Learning Visually-Grounded Semantics from Contrastive Adversarial Samples
Haoyue Shi\textsuperscript{1}, Jiayuan Mao\textsuperscript{2}, Tete Xiao\textsuperscript{1}, Yuning Jiang\textsuperscript{3} and Jian Sun\textsuperscript{3} \textsuperscript{1}: Peking University \textsuperscript{2}: Tsinghua University \textsuperscript{3}: Megvii, Inc (hys@pku.edu.cn, majy14@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\cite{2} 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.

\begin{align*}
\text{VSE [2]}: & \quad \min \ell_{\text{VSE}}(i,c) = \sum_{c'}[\alpha + s(i,c') - s(i,c)]_+ + \sum_{c'}[\alpha + s(i',c') - s(i',c)]_+ \\
\text{VSE++ [3]}: & \quad \min \ell_{\text{VSE}^*}(i,c) = \max_{c' \in C(i)}[\alpha + s(i,c') - s(i,c)] + \max_{c' \in C(i')}[\alpha + s(i',c') - s(i',c)] \\
\text{VSE-C (ours)}: & \quad \min \ell_{\text{VSE}^C}(i,c) = \ell_{\text{VSE}^*}(i,c) + \max_{c' \in C(i)}[\alpha + s(i,c') - s(i,c)]
\end{align*}

\(i, c\): image, caption, \(C\): adversarial samples.

GROUNDING TEST I: WORD-OBJECT CORRELATION
Task Description:
- 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 \cite{4} to compare the word similarity (e.g., Synonyms, Hypernyms).
- Numerical/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):

\[
\begin{align*}
J &= \nabla_i s(i,c') = \nabla_i W_i^T f(i; \theta_k) \\
&\quad \cdot W_i^T g(c'; \theta)
\end{align*}
\]

GROUNDING TEST II: FILL-IN-THE-BLANK

REFERENCES