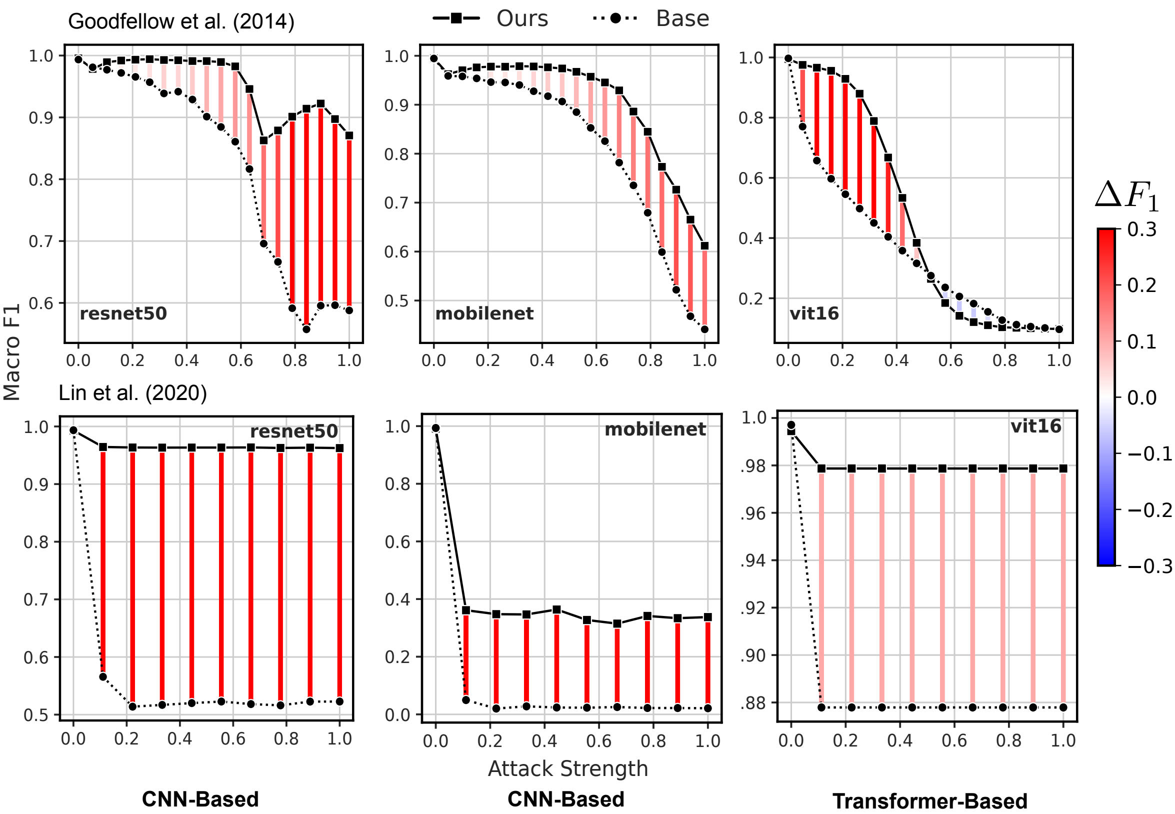

Figure 1. Regulating knowledge continuity on a host of vision models (ResNet50, MobileNetV2, and ViT16). Base models are trained with cross-entropy loss. KCReg (Our) models are finetuned with the additional regularization objective described in Alg. 2 (Lines 805-806). Two adversarial attacks are then performed: the fast-gradient sign method from [1], and an iterative attack SI-NI-FGSM from [2]. We see that regulating knowledge continuity consistently improves/stabilizes robustness. Performance is measured using F1 and the attack strength corresponds to the maximum perturbation magnitude in L2 allowed. Since the pixel values of the images are bounded between [0,1], we also constrain the attack strength to be between [0,1].

