Unreasonable Effectiveness of OCR in Visual Advertisement Understanding

Mayu Otani, Yuki Iwazaki, Kota Yamaguchi
CyberAgent, Inc.

CyberAgent®
AI Lab

Challenges in visual ads

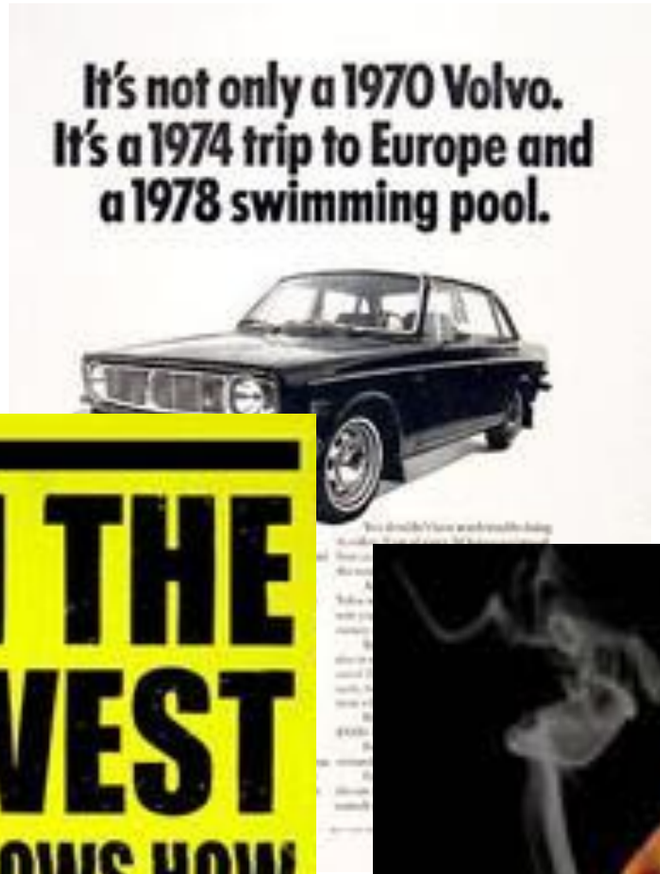
Symbolism

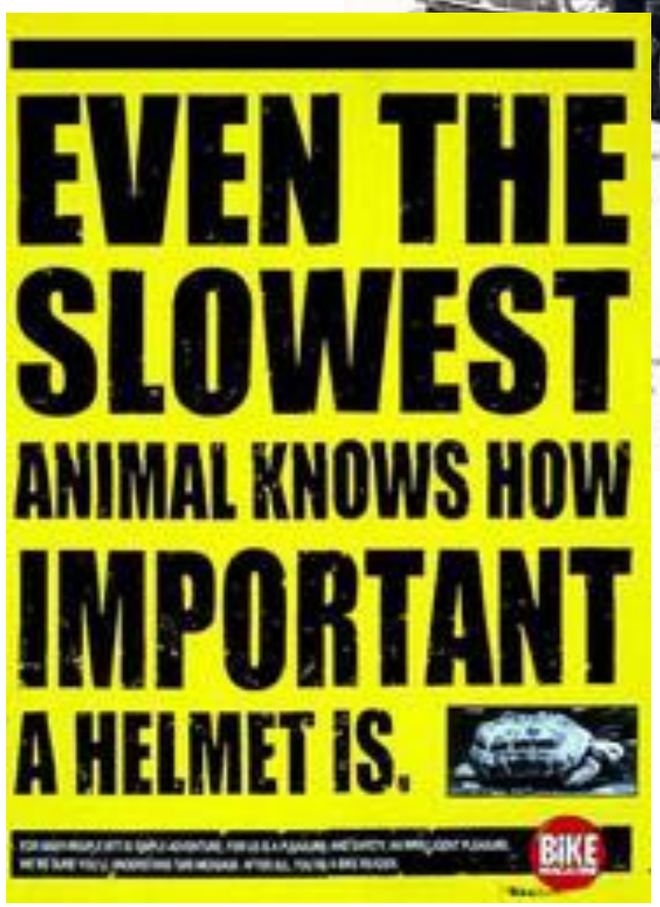
Illustrations

Complecated Compositions

Key observations

Texts on ads seem the most powerful clues





Our approach

1. Extract image2text and ocr2text relevance
2. Rank the statement given both scores

Statement text

I should shop for jeans at this store Because they use their profits to build homes for people

