Consistent Accelerated Inference via Confident Adaptive Transformers

Tal Schuster*, Adam Fisch*, Tommi Jaakkola, Regina Barzilay
Number of parameters in NLP models

![Graph showing the increase in number of parameters over years, with SVM indicated near the start and a lot at a later point.](image-url)
Is the full capacity always needed?

Movie review sentiment analysis:

"Everything of any interest was thoroughly covered in the original film, but like many people who have nothing to say, Part II won't shut up."

"This movie is fantastic!"

Can we use fewer layers?
Confident Adaptive Transformers

Classifier $F$ on top of the last layer $l$:
$$F(x) := H_l(T_l(T_{l-1}(...(T_1(x))))$$

Earlier classifiers:
$$F_1(x) := H_1(T_1(x))$$
$$F_2(x) := H_2(T_2(T_1(x)))$$
$$F_k(x) := H_k(T_k(...(T_1(x)))) , k < l$$

Create an amortized model $G(x)$ that can choose from $F_1,\ldots,F_l$
Our goal

Reduce computational effort (fewer layers when possible) while guaranteeing consistency with original classifier:

\[ \mathbb{P}(G(X_{n+1}) = F(X_{n+1})) \geq 1 - \epsilon \]
Challenges

How to measure confidence?
When can we exit?
How to measure confidence?

Previous models rely on intrinsic measures

- Softmax response (Huang et al., 2018; Schwartz et al. 2020; Xin et al., 2020)
- Entropy (Liu et al., 2020; Geng et al., 2021)
- Patience (Zhou et al., 2020)

- Doesn’t directly measure consistency
- Doesn’t support non-classification tasks
Meta early exit classifier

Directly **estimates the consistency**
A binary MLP $M_k(x)$ that predicts $1\{F_k(x) = F(x)\}$

Input to $M_k$:
- Early predictor hidden state: $\phi \left( W_p^{(k)} h_{[CLS]}^{(k)} \right)$
- Meta features:
  - Current prediction
  - History of predictions
  - Probability of current prediction
  - Difference in probability of top two predictions
Meta early exit classifier

(Ex.1) **Claim:** All airports in Guyana were closed for all international passenger flights until 1 May 2020.
**Evidence:** Airports in Guyana are closed to all international passenger flights until 1 May 2020.

(Ex.2) **Claim:** Deng Chao broke sales record for a romantic drama.
**Evidence:** The film was a success and broke box office sales record for mainland-produced romance films.
When can we exit?

Previous models use arbitrary thresholds

We are interested in a **marginal consistency guarantee**

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

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

\(\tau = (\tau_1, \ldots, \tau_{l-1})\) are confidence thresholds
When can we exit?

Pick one of the layers that are **consistent** with $F$

$$T(x) := \{i: F_i(x) = F(x)\}, \quad i \in [1, l - 1]$$

**Conformal prediction**

V. Vovk, A. Gammerman, and G. Shafer (2005)

Given $n$ calibration examples $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ and a desired tolerance level $\epsilon$, for a new input $X_{n+1}$:

- return a **set of predictions** $C_{n,\epsilon}(X_{n+1})$, such that

$$\mathbb{P}\left(Y_{n+1} \in C_{n,\epsilon}(X_{n+1})\right) \geq 1 - \epsilon$$

Meaning, $C_{n,\epsilon}$ contains the correct answer at least $1 - \epsilon$ of the time
Regular conformal sets don’t work

Example:

\[ T(x) = \{3, 5, ..., l - 1\} \]

Valid prediction set (contains the right answer):

\[ C_{n,\epsilon}(x) = \{2, 3, 4, l - 1\} \]

can lead to false predictions

Instead, we predict the inconsistent layers and avoid them

\[ I(x) := \{i: i \notin T(x)\}, \ i \in [1, l - 1] \]
Conformalized early exits

We look at the inconsistent layers:
\[ I(x) := \{i: F_i(x) \neq F(x)\}, \quad i \in [1, l - 1] \]

\( G \) is \( \epsilon \)-consistent if it avoids any \( I(x) \) layers more than \( \epsilon \)-fraction of the time

We obtain a conservative prediction \( C_\epsilon \):
\[ \mathbb{P}(I(X) \subseteq C_\epsilon(X)) \geq 1 - \epsilon \]

For \( K := \min\{j: j \in \overline{C_\epsilon}(X)\} \), we have: \( \mathbb{P}(F_K(X) = F(X)) \geq 1 - \epsilon \)
Independent calibration

For each layer, compute the empirical distribution of inconsistent scores:
\[ \nu_k^{(1:n,\infty)} = \{M_k(x_i): x_i \in D_{\text{cal}}, F_k(x_i) \neq F(x_i)\} \cup \{\infty\} \]

And set the threshold by its quantile:
\[ \tau_k^{\text{ind}} = \text{Quantile} \left( 1 - \alpha_k, \nu_k^{(1:n,\infty)} \right) \]

Let \( \alpha_k = \omega_k \cdot \epsilon \), where \( \sum_{k=1}^{l-1} \omega_k = 1 \), then \( C_{\epsilon}^{\text{ind}}(X) = \{k: M_k(x) \leq \tau_k^{\text{ind}}\} \) is valid

- In practice, we use uniform Bonferroni correction: \( \omega_k = 1/(l-1) \)

**Limitation:** Becomes very conservative as \( l \) grows
Shared calibration

Calibrating for the worst-case across inconsistent layers:

\[ m^{(1:n,\infty)} = \{ M_{\max}(x_i): x_i \in D_{\text{cal}}, \exists k \text{ s.t. } F_k(x_i) \neq F(x_i) \} \cup \{ \infty \} \]

Where \( M_{\max}(x) = \max_{k \in [l-1]} \{ M_k(x): F_k(x) \neq F(x) \} \)

Again, use quantile:

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

\[ C_{\epsilon}^{\text{share}}(x) = \{ k: M_k(x) \leq \tau^{\text{share}} \} \text{ is valid} \]
Evaluation

Baselines

- **Static**: Fixed number of layers for any input (tuned on calibration set)
- **Threshold**: Simply exit when the confidence score is over $1 - \epsilon$

  Confidence scores:
  - **SM**: Softmax value (only classification)
  - **Meta**: Our meta early exit score

Metrics

- **Consistency**: Prediction is similar to $F$
- **Layers**: Number of Transformer layers used

No marginal guarantees
Results per $\epsilon$ (dev)
Results per $\epsilon$ (dev) – regression task

Softmax-based baselines are invalid
Model agnostic performance

ALBERT-Xlarge

ALBERT-Base

RoBERTa-Large
Example test results (AG news)
Exit layer distribution per $\epsilon$ (IMDB)

This movie was obscenely obvious and predictable. The scenes were *poorly* written and *acted even worse*.

Hypothetical situations abound, one-time director Harry Ralston gives us the ultimate post-apocalyptic glimpse with the world dead...
Implementation options

**Synchronous**

- Layer 3
  - Exit?: yes
  - no
- Layer 2
  - Exit?: yes
  - no
- Layer 1
  - Exit?: yes

**Concurrent**

- Layer 3
  - Exit?: yes
- Layer 2
  - Exit?: yes
- Layer 1

stop
Speedup (AG news, $1 - \epsilon = 0.9$)

Amortized time (naïve synchronous implementation):

Amortized MACs:
Conclusion

• Dynamic computational effort per input “difficulty”

• Controllable consistency guarantees with the full model

• Meta early exit classifier

• Empirically demonstrated gains on four classification & regression tasks

Code: Github.com/TalSchuster/CATs