Reasoning about pragmatics with neural listeners and speakers

Jacob Andreas and Dan Klein
UC Berkeley

Presentation:
Xingyi Zhou

Goal: Reference Game

- Input: A target image and a distractor image
- Output: A sentence that distinguish target image from distractor image
- Evaluation: Human evaluation on AMT

(a) target (b) distractor

the owl is sitting in the tree

(c) description

the owl is wearing a hat

the owl is sitting in the tree

Reference Game Formulation

Defined on a speaker **S** and a Listener **L**

1. Reference candidates r1 and r2 are revealed to both players.
2. S is secretly assigned a random target $t \in \{1, 2\}$.
3. S produces a description d = S(t, r1, r2), which is shown to L.
4. L chooses c = L(d,r1,r2).
5. Both players win if c = t.

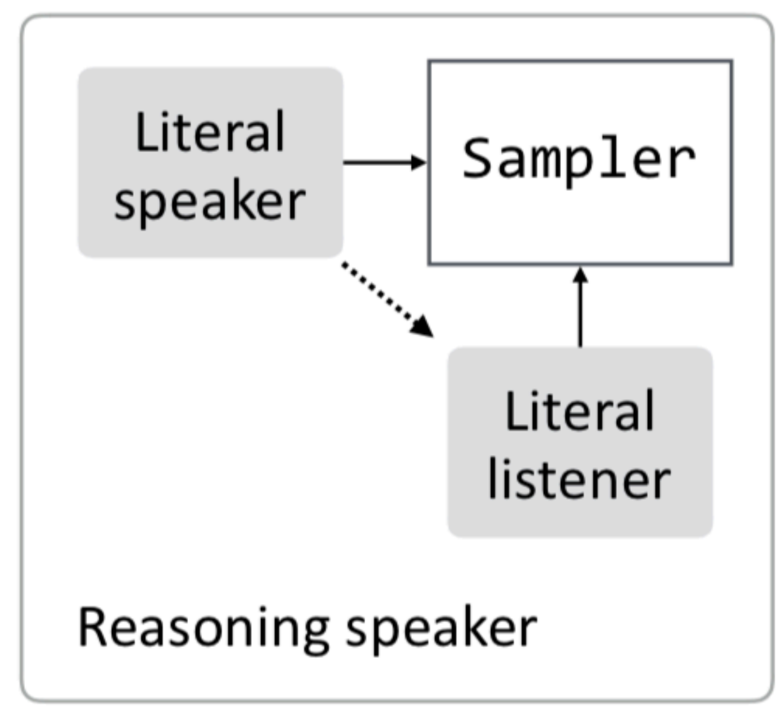
Previous Methods

- Direct approach (supervised learning)
  - Imitate human play without listener representation.
  - No domain knowledge needed.
  - Require a large training samples, which are scarce.
- Derived approach (optimizing by synthesis)
  - Initialize a listener model and then maximize the accuracy of this listener.
  - pragmatic free.
  - Require hand-engineering (on grammar) listener model.

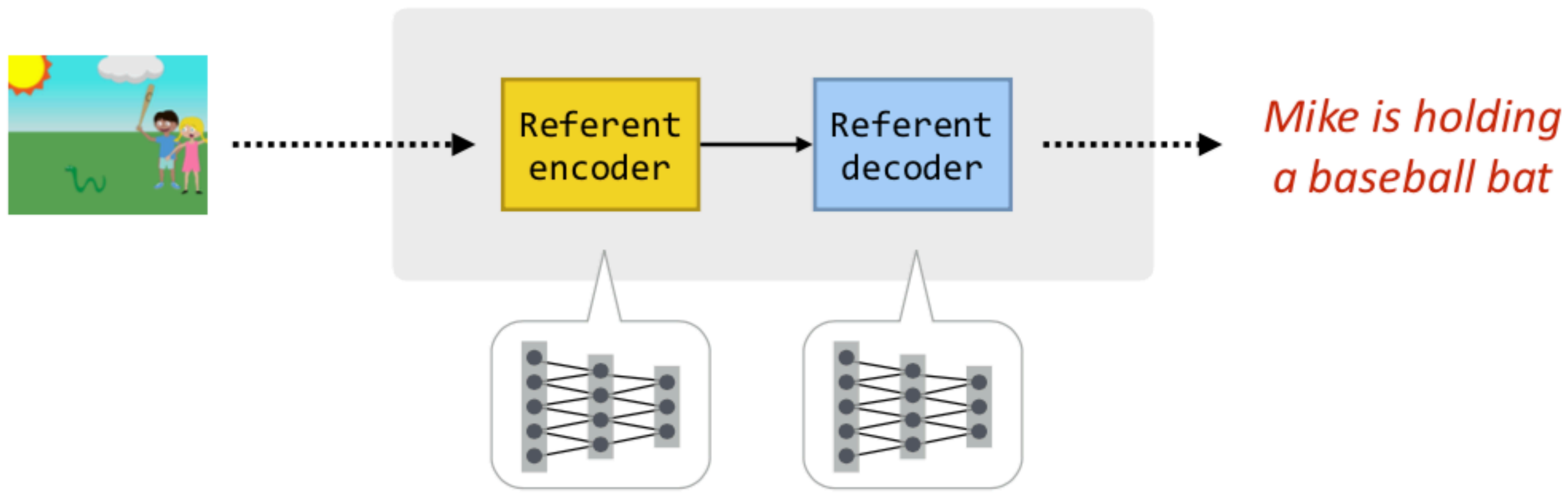
pragmatic: concerned with practical matters / it must be informative, fluent, concise, and must ultimately encode an understanding of L's behavior

Overview of the Proposed approach

- Combine the benefits of both direct and derived models.
    - Use direct model to initialize a Literal listener and a Literal speaker without domain knowledge
    - Embed the initialization with a higher-order model that reason about listener responses



