Consistent Accelerated Inference via Confident Adaptive Transformers

Tal Schuster*, Adam Fisch*, Tommi Jaakkola, and Regina Barzilay (*= equal contribution)

Overview

Goal: Reduce the computational effort of multi-layered deep models while ensuring consistency with the original model $\mathcal{F}$ that uses l layers.

Approach: A new model $\mathcal{G}$ that can exit at layer $\leq l$ and guarantees:

$$\mathbb{P}(\mathcal{G}(X_{n+1}) = \mathcal{F}(X_{n+1})) \geq 1 - \epsilon$$

Specifically, our model is defined as:

$$\mathcal{G}(x;\boldsymbol{\tau}) := \begin{cases} \mathcal{F}_1(x) & \text{if } \mathcal{M}_1(x) > \tau_1, \\ \mathcal{F}_2(x) & \text{else if } \mathcal{M}_2(x) > \tau_2, \\ \quad \vdots & \\ \mathcal{F}_l(x) & \text{otherwise,} \end{cases}$$

Where $\mathcal{M}$ are confidence scores and τ are stopping thresholds.

Challenges:

1) *What is a good confidence score?* Meta predictor

2) *How to calibrate the thresholds?* Conformal calibration over inconsistent layers

Conformalized Early Exit

We look at **inconsistent** layers:

$$\mathcal{I}(x) := \{i : \mathcal{F}_i(x) \neq \mathcal{F}(x)\}, \quad i \in [1, l-1]$$

and obtain a conservative prediction set $\mathcal{C}_\epsilon$ s.t.:

$$\mathbb{P}\left(\mathcal{I}\left(X_{n+1}\right) \subseteq \mathcal{C}_\epsilon\left(X_{n+1}\right)\right) \geq 1 - \epsilon\,.$$

The first layer in the complement set provides:

$$\mathbb{P}\left(\mathcal{F}_K\left(X_{n+1}\right) = \mathcal{F}\left(X_{n+1}\right)\right) \geq 1 - \epsilon,$$

where $K := \min\left\{j : j \in \mathcal{C}_\epsilon^c\left(X_{n+1}\right)\right\}.$

Independent calibration: For each layer, compute the empirical distribution over a calibration set:

$$v_k^{(1:n,\infty)} := \left\{\mathcal{M}_k\left(x_i\right) : x_i \in \mathcal{D}_{\text{cal}}\,, \mathcal{F}_k\left(x_i\right) \neq \mathcal{F}\left(x_i\right)\right\} \cup \{\infty\}.$$

Take the quantile after MHT correction:

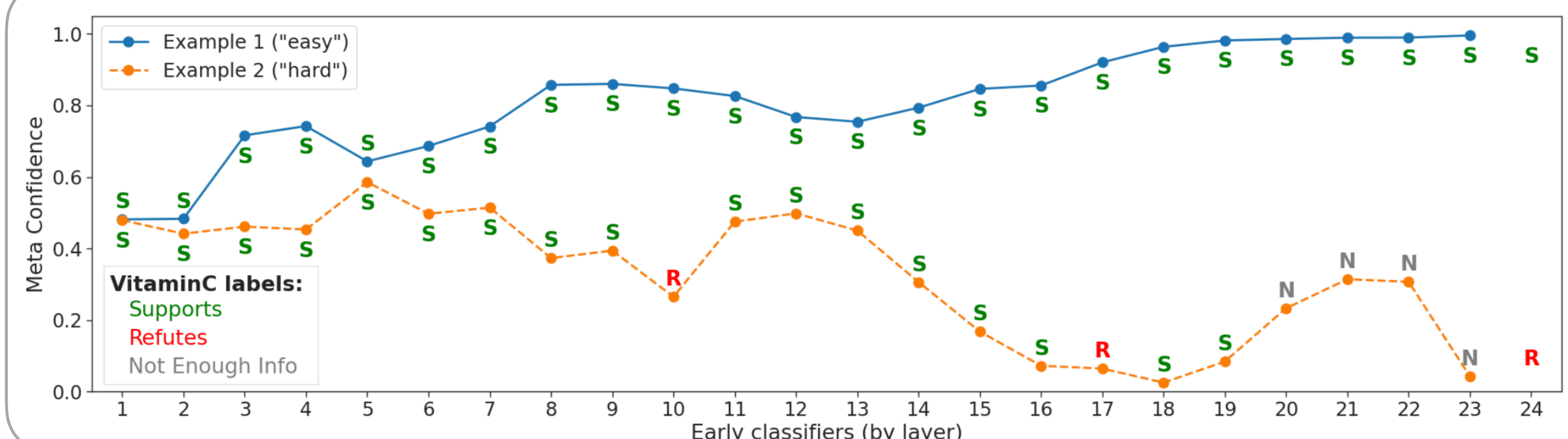
$$\tau_k^{\text{ind}} = \text{Quantile}\left(1 - \alpha_k, v_k^{(1:n,\infty)}\right).$$

