DISCOVERING TRANSFORMATIONS FOR ROBUST SEMANTIC VISION

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Tejas Gokhale

https://tejasgokhale.com

tgokhale@asu.edu

@trgokhale

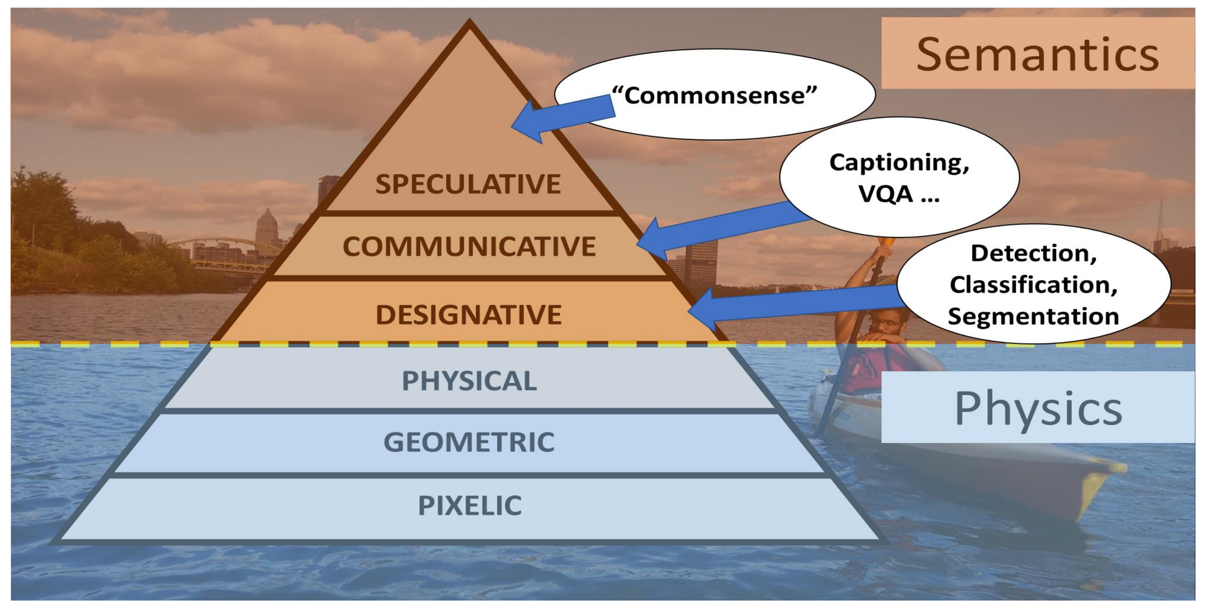
On the Academic Job Market for 2023!

Semantic vision seeks to assign "meaning" to what we see. My PhD thesis addresses several aspects of robustness in semantic vision, by:

- ❑ Identifying Failure Modes of Semantic Vision Models
- ❑ Creating evaluation tools, datasets, and benchmarks to diagnose failures
- ❑ Developing algorithms that discover transformations to improve robustness

Functional Adversarial Transformations for Robust Image Classification

Semantic Adversarial Transformations for Vision+Language Robustness



[AAAI 2021] Attribute-Guided Adversarial Training

- ❑ In real-world scenarios, test examples can **vary along known attributes** such as size, shape, colors, geometric transforms (rotation / translation / scaling)
- ❑ Such shift is larger than pixel-level noise (prior work on adversarial robustness) ⇒ data augmentation via norm-bounded adv. perturbations is ineffective.

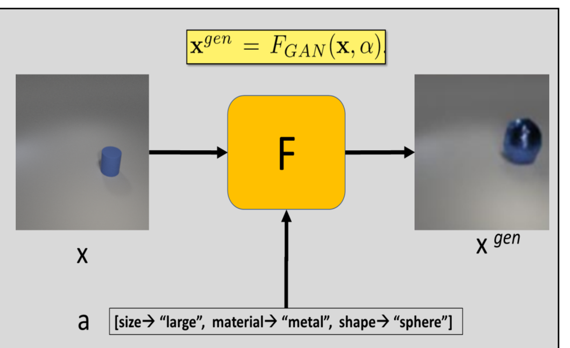
❑ **AGAT** (1) Parameterizes input space by attributes α