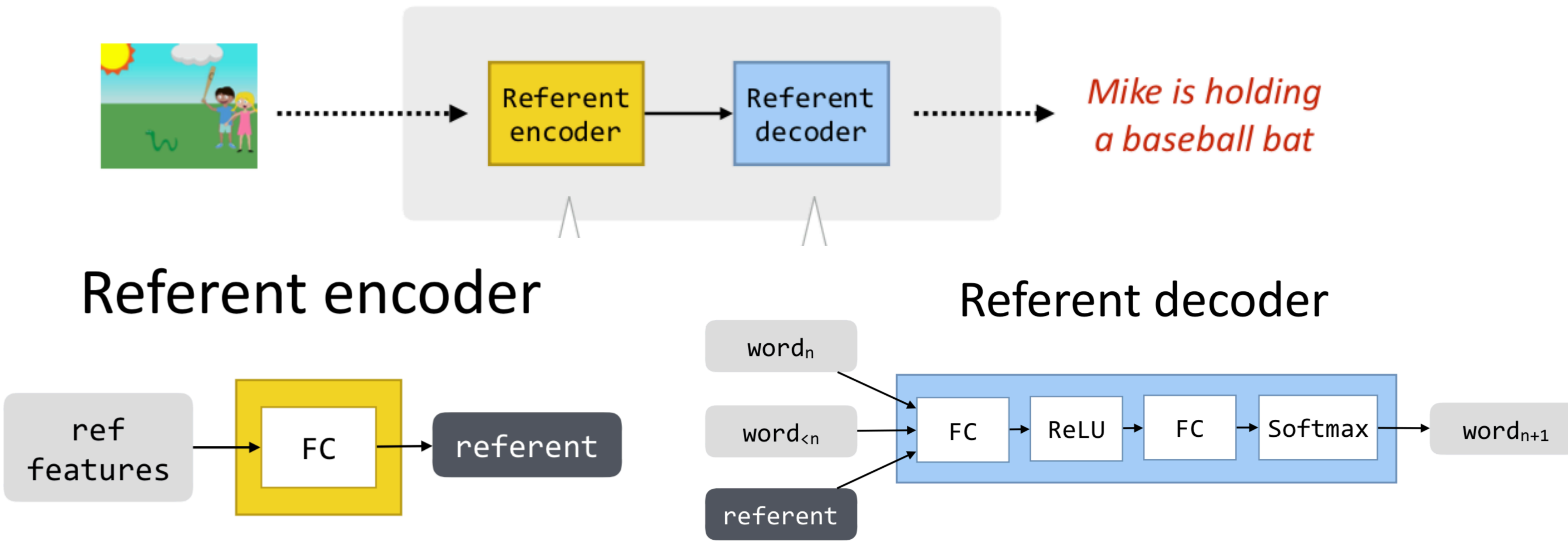
Initialize the Literal Speaker(S0)



- Only have non-contrastive captions for training
- Image features: indicator features provided by the dataset, not CNN features but easy to replace
- Use a decoder to recursively generate a sentence (similar to RNN)
- The literal Speaker itself is sufficient for referring game.

Slides credit:
Andreas and Klein

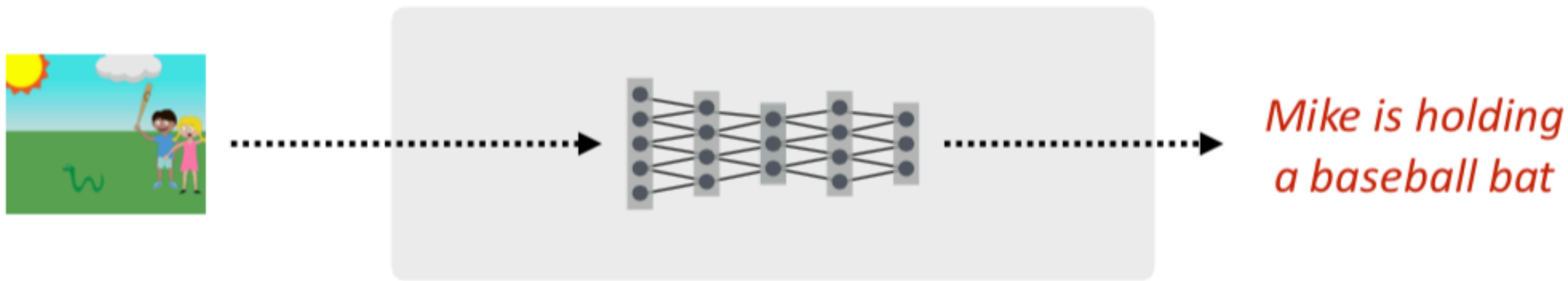
Initialize the Literal Speaker(S0)



Slides credit: Andreas and Klein

Initialize the Literal Speaker(S0)

