**Motivation and Idea:**

**Goal:** Improve object localization capability of image classification networks.

- Training image “dog”
  - Network focuses only on the most discriminative part (dog’s face) for classification

- Training image “dog”
  - Hide patches to force the network to seek other relevant parts
    - Existing object localization methods (e.g., Oquab 2015, Zhou 2016) tend to localize only the most discriminative part.
    - Masking image pixels has been used for visualizing CNNs (Zeiler 2014), semantic segmentation (Wei 2017), and generating occlusion training examples (Wang 2017).

**Approach:**

**Training**

- For the same image, we randomly hide a different set of patches in each training epoch.
  - This allows the network to learn multiple relevant object parts for each image.

**Testing**

- During testing, the full image without any hidden patches is given as input.

**Assigning value to hidden pixels:**

- Patches are hidden only during training; during testing full image is given as input.
  - Activations of 1st conv layer will have different distribution during training and testing.
  - Assigning \( \mu \) (mean RGB value of all pixels in dataset) to each hidden pixel ensures same activation (in expectation) during training and testing:
    \[
    E[\sum_{i=1}^{k \times k} w_i x_i] = \sum_{i=1}^{k \times k} w_i \mu
    \]

**Qualitative Results:**

- AlexNet-GAP
  - [Zhou et al. CVPR 2016]

- Ours

**Failure Cases:**

- AlexNet-GAP
- Ours

- Merges spatially close instances together (first row).
- Localizes co-occurring context of a class (second row).

**Quantitative Results:**

- **Object localization results on ImageNet validation data:**
  - Methods | GT-known Loc (AlexNet) | Top-1 Loc (AlexNet) | GT-known Loc (GAP) | Top-1 Loc (GAP)
  - Backprop [Simonyan 2014] | - | 34.83 | - | 38.69
  - GAP [Zhou 2016] | 54.90 | 36.25 | 58.41 | 43.60
  - Ours | 58.68 | 37.65 | 60.29 | 45.21

- **Comparison with dropout:**
  - Methods | GT-known Loc (AlexNet) | Top-1 Loc (AlexNet) | GT-known Loc (GAP) | Top-1 Loc (GAP)
  - AlexNet | 54.90 | 36.25 | 58.41 | 43.60
  - AlexNet-Hs16 | 57.86 | 36.77 | 60.29 | 45.21
  - AlexNet-Hs32 | 58.75 | 37.33 | 61.38 | 46.31
  - AlexNet-Hs44 | 58.55 | 37.54 | 62.14 | 47.45
  - AlexNet-Hs56 | 58.43 | 37.34 | 62.93 | 48.55
  - AlexNet-Hs-mix | 58.68 | 37.65 | 63.11 | 48.76

- **Results with different patch sizes:**
  - Methods | GT-known Loc (AlexNet) | Top-1 Loc (AlexNet) | GT-known Loc (GAP) | Top-1 Loc (GAP)
  - AlexNet | 54.90 | 36.25 | 58.41 | 43.60
  - AlexNet-dropout-trainonly | 42.17 | 7.65 | 55.84 | 37.54
  - AlexNet-dropout-train + test | 53.48 | 31.68 | 67.25 | 49.31

- **Action localization results on THUMOS 2014:**
  - Pre-trained AlexNet
  - Pre-trained AlexNet with Hide-and-Seek

- **Conclusions:**
  - Simple idea of Hide-and-Seek to improve weakly-supervised object and action localization.
  - Only need to change the input without modifying the network.
  - Generalizes to multiple network architectures, input data, and tasks.

**Acknowledgement:** This work was supported in part by Intel Corp, Amazon Web Services Cloud Credits for Research, and GPUs donated by NVIDIA.