Learning Visually-Grounded Semantics from Contrastive Adversarial Samples

Haoyue Shi1, Jiayuan Mao2, Tete Xiao4, Yuning Jiang3 and Jian Sun3 1: Peking University 2: Tsinghua University 3: Megvii, Inc
(hyshi, jasonhsiaoj97)@pku.edu.cn, mjyl14@mails.tsinghua.edu.cn, (jyn, sunjian)@megvii.com

INTRODUCTION
Visual-Semantic Embeddings (VSE)
• Use parallelized image-caption pairs and embed texts and images into a joint space.
• Several datasets have been created for such purpose.
• However, even MS-COCO[1] is too small compared with the compositional semantic space.

VSE with Contrastive Adversarial Samples (this work)
• Show the limitation of existing datasets and frameworks through adversarial attacks.
• Close the gap with semantics-aware text augmentation.
• Evaluate the visual grounding on multiple tasks.

A SIMPLE YET EFFECTIVE APPROACH
Add the Contrastive’ Adversarial Samples to the Training Set
*: Use the online hard example mining (OHEM) technique to find “Contrastive” ones.

\[
\min \ell_{VSE}(i, c) = \sum_{c'} [\alpha + s(i, c') - s(i, c)]_+ + \sum_{c' \neq c} [\alpha + s(i', c') - s(i, c)]_+ \\
\min \ell^{VSE*}(i, c) = \max_{c \neq c'} [\alpha + s(i, c') - s(i, c)] + \max_{c' \neq c} [\alpha + s(i', c') - s(i, c)] \\
\min \ell^{VSE-C}(i, c) = \ell^{VSE*}(i, c) + \max_{c' \in C \setminus c} [\alpha + s(i, c') - s(i, c)] \\
\]

\{i, image, c, caption, C: adversarial samples.\}

GROUNDING TEST I: WORD-OBJECT CORRELATION
Task Description
Image
A table with a huge glass vase and fake flowers come out of it.
A plant in a vase sits at the end of a table.
A vase with flowers in it with long stems sitting on a table with candles.
A large centerpiece that is sitting on the edge of a dining table.
Flowers in a clear vase sitting on a table.

Captions
A table with a huge glass vase and fake flowers come out of it.
A plant in a vase sits at the end of a table.
A vase with flowers in it with long stems sitting on a table with candles.
A large centerpiece that is sitting on the edge of a dining table.
Flowers in a clear vase sitting on a table.

POSITIVE OBJECTS: table, plant, vase.
NEGATIVE OBJECTS: screen, pickle, sandwich, toy, hill, coat, cat, etc.

BEGIN WITH ADVERSARIAL ATTACKS

Semantics-aware Text Augmentation (Adversarial Samples)
• Noun: use Word-Net [4] to compare the word similarity (e.g., Synonyms, Hypernyms).
• Numeral/Indefinite Article: singularize or pluralize corresponding nouns when necessary.
• Relation: dependency-parsing based subject and object detection.

SALIENCY VISUALIZATION
Which part in the image or caption, in particular, makes them semantically different? We compute the Jacobian (we normalize the textual saliency for visualization):
\[
J = \nabla_s a(s, c') = \nabla_c W^f_{ji}(s, c) \cdot W^g_{gj}(c', \theta) \\
\]

GROUNDING TEST II: FILL-IN-THE-BLANK

REFERENCES

ACKNOWLEDGEMENTS
This work was done when HS, JM and TX were intern researchers at Megvii Inc. HS, JM and TX contribute equally to this paper.