Training

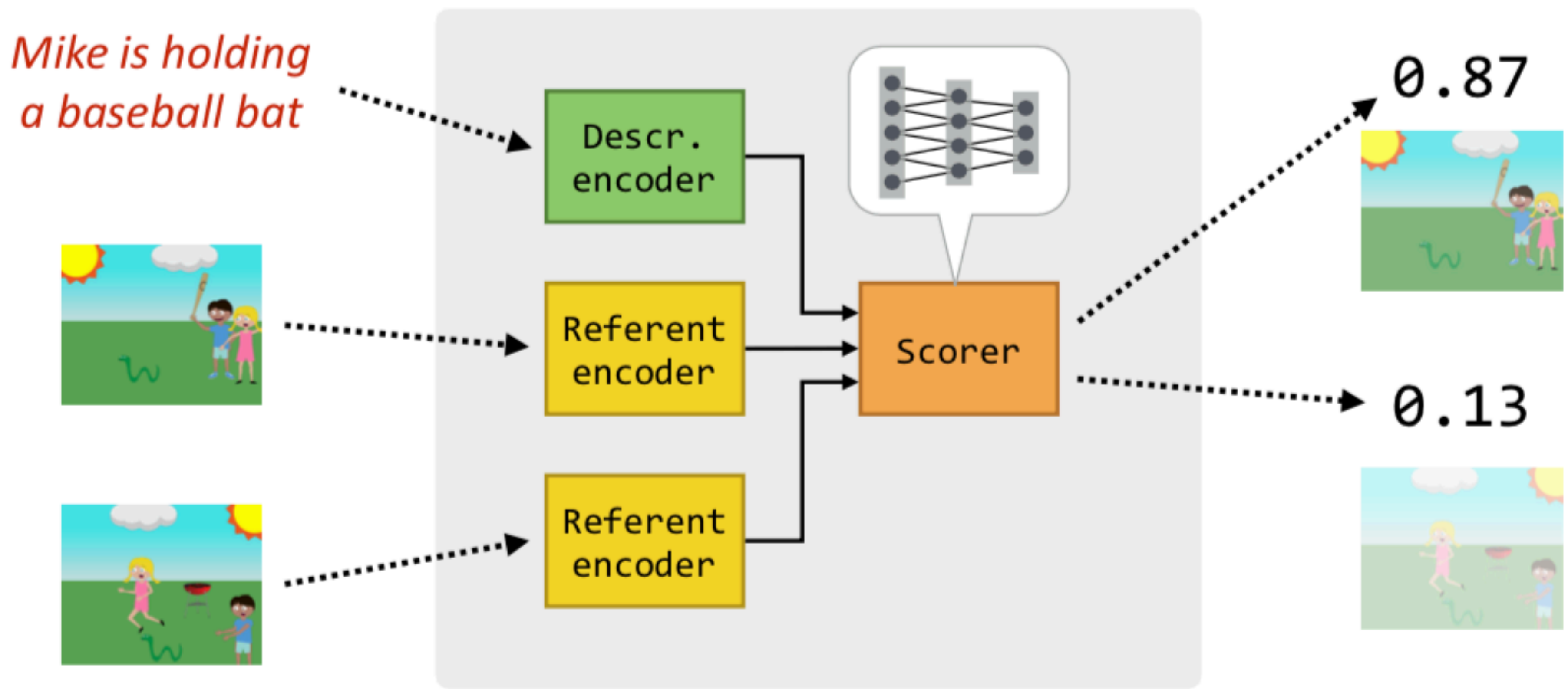


Testing

Slides credit:
Andreas and Klein

Produce the sentence and its confidence score during testing

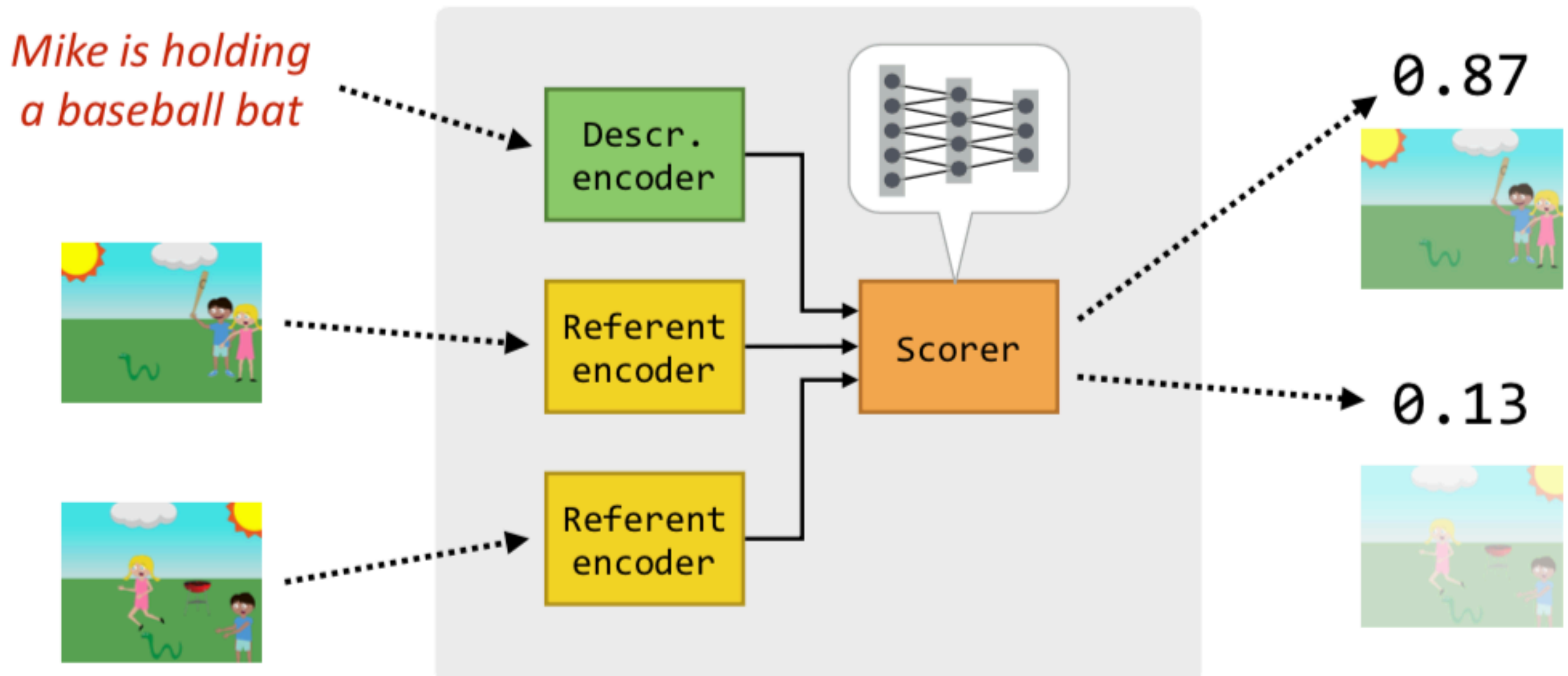
Initialize the Literal Listener(L0)



- Random sample distractor image as negative sample.
- Take n-gram feature as sentence representation.