Algorithm 3. Pseudocode for certifying the robustness of a neural network using Alg. 1 and Theorem 4.1

// Probability that a δ perturbation in the j^{th} hidden layer will induce an absolute η change in accuracy

function CERTIFY($f, \{(x_i, y_i)\}_{i=1}^n, k, j, \alpha, \delta, \eta$)

  Let $\mathcal{L}$ be the 0-1 loss function

  $\varepsilon_U \leftarrow$ UPPERCONFBOUND($f, \mathcal{L}, \{(x_i, y_i)\}_{i=1}^n, k, j, \alpha$)

  $B \leftarrow \max_{1 \le a,b \le n} d_j(f^j(x_a), f^j(x_b))$

  $V \leftarrow \eta \left(1 - \exp(-2/B^2 \left(\delta - B\sqrt{\frac{1}{2}\log n}\right)^2\right)$

  return CLIP$(1 - \varepsilon_U \delta / V, 0, 1)$

end function

// Upper bound the knowledge continuity of f in the j^{th} layer using a $(1-\alpha)$ one-sided normal confidence interval

function UPPERCONFBOUND($f, \mathcal{L}, \{(x_i, y_i)\}_{i=1}^n, k, j, \alpha$)

  $U \leftarrow \mathbf{0}_k$ *// dimension-k zero-vector*

  for $i = 1 \ldots k$ **do** *// Boostrapping with k straps*

    $S \leftarrow$ *sample w/ replacement n points from* $\{(x_i, y_i)\}_{i=1}^n$

    $U_i \leftarrow$ *(Alg. 1)*$(S, \mathcal{L}, f, j)$ *// see Alg. 1 Lines 794-795*

  end for

  return $\frac{1}{k}\sum_{\ell=1}^k U_k + \Phi^{-1}(\alpha)\mathrm{std}(U)/\sqrt{k}$ *// std: standard deviation*

end function

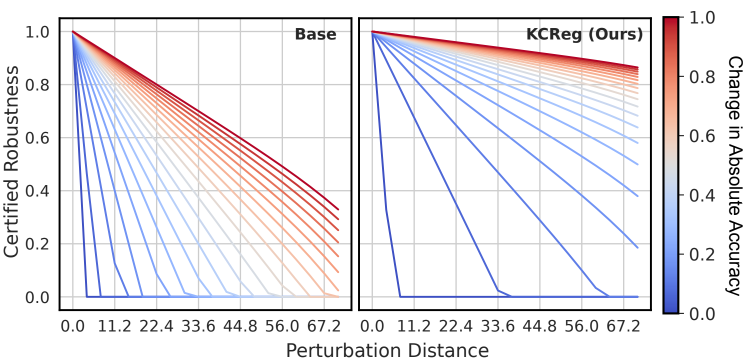

Figure 2. Certification of robustness for GPT2, layer=6. We apply Alg. 3 to certify robustness of the model before and after regularization with Alg. 2 (Line 805-806). Each line corresponds to the change in absolute accuracy for a set of examples to be considered non-robust. The y-axis corresponds to the certified probability measure of the set of non-robust examples under this criterion and the x-axis corresponds to the maximum perturbation distance in the representation space.

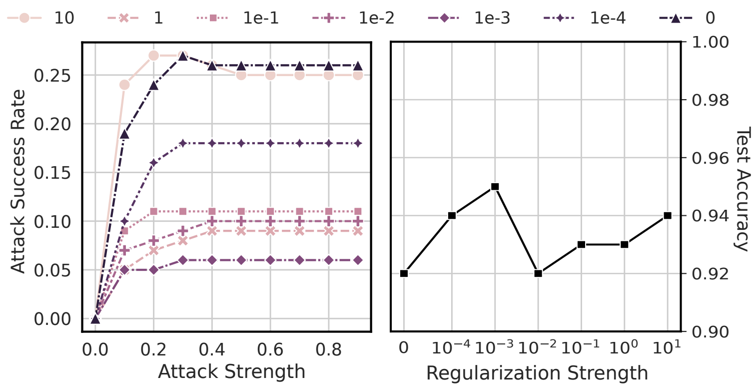

Figure 3. On the left, ablation over the strength of regularization and its effect on the attack strength-attack success rate curves. On the right, ablation over the regularization strength and its effect on test accuracy. This same curve can be observed in Fig. 7 (Lines 854-855, Pg. 29). We see that moderate regularization significantly improves robustness across all attack strengths. This improvement does not come at the expensive of test accuracy. The attack-strength is measured using the minimum angular similarity between the perturbed and original text.