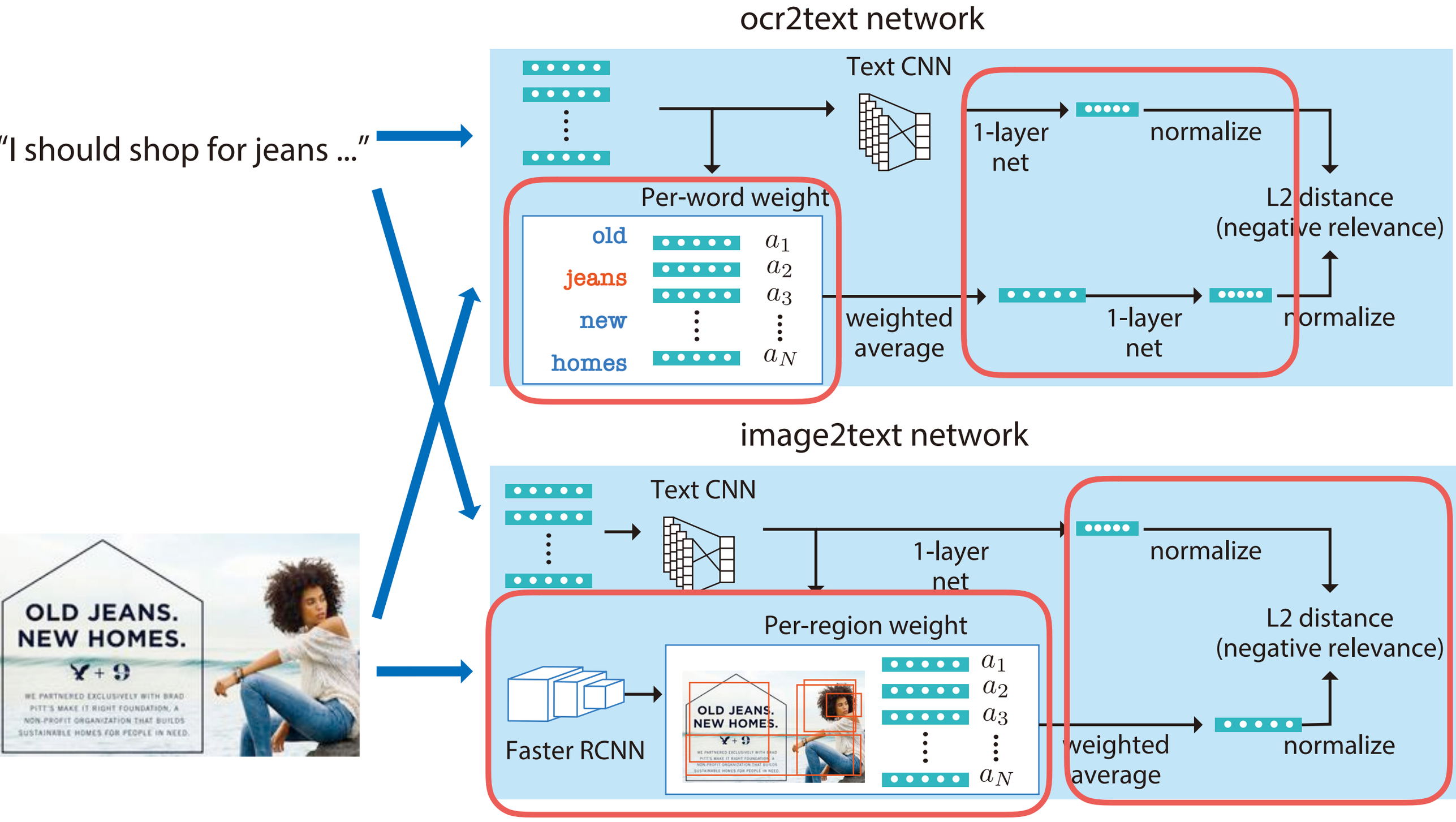
Visual and OCR clues from ad

old jeans new homes

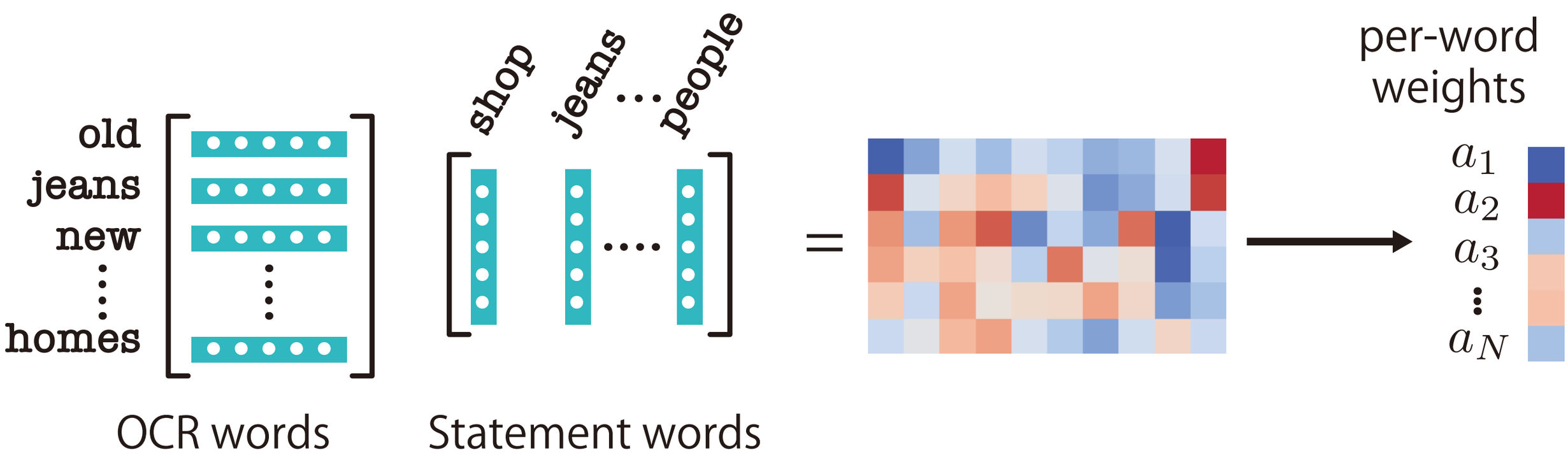
Image2text relevance

Ocr2text relevance

Our model


ocr2text network
"I should shop for jeans ..."
Text CNN
Per-word weight
old
jeans
new
homes
a_1
a_2
a_3
a_N
weighted average
1-layer net
normalize
L2 distance (negative relevance)
1-layer net
normalize
image2text network
Text CNN
1-layer net
normalize
L2 distance (negative relevance)
Per-region weight
Faster RCNN
OLD JEANS. NEW HOMES.
a_1
a_2
a_3
a_N
weighted average
normalize

Per-word weights



Vector similarity between OCR detected words and words in a statement text

Image embedding module

I should shop for jeans...

Faster RCNN

Attention network

1-layer network

1-layer network

Per-region weights

a_{r_1}
a_{r_2}
a_{r_3}
a_{r_M}

Weighted average

Text pre-processing

Long-term dependency harms. We trim texts.

I should shop for jeans at this store Because they use their profits to build homes for people