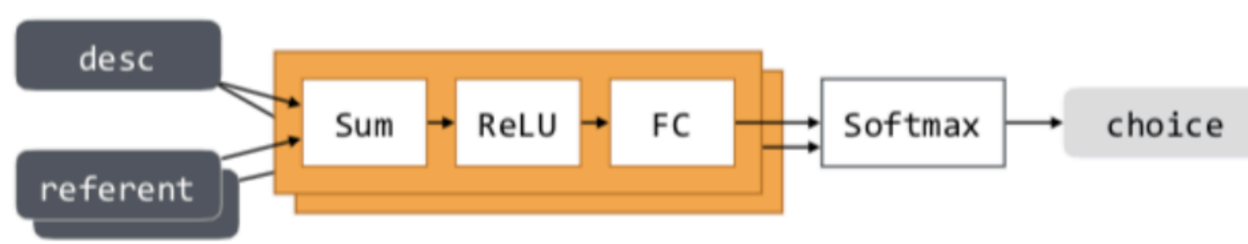
Slides credit:
Andreas and Klein

Initialize the Literal Listener(L0)



Referent encoder

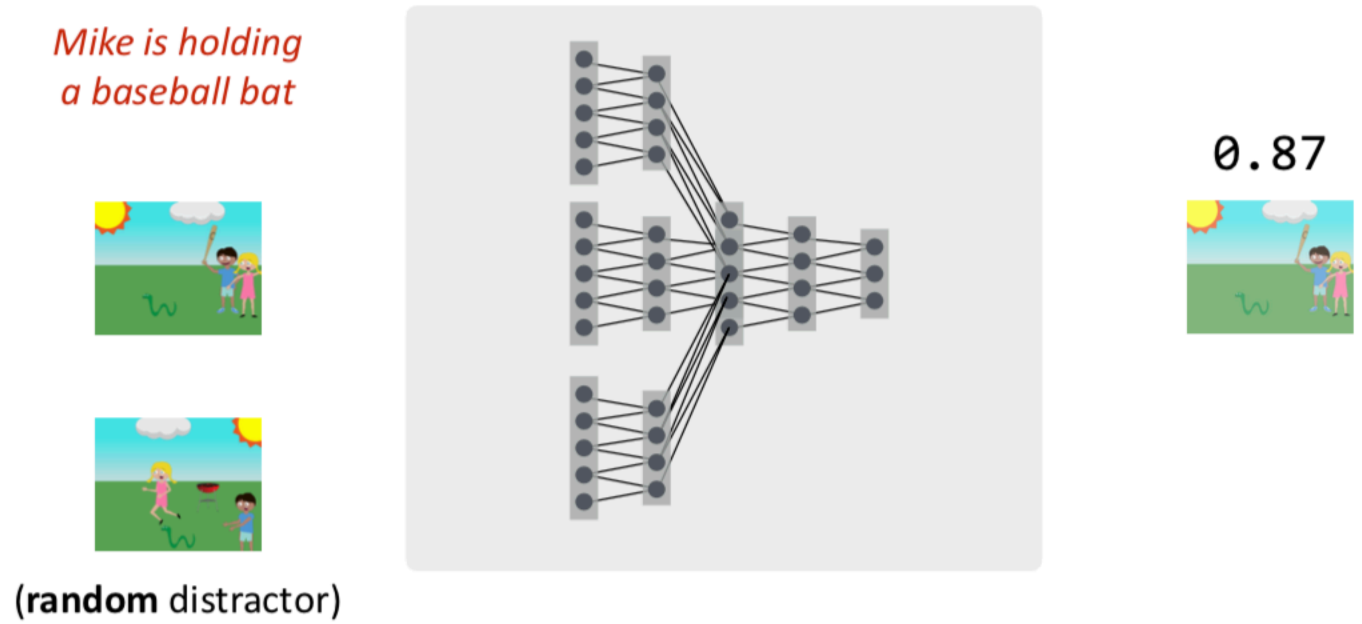
Referent decoder



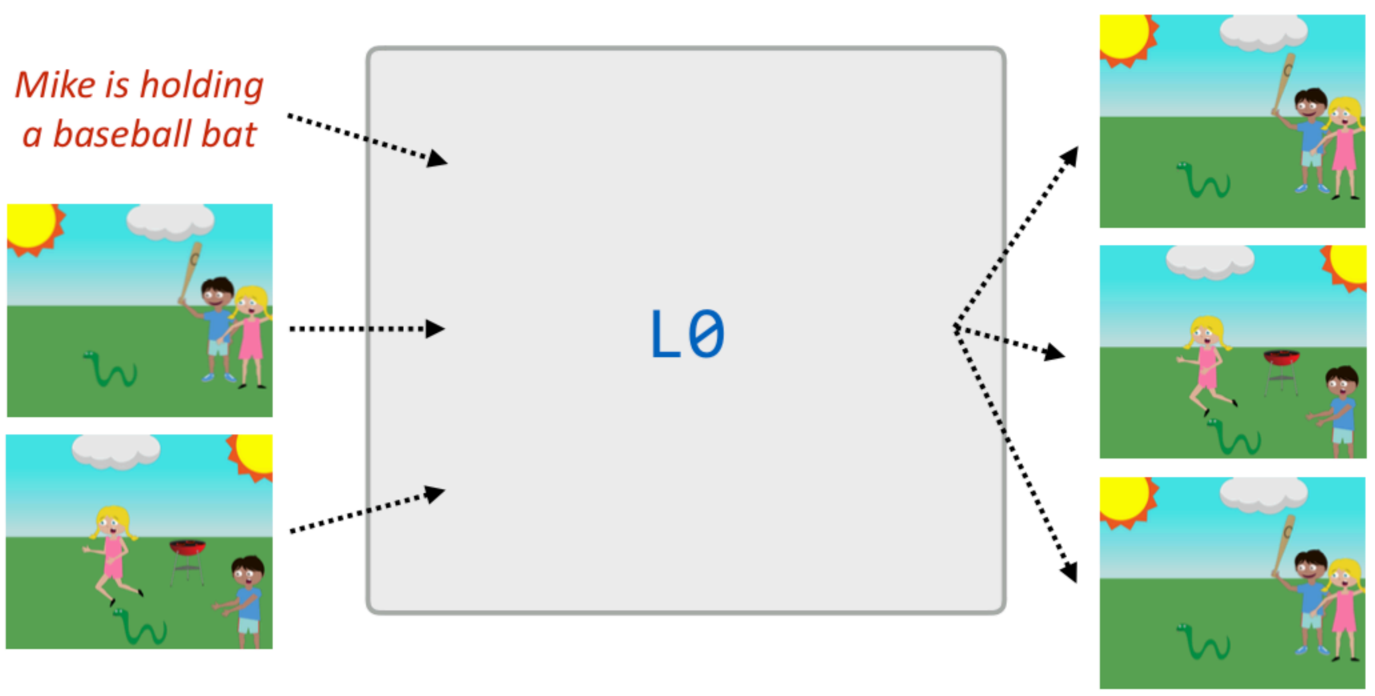
Slides credit:
Andreas and Klein

Initialize the Literal Listener(L0)