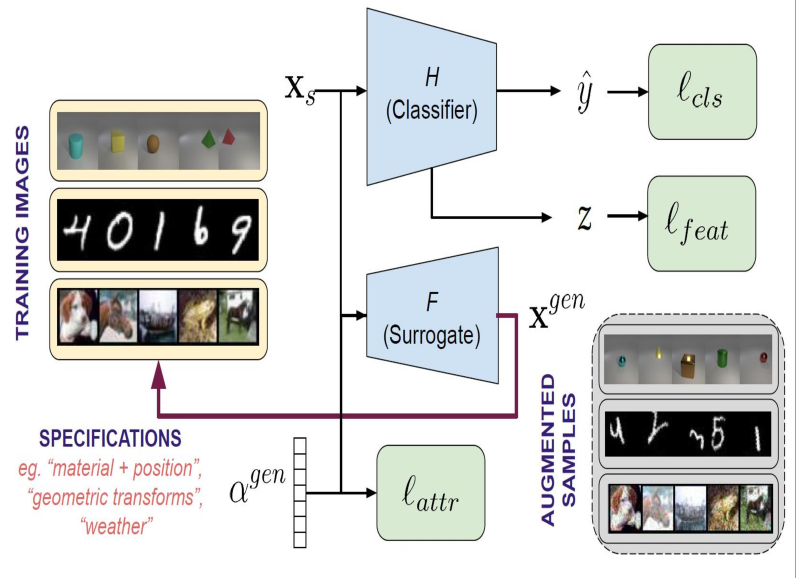
❑ (2) adversarially perturbs attribute space

$$\min_{\theta} \underset{(x,y)\in\mathcal{D}}{\mathbb{E}} \max_{d(\hat{\alpha},\alpha_x)<\epsilon} \ell(\theta;(F(x,\hat{\alpha}),y))$$

Inner Maximization → finds adversarial attributes α

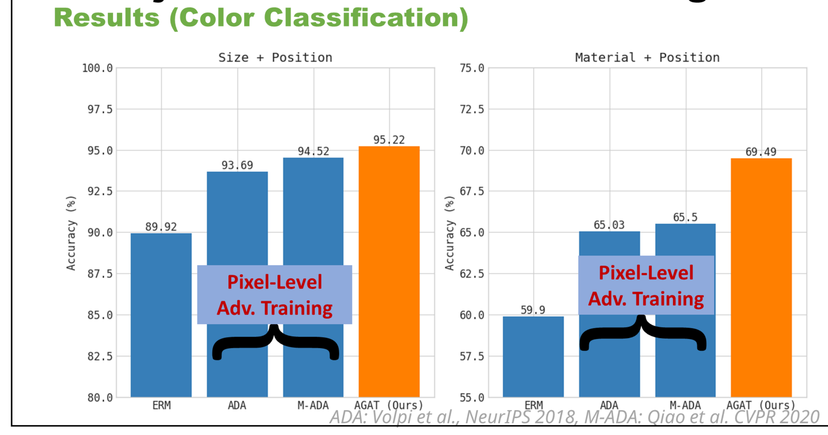
Outer Minimization: → updates classifier on $x^{gen} = F(x,\hat{\alpha})$





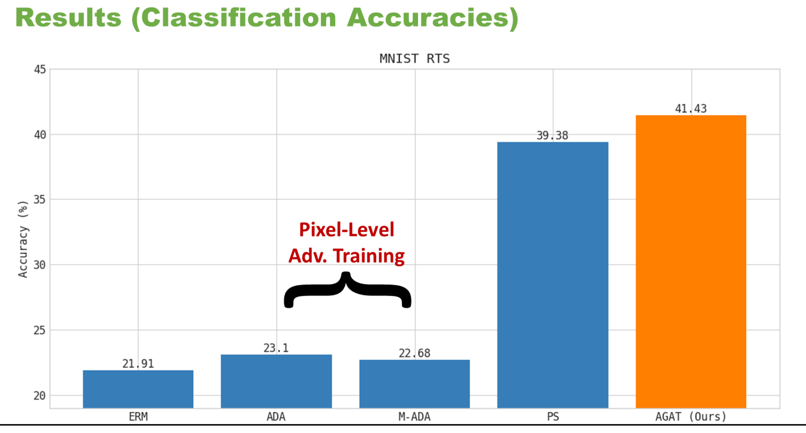
1. Object-Level Shift: CLEVR-Singles

Results (Color Classification)