I should <action> because <reason>

shop for jeans at this store

they use their profits to build homes for people

Action to image / OCR relevance estimation

Reason to image / OCR relevance estimation

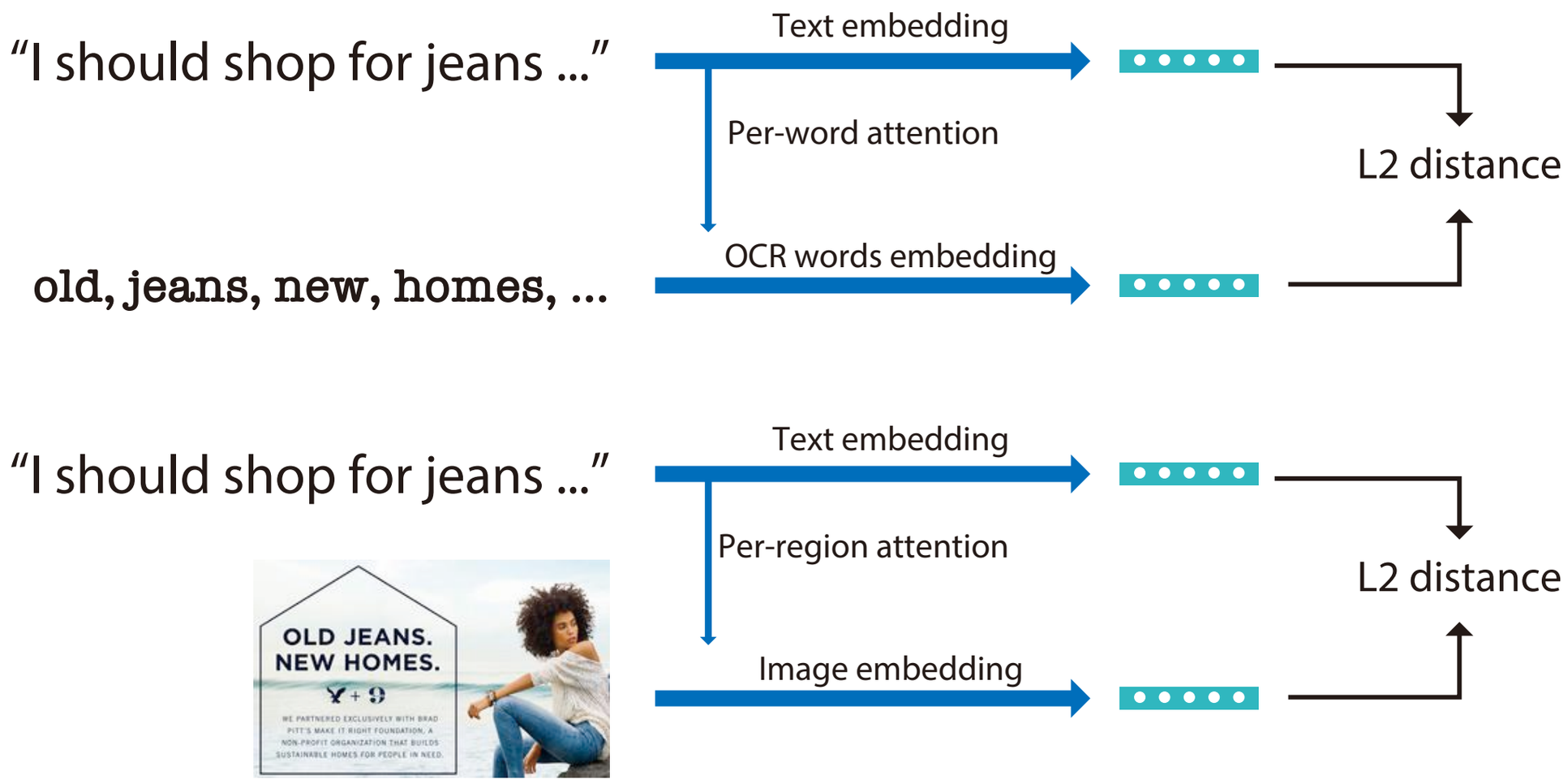
Training

Contrastive Loss

$$L_n = \frac{1}{2}\left(y_n d_n^2 + (1 - y_n)\max(\mu - d_n, 0)^2\right)$$

y_n : Label. 1=correct text, 0=incorrect text

d_n : Distance between text and image/OCR embeddings



Examples of QA results

A. I should shop for jeans at this store Because they use their profits to build homes for people

Per-word weights for action

old jeans new homes we partnered exclusively with brad make it right foundation non profit organization that builds sustainable homes for people in need