Training

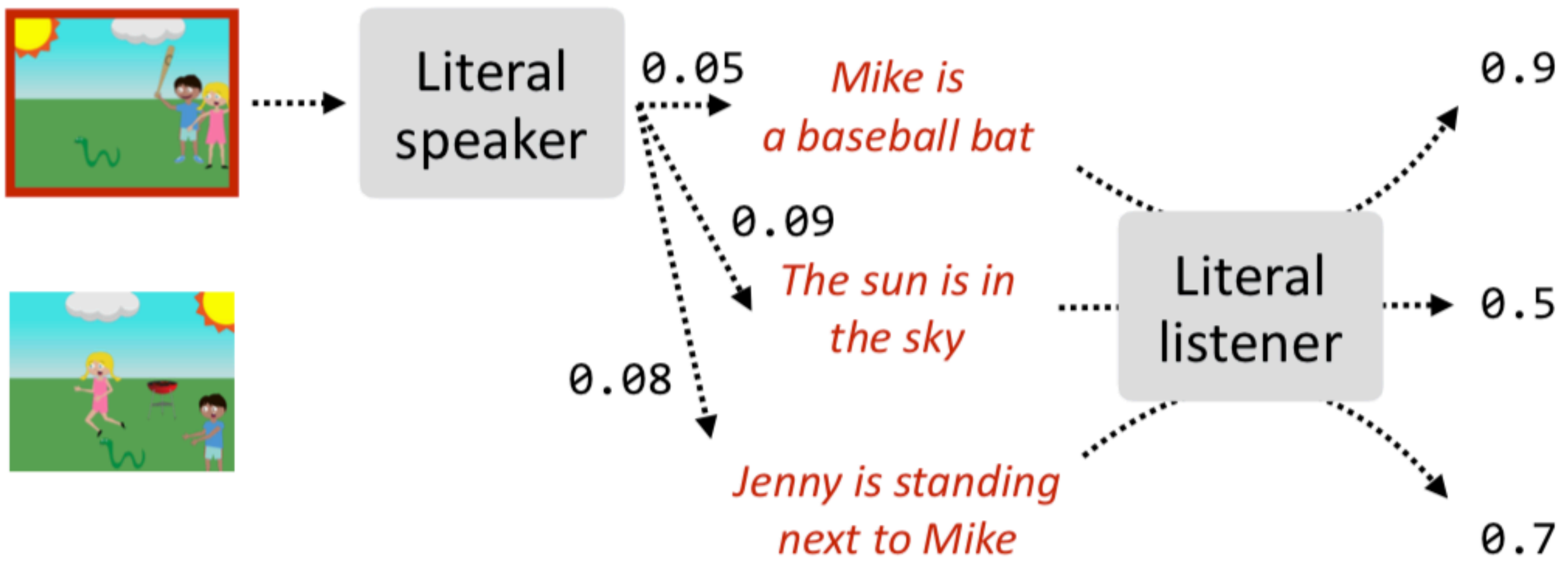


Testing



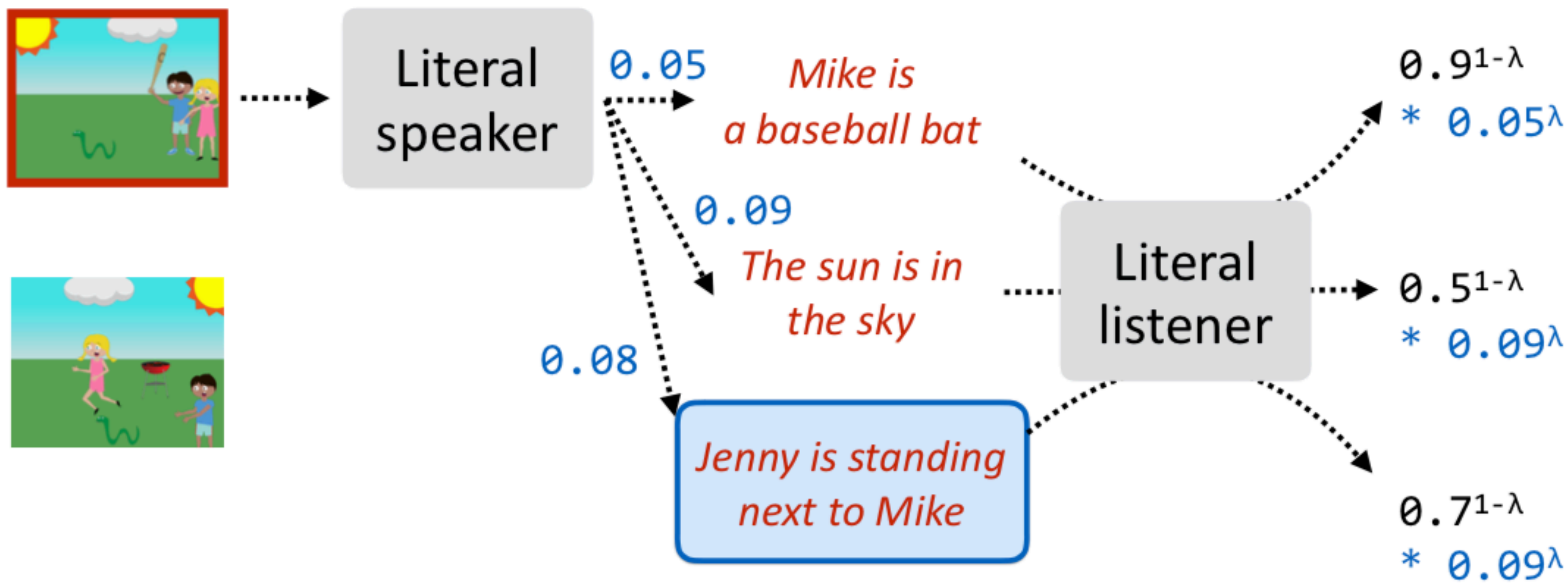
Slides credit:
Andreas and Klein

Reasoning speaker(S1)



Slides credit: Andreas and Klein

Reasoning speaker(S1)