Shared Calibration: Calibrate for the *worst-case* across inconsistent layers (the maximum score):

$$m^{(1:n,\infty)} := \left\{\mathcal{M}_{\max}\left(x_i\right) : x_i \in \mathcal{D}_{\text{cal}}\,, \exists k \mathcal{F}_k\left(x_i\right) \neq \mathcal{F}\left(x_i\right)\right\} \cup \{\infty\};$$

$$\tau^{\text{share}} = \text{Quantile}\left(1 - \epsilon, m^{(1:n,\infty)}\right).$$

Meta confidence per Transformer layer

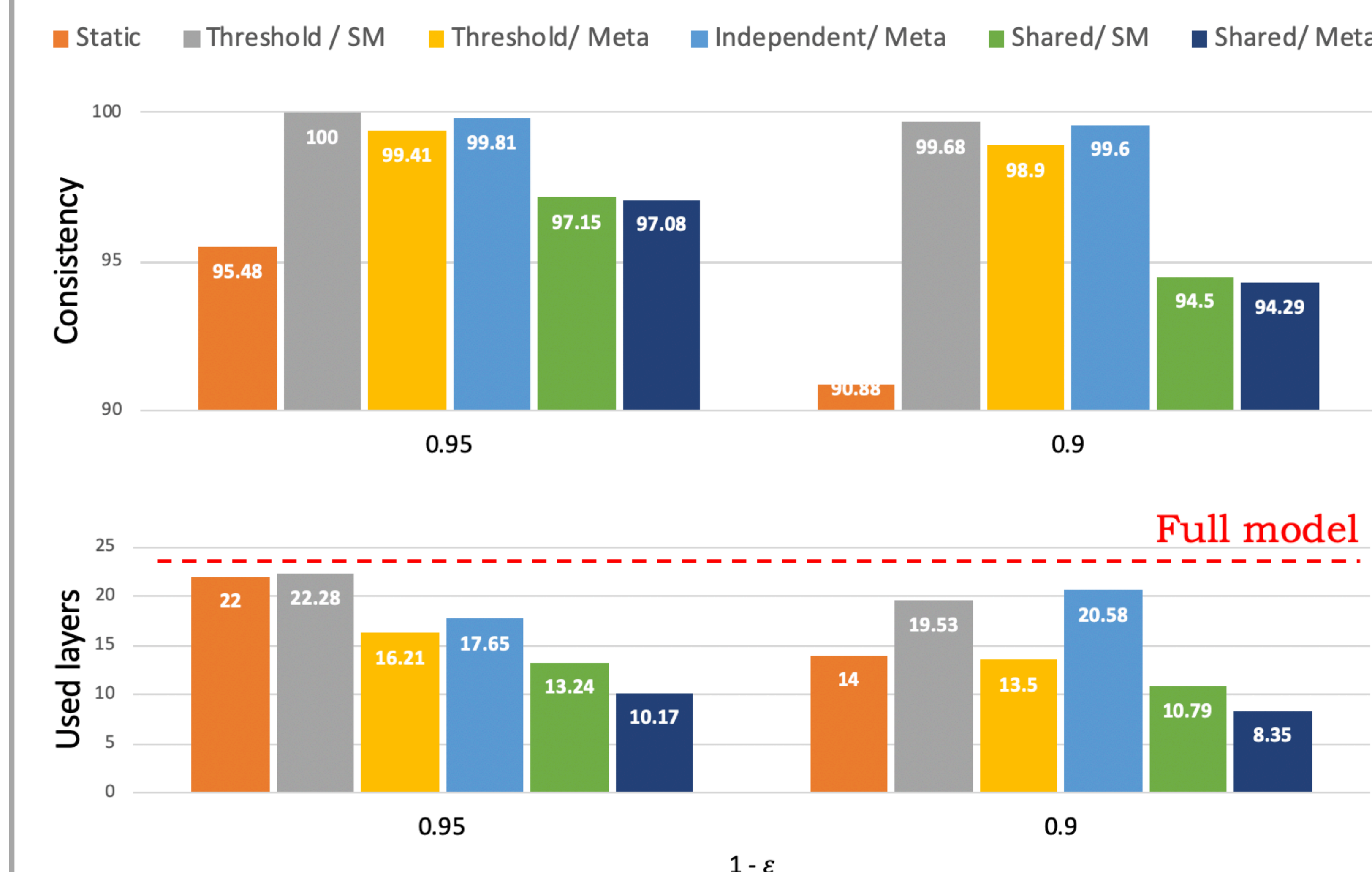