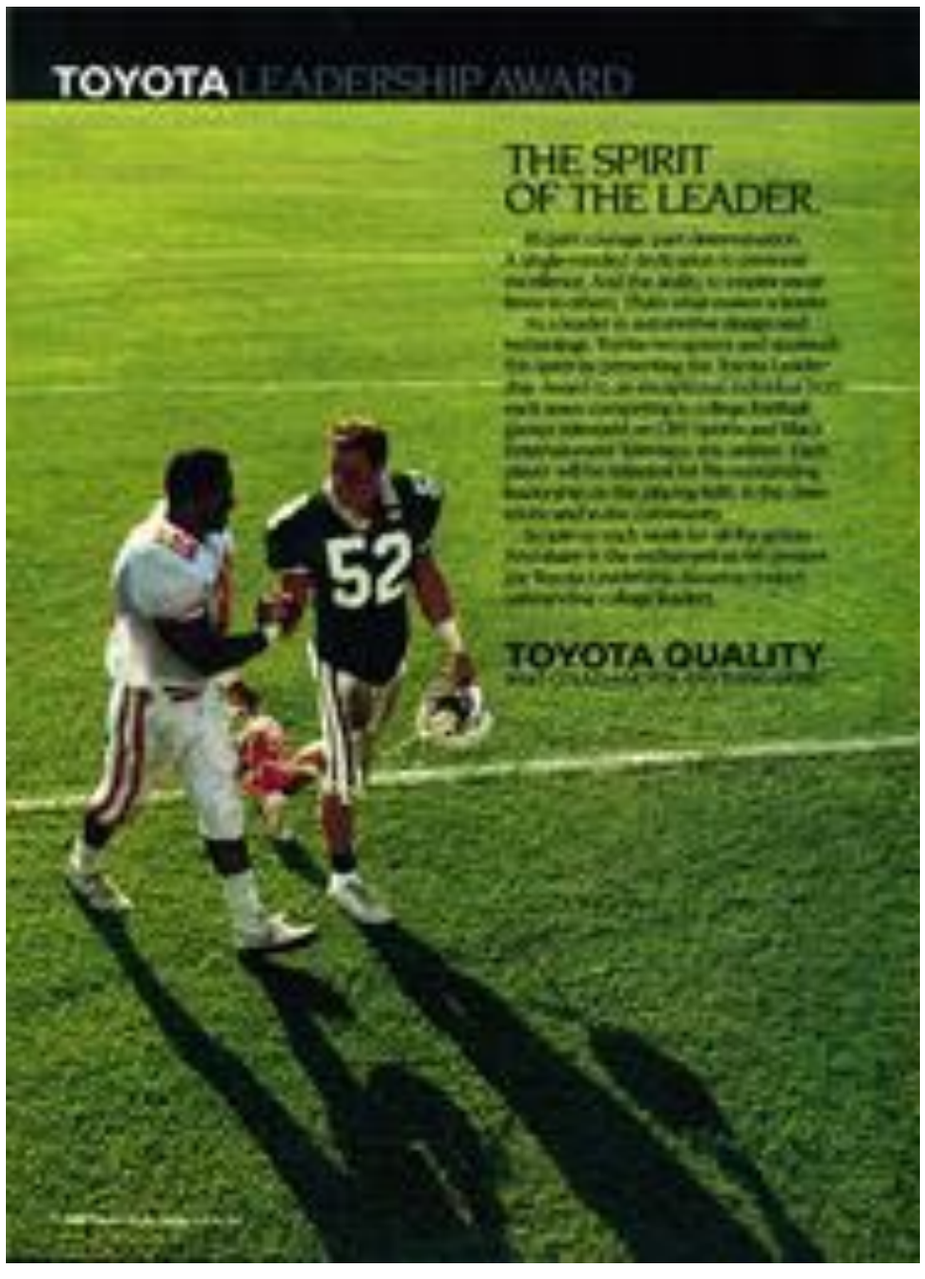
Per-word weights for reason

old jeans new homes we partnered exclusively with brad make it right foundation non profit organization that builds sustainable homes for people in need

Examples of QA results

Per-word weights for reason

toyota leadership award the spirit of the leader part courage part determination single minded dedication to personal excellence and the ability to inspire excel in others what makes leader as leader in automotive design and technology toyota recognizes and this spirit by presenting the toyota leader ship award to an each team competing in college football games …



A. I should buy Toyota Because they support college leaders

Examples of QA results

No words detected

A. I should shop at the gap Because it will make me sexy

How much does OCR help?