λ :Trade of between L0 and S0

$$p_k = p_{\mathrm{S0}}(d_k|r_i)^{\lambda} \cdot p_{\mathrm{L0}}(i|d_k, r_1, r_2)^{1-\lambda}$$

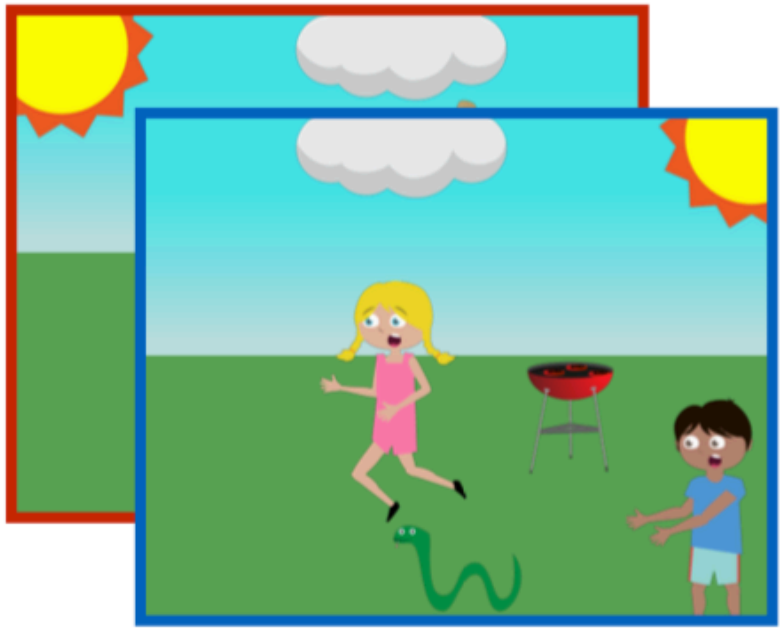
Slides credit:
Andreas and Klein

Reasoning speaker(S1)

$$p_k = p_{\mathrm{S0}}(d_k|r_i)^{\lambda} \cdot p_{\mathrm{L0}}(i|d_k, r_1, r_2)^{1-\lambda}$$

- S0: Ensure that the description conforms with patterns of human language use and align with the image.
- L0: Ensure that the description contains enough information and take account of the contrastive image.

