MODULARIZED TEXTUAL GROUNDING FOR COUNTERFACTUAL RESILIENCE

{ ZHIYUAN FANG\textsuperscript{1}, SHU KONG\textsuperscript{2}, CHARLESS FOWLKES\textsuperscript{2}, YEZHOU YANG\textsuperscript{1} }

\textsuperscript{1}ARIZONA STATE UNIVERSITY, \textsuperscript{2}UNIVERSITY OF CALIFORNIA, IRVINE

---

**OVERVIEW**

Text: I am looking for a boy in blue.

- Counter-factual Phrase Grounding: woman in brown

**INTRODUCTION**

We propose a visual grounding system which is:

- End-to-end trainable in a weakly supervised fashion with only image-level annotations.
- Counterfactually resilient owing to the modular design.

**TRAINING**

We stack word2vec of \( C \) attributes as a dictionary \( D = \{d_1, \ldots, d_C\} \), and have scoring function for \( i^{th} \) attribute:

\[
y_i = \frac{2}{1 + \exp\left(\|\psi(D_i) - \phi(x)\|^2\right)}
\]

, where \( \psi \) and \( \phi \) are transformation layers. We apply the above scoring function at pixel level training.

**RESULTS**

![Entity Grounding](image1)

- **Entity Grounding**: Image
  - Person
  - Train
  - Cat

- **Attribute/Color Grounding**: Semantic-Attribute, Color
  - Person
  - White+Blue
  - Blue+Green
  - Purple

- **Textual Grounding**: Entity, Attribute, Color Grounding
  - Many sorry men in a line holding grass in green sets walking down a concrete road with a big building in the background.
  - A young girl in red leotard pants is holding a very small baby on her lap.
  - Girl in yellow coat.

- **Textual Grounding in Videos**: Girl in yellow coat.
  - A man with white shirt.
  - Lady with red dress.

**ARCHITECTURE**

![Architecture Diagram]

- **Entity Module**: Attention Branch, “Person”
- **Semantic Attribute Module**: “older man”
- **Color Module**: “blue”
- **Attention Map**: Bounding Boxes Selection

**COLOR GROUNDING MODULE**

- **Input**
- **Pixel Level Classification**
- **Res Backbone**
- **Feature Maps**
- **Pixel Embedding**

**REFERENCES**


**CONTACT AUTHOR**

![QR Code]