ShapeGlot: Learning Language for Shape Differentiation

Panos Achlioptas[1], Judy Fan[2], Robert Hawkins[3], Noah Goodman[1], Leonidas Guibas[1]
[1]Stanford University, [2]UC San Diego, [3]Princeton University

Background & Motivation

- People can refer to the **shape** (geometry, topology) of an object to *distinguish* it among others objects.
- **Existing studies** rely primarily on properties like *color* and *spatial location* to refer to an object.
- **Existing studies** work explicitly with 2D images and are 'blind' to the *part-based* compositionality of 3D objects, or their *fine-grained* geometry.

This Work

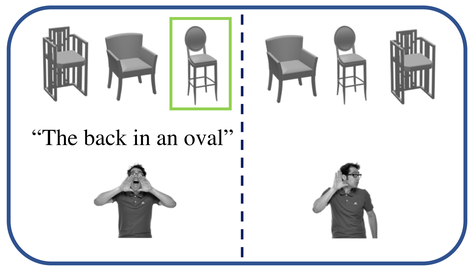
- Builds a large scale **multi-modal** dataset calibrated for shape-based reference, aka ShapeGlot!
- Introduces *novel* ***speaker-listeners*** considering the effect of using:
  a) 2D & 3D object representations
  b) context-based discrimination
  c) neural word-attention
  d) pragmatic referential reasoning
- Discovers a plethora of **surprising generalization** scenarios.

Code & Data

www.bit.ly/shapeglot

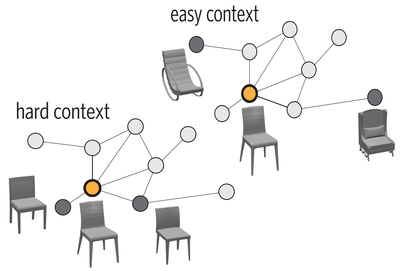
Making ShapeGlot

- Tap on 'pure' 3D meshes to make a reference game



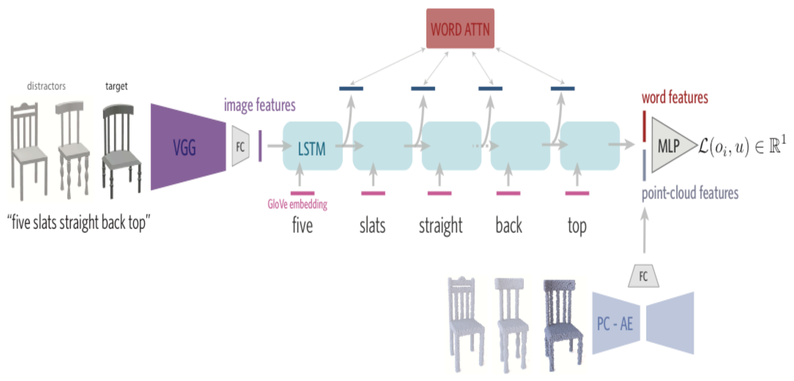
✓Free-form language
✓By construction: **only** about shape