Experiments - Dataset



Abstract Scenes Dataset

1000 scenes

10k sentences

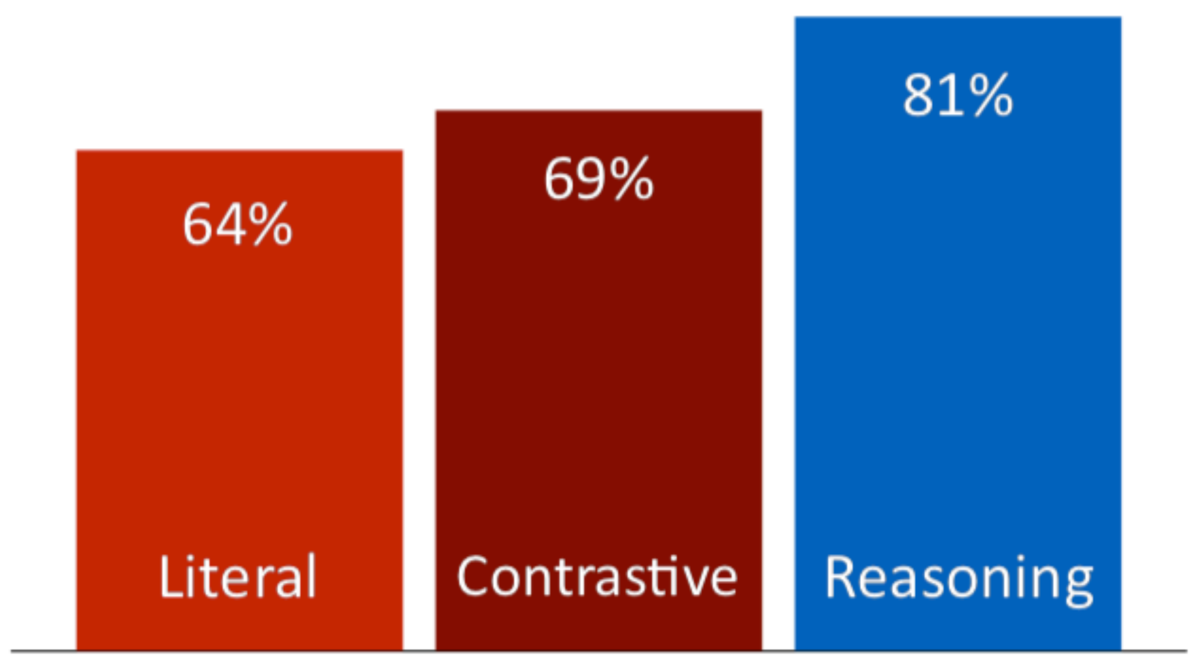
Feature representations

Evaluation: Human evaluation on AMT

Slides credit:
Andreas and Klein

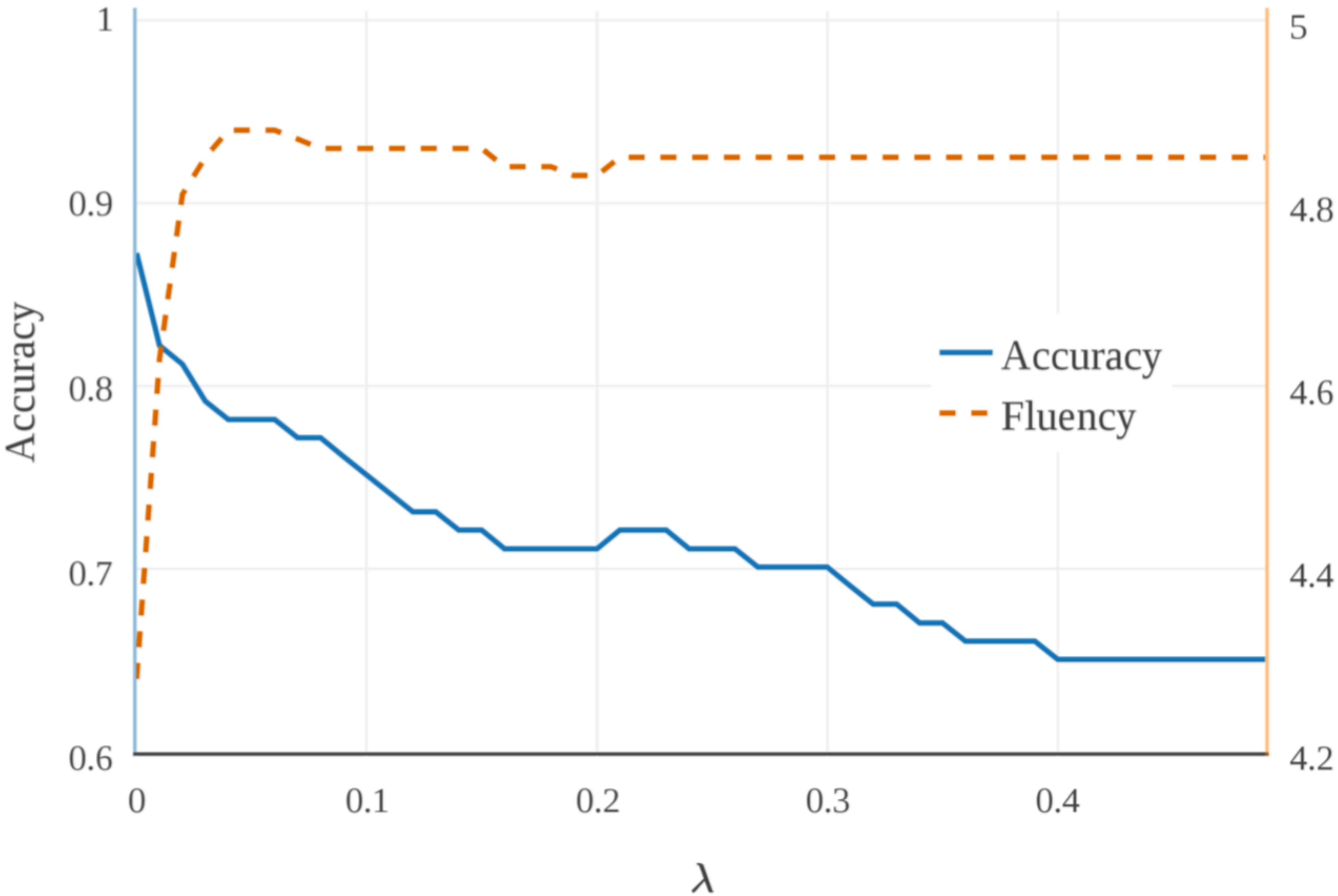
Experiments - Baselines & Results

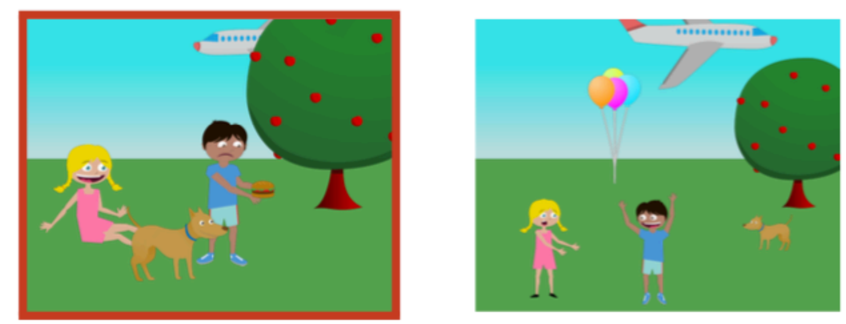
- Literal: the S0 model by itself
- Contrastive: a conditional LM trained on both the target image and a random distractor [Mao et al. 2015]