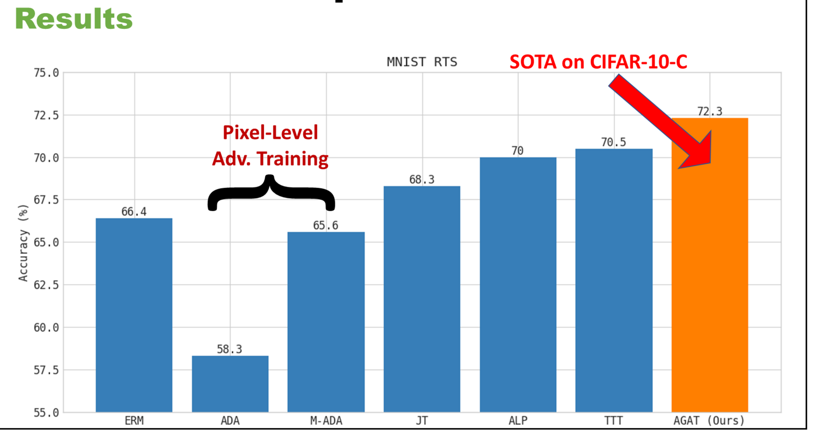
2. Geometric Transforms: MNIST-RTS

Results (Classification Accuracies)



3. Natural Corruptions: CIFAR-10-C

Results



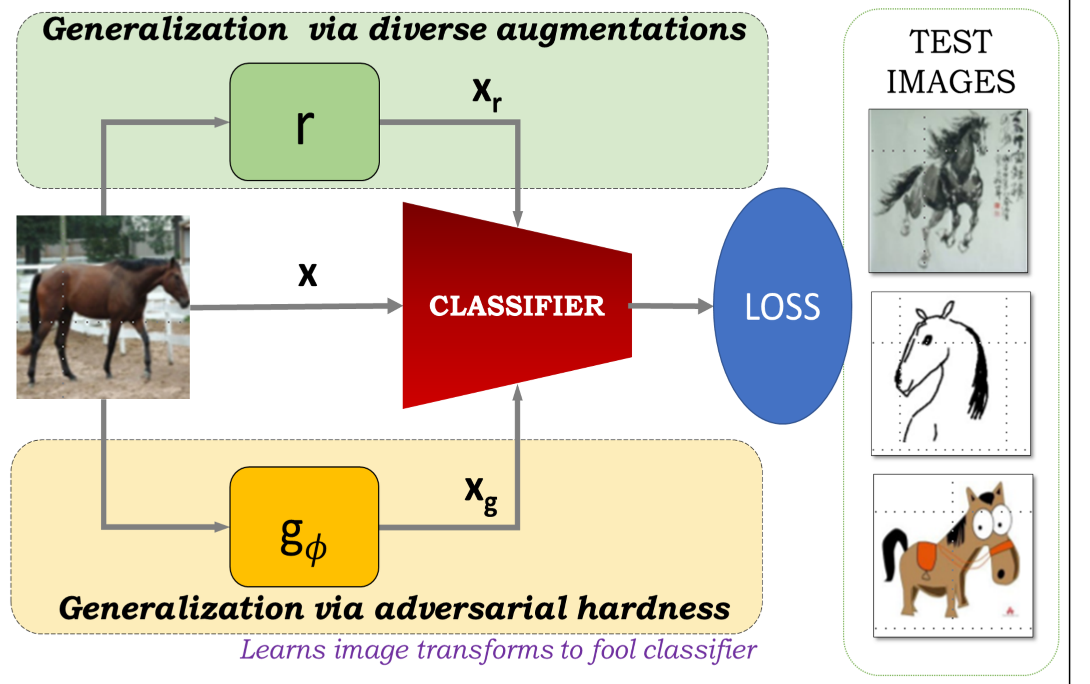
ADA: Volpi et al. NeurIPS 2018, M-ADA: Qiao et al. CVPR 2020

ALT: Improving Diversity with Adversarially Learned Transformations

- ❑ In Single Source Domain Generalization (SSDG), classifier is trained on a single domain, but expected to generalize to unseen domains
- ❑ Success of SSDG depends on maximizing diversity of training data.
  ⇒ **Data Augmentation is one of the main sources of diversity!**
- ❑ But what augmentation method should we choose? Test domains are unknown! Standard data augmentations only help on some benchmarks, not on others.
- ❑ **ALT discovers adversarial transformations that are also diverse using** an **image-to-image neural network with learnable weights ϕ**
- ❑ Instead of perturbing images directly in pixel-space, ALT learns perturbations of ϕ to generate adversarial images, and the classifier s trained on these. $x_g = \max_{\phi} \mathcal{L}_{CE}(f(g(x;\phi);\theta), y) - \mathcal{L}_{TV}(g(x;\phi))$.