- Use *latent*-based context formation

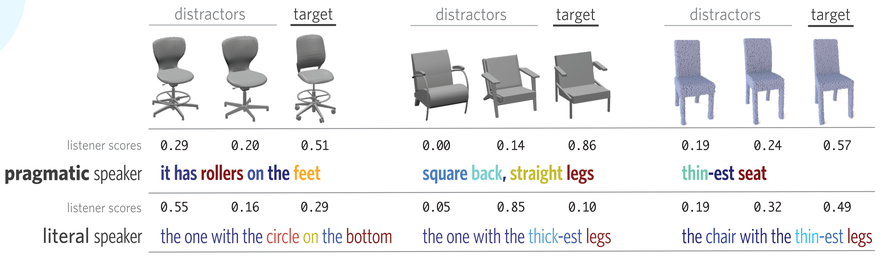


✓80K Utterances, 4K contexts

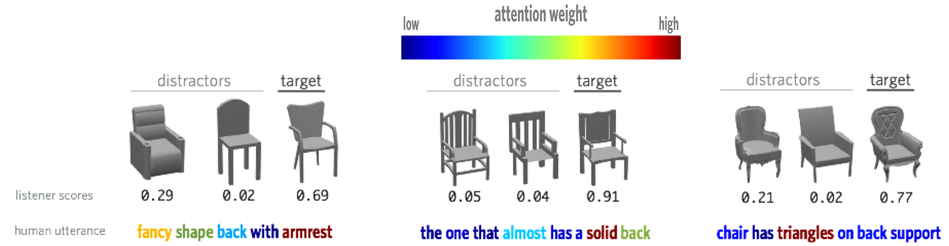
Multi-modal Attentive Neural Listeners



Pragmatic Neural Speakers



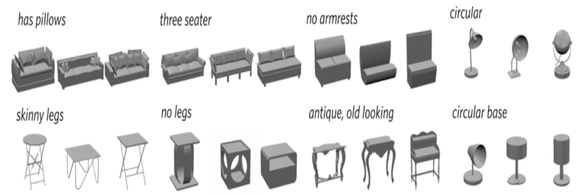
Listener's "fit" $\beta \log(P_L(t|U,O)) + \frac{(1-\beta)}{|U|^\alpha} \log(P_S(U|O,t))$ Speaker's "fit"



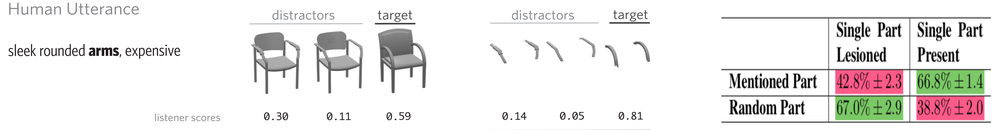
Neural attention provides intuitive model interpretation

	Input Modality	Language Task	Object Task
No Attention	Point Cloud	$67.6 \pm 0.3\%$	$66.4 \pm 0.7\%$
	Image	$81.2 \pm 0.5\%$	$77.4 \pm 0.7\%$
	Both	$83.1 \pm 0.4\%$	$78.9 \pm 1.0\%$
With Attention	Point Cloud	$67.4 \pm 0.3\%$	$65.6 \pm 1.4\%$
	Image	$81.7 \pm 0.5\%$	$77.6 \pm 0.8\%$
	Both	$\mathbf{83.7} \pm 0.3\%$	$\mathbf{79.6} \pm 0.8\%$

Listening Ablations



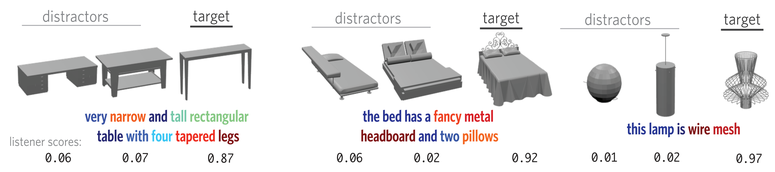
Listeners for *novel* shape-retrieval



	Single Part Lesioned	Single Part Present
Mentioned Part	$42.8\% \pm 2.3$	$66.8\% \pm 1.4$
Random Part	$67.0\% \pm 2.9$	$38.8\% \pm 2.0$

If used, object parts (e.g. visual 'arms') are necessary & sufficient for disambiguation

Zero-shot Listening (in unseen class **&** lang.)



Zero-shot Speaking (in catalogue models)

Speaker Architecture	Modality	Neural Listener	Human Listener
Context Unaware	Point Cloud	$59.1 \pm 2.0\%$	-
	Image	$64.0 \pm 1.7\%$	-
Literal	Point Cloud	$71.5 \pm 1.3\%$	66.2
	Image	$76.6 \pm 1.0\%$	68.3
Pragmatic	Point Cloud	$90.3 \pm 1.3\%$	69.4
	Image	$\mathbf{92.2} \pm 0.5\%$	**78.7**

Speaking Ablations

Key Take Away Points

- Shape-based referential language is **robust** across classes (e.g. ZSL from 'chairs' to 'lamps').
- Language *alone* enables part-based ***visual*** reasoning.
- Pragmatic neural agents perform *significantly* better than literal ones.