Slides credit: Andreas and Klein

Tradeoff between speaker and listener models



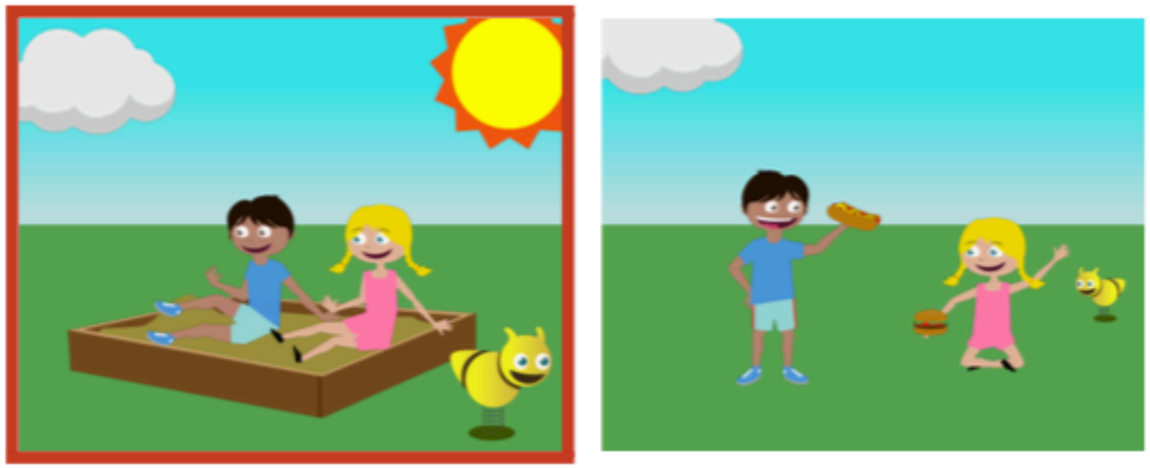


(a) target (b) distractor

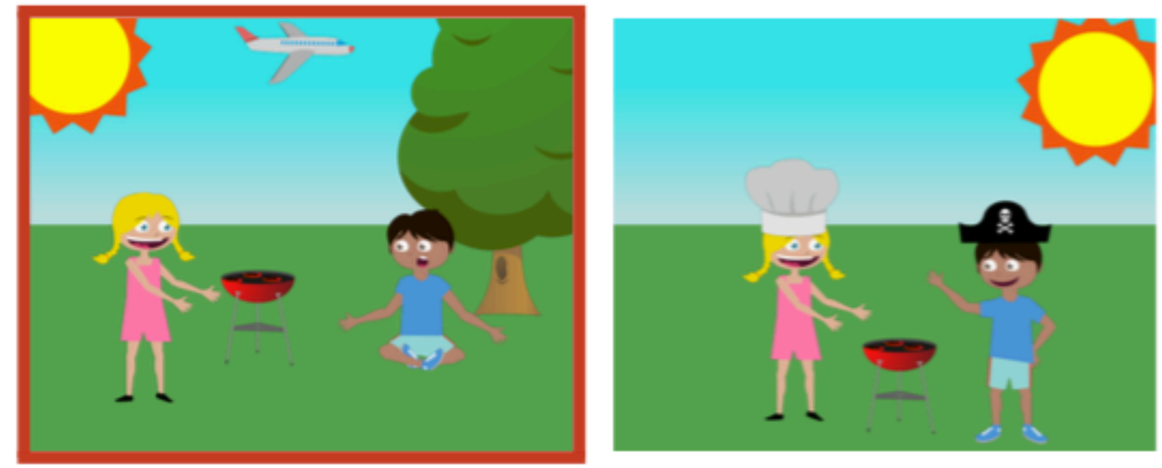
(prefer L0)	0.0	*a hamburger on the ground*
	0.1	*mike is holding the burger*
(prefer S0)	0.2	*the airplane is in the sky*

- Merely rely on Listener gives the highest accuracy but degraded fluency.
- Add only a small speaker weight achieves a good balance.

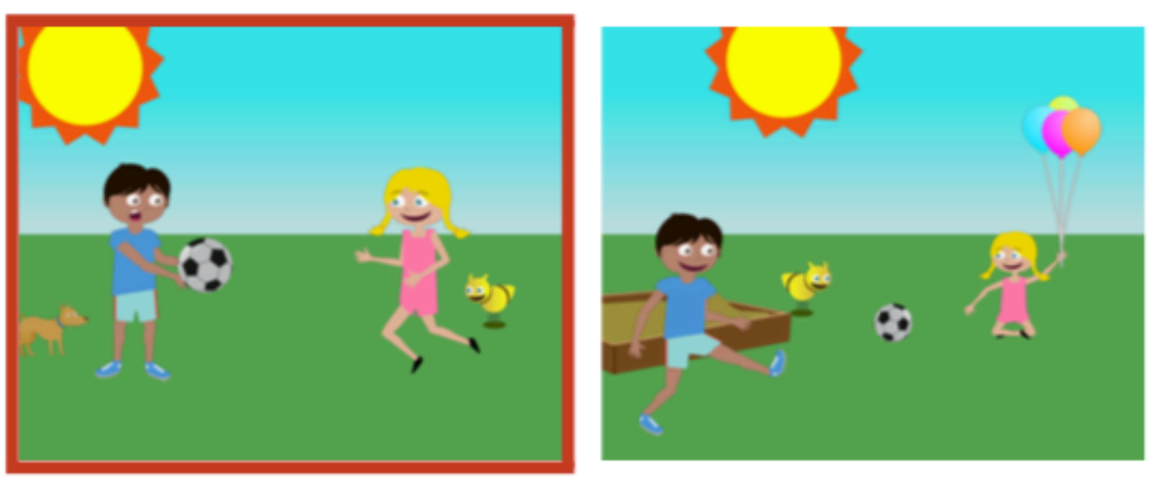
Qualitative Results



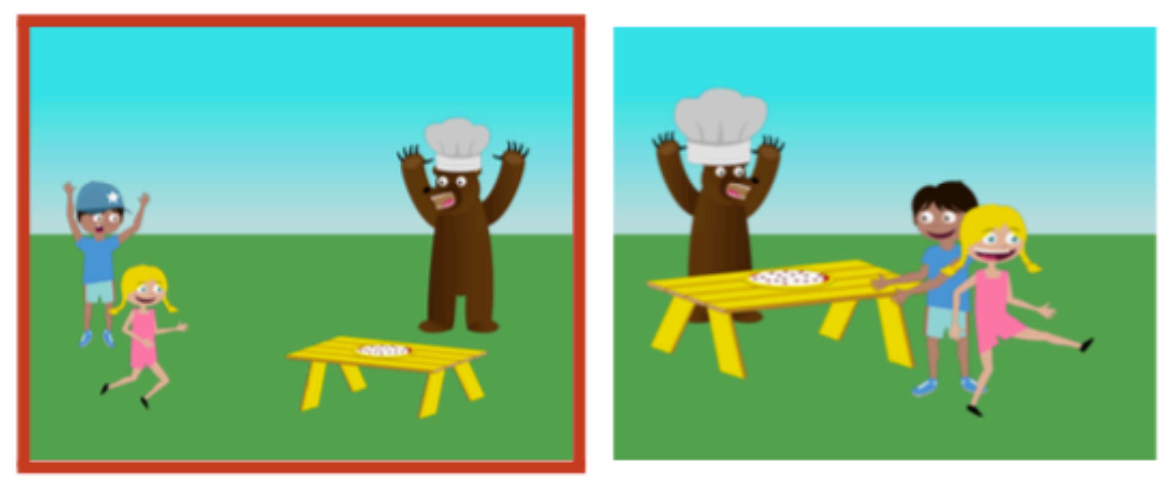
(a) *the sun is in the sky*
[contrastive]



(d) *the plane is flying in the sky*
[contrastive]



(c) *the dog is standing beside jenny*
[contrastive]



(b) *mike is wearing a chef's hat*
[non-contrastive]

Qualitative Results - contrastive

(a) (b) (c)

(b vs. a) *mike is holding a baseball bat*

(b vs. c) *the snake is slithering away from mike and jenny*

- The model is able to produce contrastive description even though the speaker is trained on non-contrastive images.

Comments

- Pros:
  - A good practice to combine two streams of the literatures.
  - All the sub-modules are several linear layers, making the system clear and efficient. And the qualitative results are fairly good.
- Cons:
  - The model achieve best accuracy with L0, making it hard to claim that language fluency is important for referring games.
  - The speaker is still not contrastive, this may lead to an inherent difficulty for fine-grained scenes.
  - The human evaluation is infeasible and unfair. Is there better evaluation for referring game?
  - The training is based on hand-craft features and not end-to-end.