❑ALT can also be combined with diversity-based augmentation functions (e.g. AugMix, RandConv) – the classifier is trained to be consistent on all transformed versions of image x

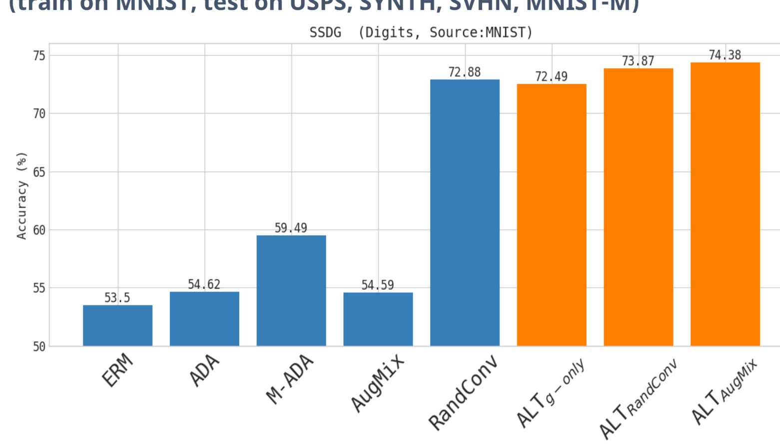
$$\mathcal{L}_{KL} = D_{KL}(p_{mix}\|p_c) + w_r D_{KL}(p_{mix}\|p_r) + (2-w_r) D_{KL}(p_{mix}\|p_g).$$



- ➢ ALT beats prior baselines on 3 SSDG benchmarks.
- ➢ ALT is significantly better than pixel-level AT
- ➢ ALT is significantly better than standard data augmentation techniques (e.g. AugMix, RandConv).
- ➢ When combined, ALT further boosts performance !

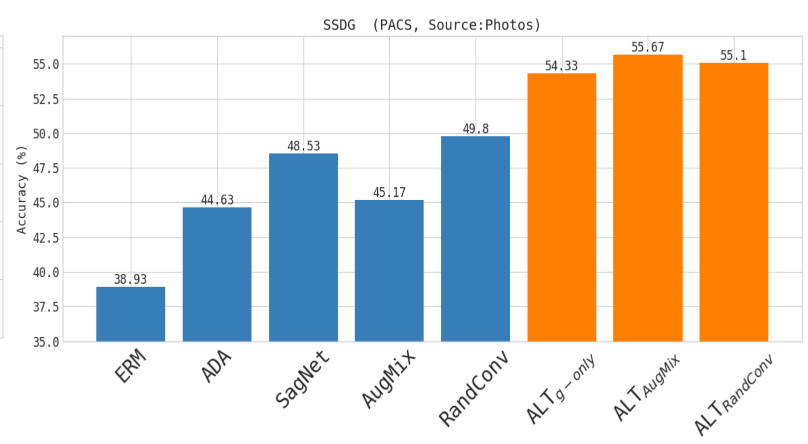
Results: SSDG on Digits (10 classes)

(train on MNIST, test on USPS, SYNTH, SVHN, MNIST-M)



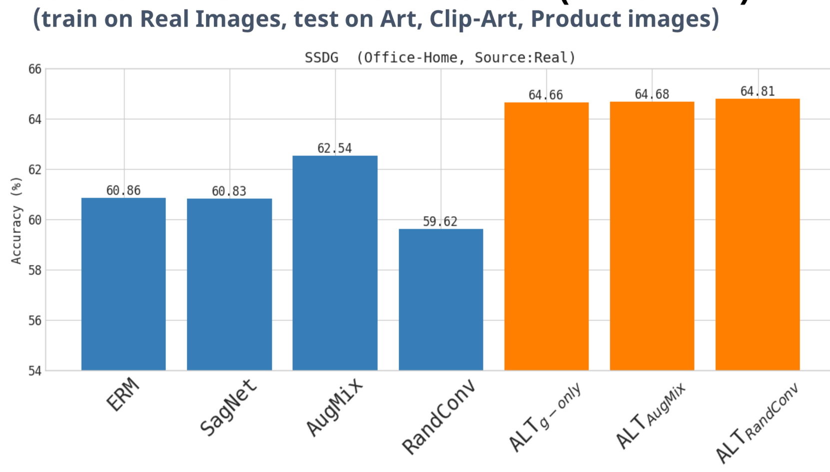
Results: SSDG on PACS (7 classes)