Experiments

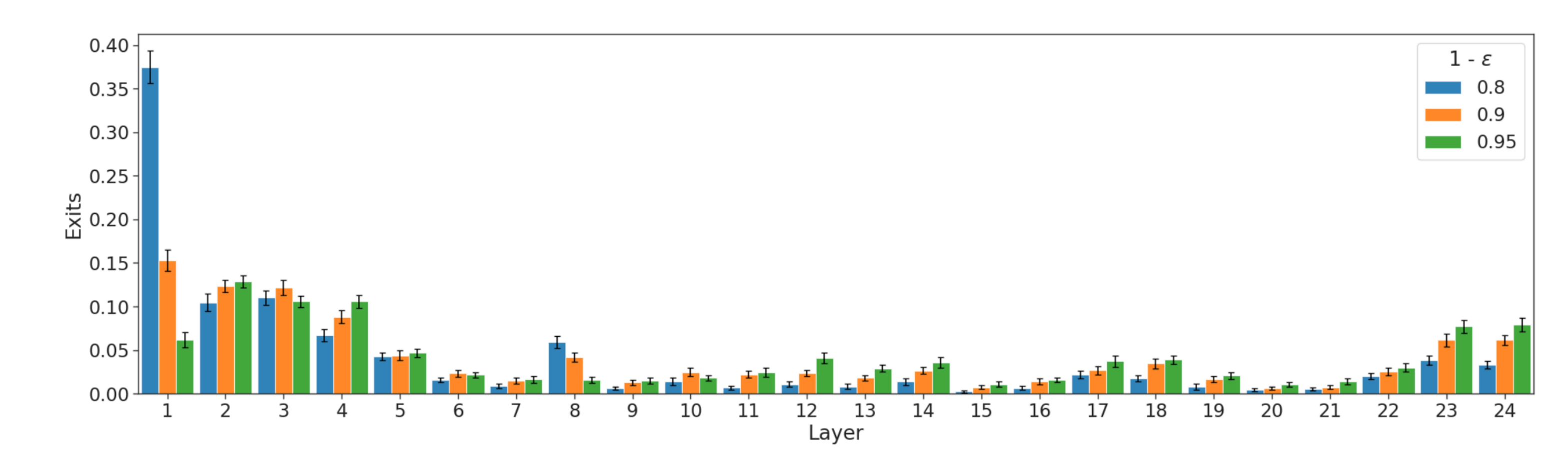
Experimental setting: we implement our **C**onfident **A**daptive **T**ransformers (CATs) on top of popular models (e.g. Albert). We evaluate on four NLP tasks covering both classification and regression.

Evaluated on IMDB, VitaminC, AG news, STS-B

Results on AG news:



The distribution of exit layers is dynamically controlled by the user-defined tolerance level:



Amortized time $(1 - \epsilon = 0.9)$: