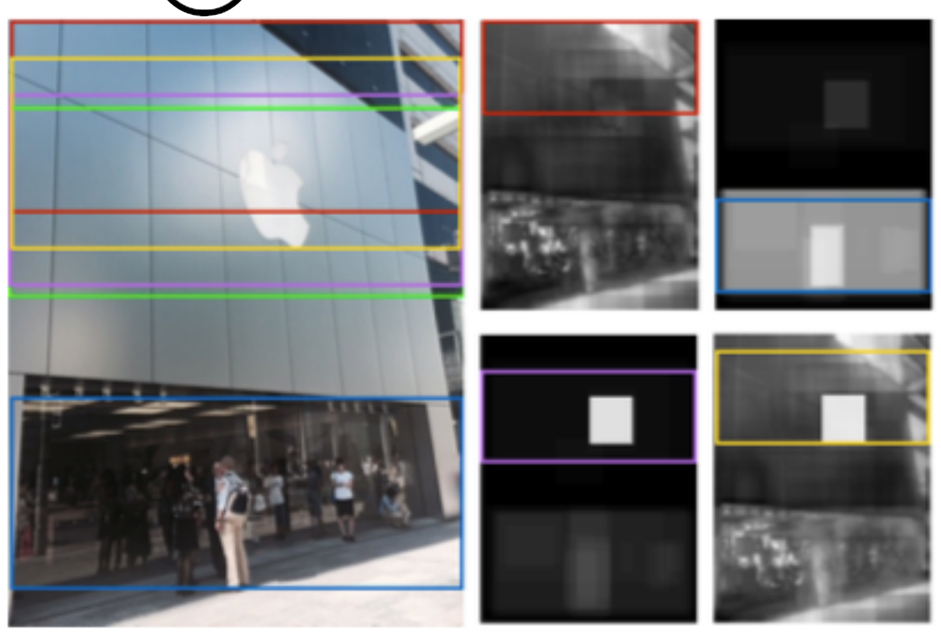


Figure 1: Left: Original image, green rectangle is ground truth. Right: left top is Saliency map, right top is RCNN-based, left bottom is Multimodal, right bottom is Saliency map+Multimodal. Article: "iPhone6s is now on sale … Apple showed the patent… "

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, accuracy may be better

---

- Create a bigger open dataset (now preparing)
- Consider better approach
  - Deep attention model / Submodular optimization
- Summarize both images & texts simultaneously

References

[1] A. Karpathy et al., Deep visual-semantic alignments for generating image descriptions., *In CVPR*, 2015.
[2] R. Girshick et al., Rich feature hierarchies for accurate object detection and semantic segmentation., *In CVPR*, 2014.
[3] M. Schuster et al., Bidirectional recurrent neural networks., *IEEE Transactions on SP*, 1997.
[4] L. Itty et al., A model of saliency-based visual attention for rapid scene analysis., *IEEE Transactions on PAMI*, 1998.