(train on Photos, test on Art, Cartoons, Sketches)

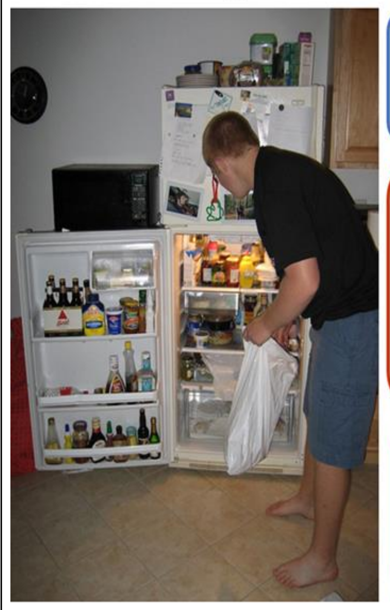


Results: SSDG on Office-Home (65 classes)

(train on Real Images, test on Art, Clip-Art, Product images)



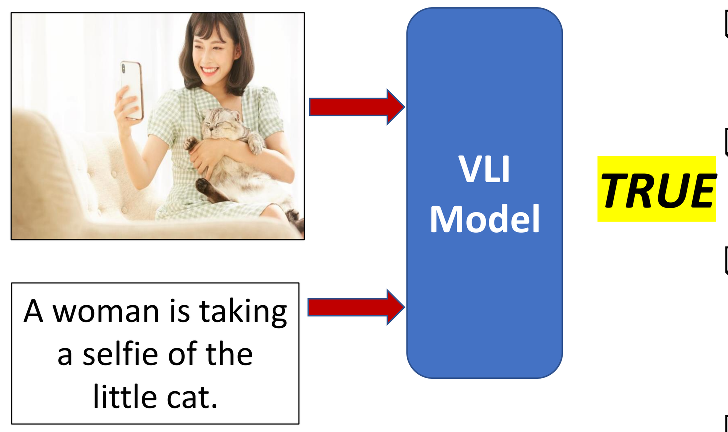
[ECCV 2020] Visual Question Answering under the Lens of Logic



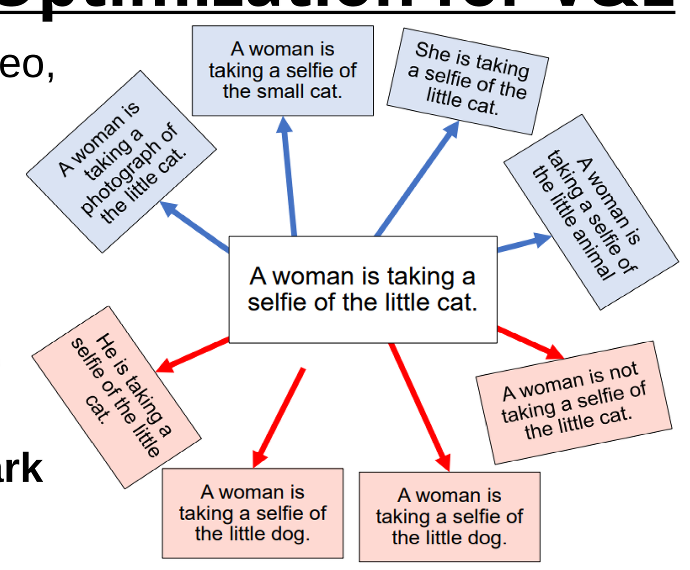
	Question	Predicted Answer	Accuracy (%) SOTA	LOL
VQA			88.20 ✓	86.55 ✓
Q_1:	Is there beer?	YES (0.96)		
Q_2:	Is the man wearing shoes?	NO (0.90)		
VQA-Compose			50.69	82.39
$\neg Q_2$:	Is the man *not* wearing shoes?	NO (0.80)		
$\neg Q_2 \wedge Q_1$	Is the man *not* wearing shoes *and* is there beer?	NO (0.62)		
$Q_1 \wedge C$	Is there beer and does this seem like a man bending over to look inside of a fridge?	NO (1.00)		
VQA-Supplement			50.61	87.80
$\neg Q_2 \vee B$	Is the man not wearing shoes or is there a clock?	NO (1.00)		
$Q_1 \wedge anto(B)$	Is there beer and is there a wine glass?	YES (0.84)		