P@1

I should shop for jeans at this store Because they use their profits to build homes for people

0.42

I should shop for jeans at this store Because they use their profits to build homes for people

old jeans
new homes

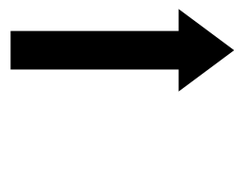
0.82

How much does appearance help?

P@1
(validation set)

I should shop for jeans at this store Because they use their profits to build homes for people

old jeans
new homes



0.83

I should shop for jeans at this store Because they use their profits to build homes for people

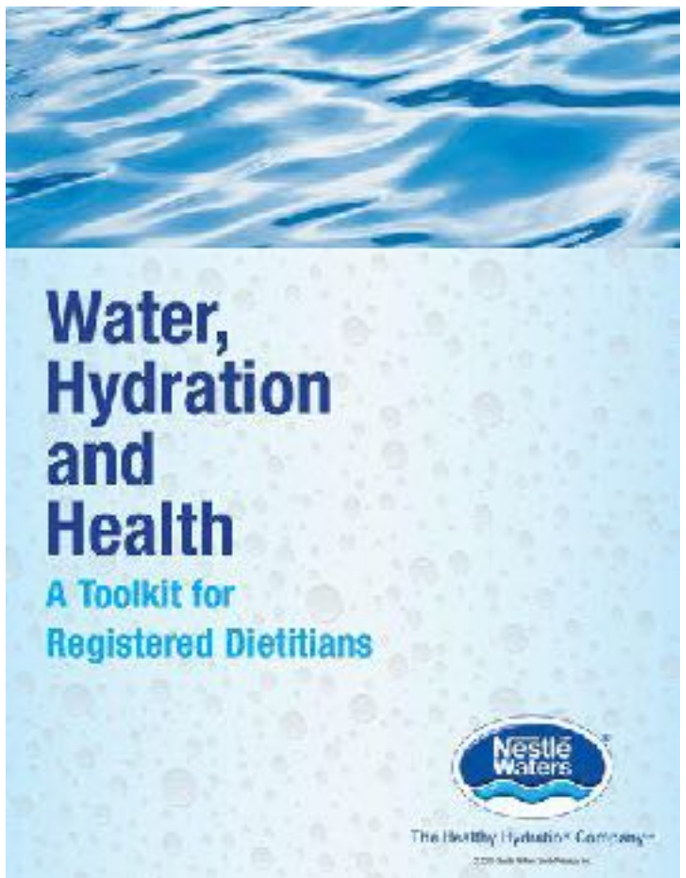
old jeans
new homes

0.85

Is OCR quality crucial?

A. Yes

	Score on validation set
Google Cloud Vision OCR	0.85
Tesseract OCR v3	0.54



Tesseract v3

```
, 1‘2. .4, f4 Water, HydraIiOn and”
```

Google Cloud Vision

```
Water,Hydration and Health
A Toolkit for Registered Dietitians
Nestle Waters
The Healthy Hydration Company
```

Things we didn't try

but look important

OCR results correction

B L AC K B E R R Y\nR E M E M B E R\n...

⬇

BLACKBERRY REMEMBER …

Multilingual support

Words detected but discarded

'誕生。美容オイル生恋軋のルージュ。\nそれは、美容オイルがスティックになった、贅沢な色艶。\nとろけて密着。色っぽくうるんだ官能の唇へ。\nマキアージュドラマティックルージュ全10色新発売\nMAKE uP DATA :pクアチイッタルー問" RD425 /トカルーアイPipp-V1233:殷定カラート\n2Mn合わ ▪0120.30-47100900-2100庫楽ギ始収定AMHEMO\nwww.shiseido.cojpmq\nレディにしあがれ。\nNEWMAQUİIAGE\nとの唇、女っぽくて、ごめんなさい。\n'

Limitations

- Multilingual support
- Decorative font
- Paintings, drawings

Misc findings

- Higher resolution helps
  - Perhaps because CNNs can literally read texts
  - Also object detection?

- Annotators really read texts
  - And perhaps logos

Symbolism in Ad

Symbolism might be used for rhetorics, but not necessarily for surface messages.

Remaining questions

- How well do humans understand ad messages without language clues?

- Effects of designs (colors, layout, font style, etc.) in message telling

- Better task / dataset design?
  - Hiding / blurring texts enough?