Textual Visual Semantic Dataset for Text Spotting

Ahmed Sabir[1], Francesc Moreno-Noguer[2], Lluís Padró[1]

[1] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain

[2] Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain

CVPR-Workshop on Text and Documents
in the Deep Learning Era

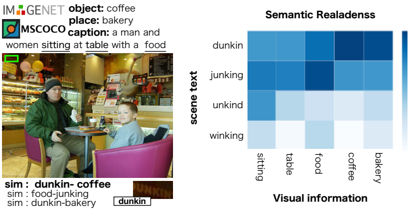
Summary

PROBLEM

- Improving the performance of pre-trained text spotting systems in a complex background with semantic information.

CONTRIBUTION

- Introducing extra prior knowledge (task and dataset) to text spotting or OCR in the wild: by reranking the candidates based on their semantic relatedness with words describing the image context.



Proposal task: Integrating Visual Semantic Information

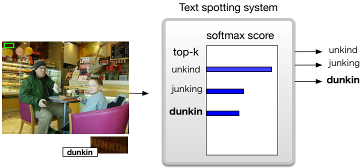
Case: complex background

Example: case of cut-off bounding box.

The correct candidate word is inside the baseline softmax $k = 3$

✓ The objective is to re-rank the correct candidate word.

- The baseline (CNN-90k dict) is trained on oxford synthetic dataset with 90k dict [Jaderberg et al., 2016]



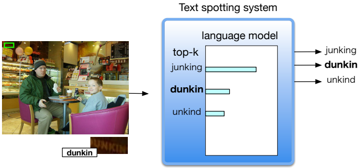
Baseline

Proposal task: Integrating Visual Semantic Information

Case: complex background

Example: case of cut-off bounding box.
The correct candidate word is inside the baseline softmax $k = 3$
✓ The objective is to re-rank the correct candidate word.

- The baseline (CNN-90k dict) is trained on oxford synthetic dataset with 90k dict [Jaderberg et al., 2016]
- The simplest approach is to add Language model ☹



Baseline+Language model

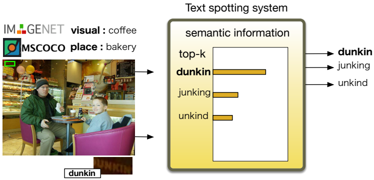
Proposal task: Integrating Visual Semantic Information

Case: complex background

Example: case of cut-off bounding box.
The correct candidate word is inside the baseline softmax $k = 3$
✓ The objective is to re-rank the correct candidate word.

- The baseline (CNN-90k dict) is trained on oxford synthetic dataset with 90k dict [Jaderberg et al., 2016]

- Our approach is to add visual semantic information (word relation) ☺

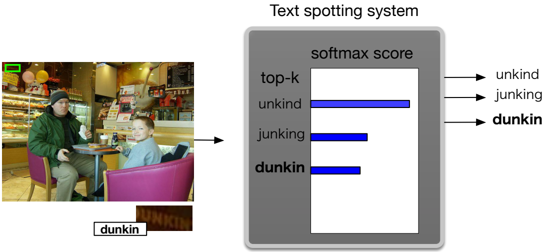


Baseline+Visual Semantic

How to integrate the semantic information?

step 1

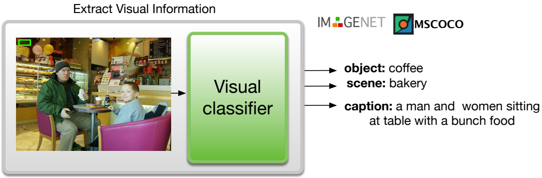
1. We extract the top-k with associate probability from the baseline.



How to integrate the semantic information?

Step 2

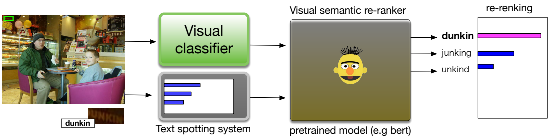
1. We extract the top-k with associate probability from the baseline.
2. We employ visual classifier (i.e object, scene and caption) to extract the visual context from the image.



How to integrate the semantic information?

Step 3

1. We extract the top-k with associate probability from the baseline.
2. We employ visual classifier (i.e object, scene and caption) to extract the visual context from the image.
3. We compute the semantic similarity between the word and its visual context and then re-rank them.



Proposed: Textual Visual Dataset

Dataset generation

Text hypothesis

- We employ several off the self pre-trained Text Spotting baselines to generate *k* text hypotheses (i.e. CNN-90K, CRNN,..)