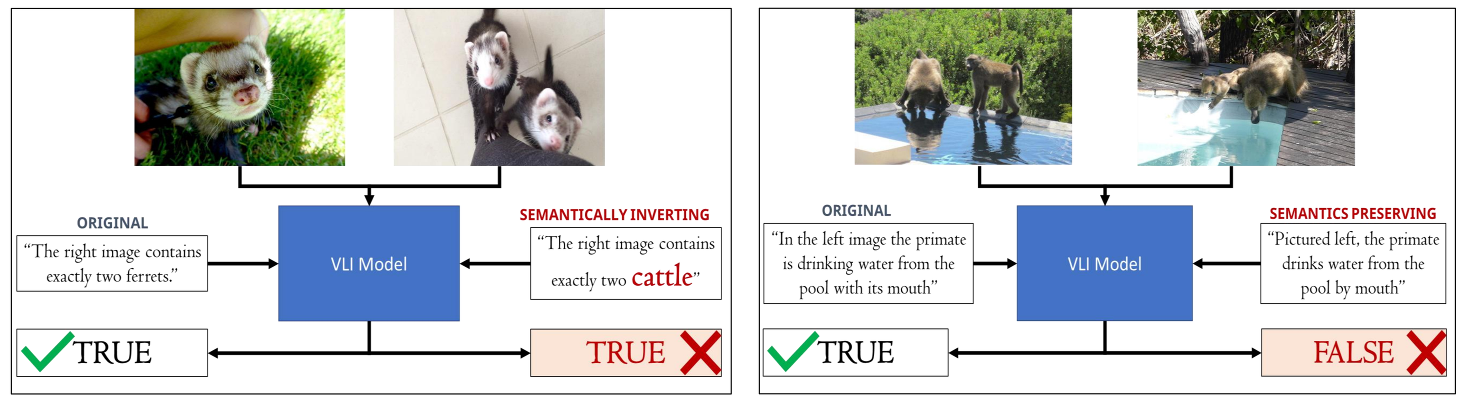
- ❑ We found VQA models to be highly susceptible to logical combinations of questions.
- ❑ VQA-LOL: a VQA benchmark for testing logical capabilities: **NEGATION, CONJUNCTION, and DISJUNCTION** of 2+ questions
- ❑ A loss inspired by Frechet Inequalities (for probabilities of events involving logical operations) improves performance compared to baselines

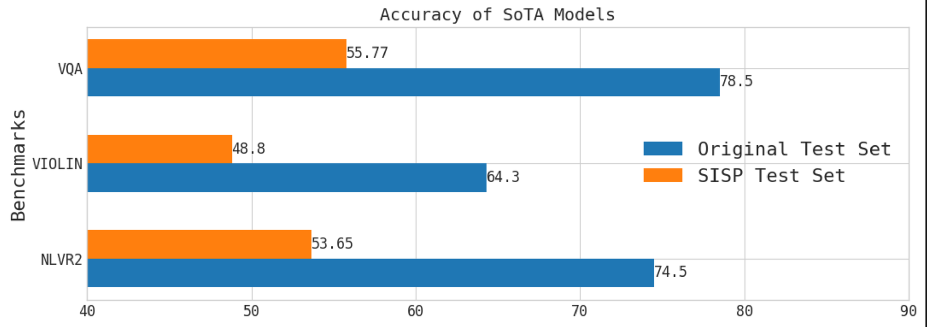
[ACL 2022] SDRO: Semantically Distributed Robust Optimization for V&L



- ❑ VLI: Vision & Language Inference (given image/video, predict if a sentence is **TRUE/ FALSE)**
- ❑ How do VLI models fare against **Linguistic Transformations of sentences**?
- ❑ To test this, we created the **SISP** benchmark using automated text transforms (SI: semantics inverting, SP: semantics preserving)
- ❑ **VLI models are unreliable on the SISP benchmark**







SDRO: SISP Transformations are ADVERSARIAL ⇒ Use them to train VLI models !

SDRO utilizes SISP transformations as the perturbation set in a DRO setting during training:

► Apply SISP to each sentence ► Find loss maximizing transformation

► Minimize classifier loss on adv samples

$$\mathcal{R}_{SDRO} = \sup_{g\in\mathcal{G}} \underset{(\mathbf{x},\mathbf{y})\sim g}{\mathbb{E}} \ell(f(\mathbf{x};\theta),\mathbf{y})$$

TEST-TIME ENSEMBLING: SISP can be leverage during inference too.

► get "N" views of input sentence using SISP ► obtain N predictions from classifier ► ensemble (avg) the predictions