Visual context

- Object classifier (Resent152, Inception-ResNet-v2) 1000 label classes
- Scene classifier (365 Resent scene classifier) 365 label classes
- Caption description (standard model) tuned on COCO-caption

Text hypothesis	Object	Scene	Caption
11 , il, j, m, ...	railroad	train	a train is on a train track with a train on it
lossing, docile, dow, **dell**, ...	bookshop	bookstore	a woman sitting at a table with a laptop
29th, 2th, **2011**, zit, ...	parking	shopping	a man is holding a cell phone while standing
happy, hooping, happily, nappy, ...	childs	bib	a cake with a bunch of different types of scissors
coke, gulp, slurp, fluky,...	plate	pizzeria	a table with a pizza and a fork on it
will, **wii**, xviii, wit,....	remote	room	a close up of a remote control on a table

Proposed: Textual Visual Dataset for COCO-text

Resulting dataset

- ICDAR-17-V: Image + Textual dataset from IC17 Task 3
- COCO-text-Visual: Image + Textual dataset from COCO-text
- COCO-Pairs: Only Textual dataset from COCO-text

Unique Count for Textual dataset								
Dataset	image #	bbox	caption	object	words	nouns	verb	adjectives
Conceptual [35]	3M	-	3M	-	34219,055	10254,864	1043,385	3263,654
MSCOCO [22]	82k	-	413k	-	3732,339	3401,489	250,761	424977
Flickr 30K [44]	30k	-	160k	-	2604,646	509,459	139128	169158
SVT [42]	350	✓	-	-	10,437	3856	46	666
COCO-Text [40]	66k	✓	-	-	177,547	134,970	770	11,393
Visual context dataset (proposed dataset)								
COCO-Text-V	16k	✓	60k	120k	697,335	246,013	35,807	40,922
IC17-V	10k	✓	25k	50k	296,500	96,371	15,820	15,023
COCO-Pairs	66k	-	-	158k	319,178	188,295	6,878	46,983

Result

Evaluation

- We evaluate our approach on different baselines: 1) CNN-90k dictionary [Jaderberg et al., 2016] 2) LSTM with visual attention [Ghosh et al., 2017]. The table shows the best results after re-ranking using different re-ranker.

Model	**CNN**			**LSTM**		
	Acc.	*k*	MRR	*Acc.*	*k*	MRR
Baseline (BL)	**Acc.:19.7**			**Acc.:17.9**		
Experiment 1 word-to-word relation (i.e. object and scene)						
BL+Word2vec [25]	21.8	5	44.3	19.5	4	80.4
BL+Glove [27]	22.0	7	44.5	19.1	4	78.8
BL+Sw2v [24]	21.8	7	44.3	19.4	4	80.1
BL+Fasttext [17]	21.9	7	44.6	19.4	4	80.3
BL+TWE [34]	22.2	7	44.7	19.5	4	80.2
BL+RWE [3]	21.9	7	44.5	19.6	4	80.7
BL+LSTMmebed [13]	21.6	7	44.0	19.2	4	79.6
Experiment 2 word-to-sentence relation (i.e. caption)						
BL+USE-T [5]	22.0	6	44.7	19.2	4	79.5
BL+BERT-feature [6]	21.7	7	45.0	19.3	4	**81.2**
BL+BERT (fine-tune) [6]	**22.7**	8	**45.9**	**20.1**	9	79.1

Examples

- Re-ranking the correct candidate word and its visual context with tuned BERT [Devlin et al. 2019] on the proposed dataset.
- BERT re-ranked the candidates based on the image description.

object: life*boat* **scene:** raft
caption: a *boat* is parked in small *boat*
text hypothesis: honor, donor, **honda**,..
bounding box: 

honda

top-w_k: *sim*(honda, parked)

object: pencil **scene:** *child*
caption: a small *child*'s toy is sitting on a table
text hypothesis: adding, adana, **adam**,..
bounding box: 

adam

top-w_k: *sim*(adam, toy)

object: *street* **scene:** downtown
caption: a *street* sign with a sign on on the side
text hypothesis: nay, **way**, may,..
bounding box: 

way

top-w_k: *sim*(way, street)

object: american **scene:** hospital
caption: a table with bunch of food on it
text hypothesis: il,xl,**7**,..
bounding box: 
7

top-w_k: *sim*(7, table), ULM(7)

Conclusion

Contributions

- We defined the task of post-processing for text spotting by exploring the semantic relation between text and scene in a textual manner. Also, introducing a visual context dataset for this problem.

Final thoughts

- Text in images is **not always related** to its visual environment, there is only a fraction of cases this approach may help solving, but given its low cost, it may be useful for domain adaptation of general text spotting systems (e.g fixing false-positive and short word)