Test-time Adaptation for Machine Translation Evaluation by Uncertainty Minimization

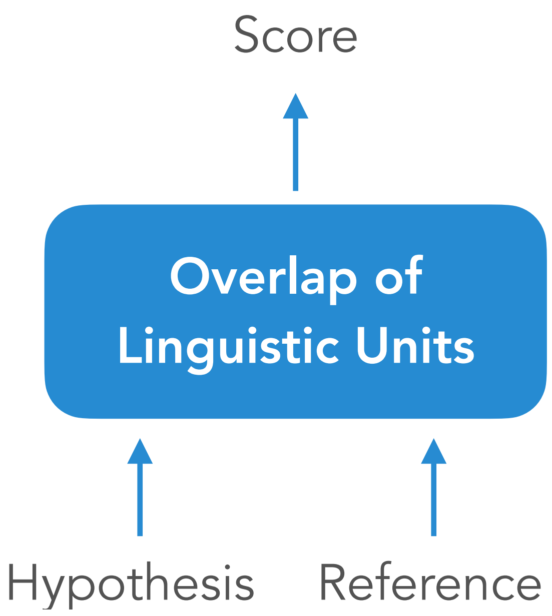
@ACL 2023

Runzhe Zhan, Xuebo Liu, Derek F. Wong, Cuilian Zhang, Lidia S. Chao, Min Zhang

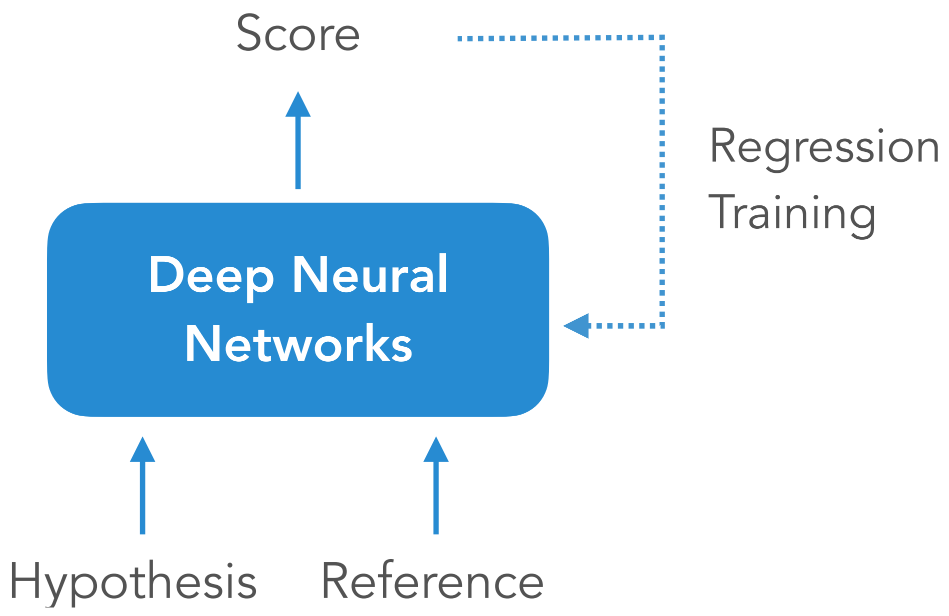
Introduction

MT Evaluation

- **Metric:** Automatically quantify the translation quality.
- **Paradigms:** Traditional metrics (e.g., BLEU), Neural metrics (e.g., COMET).



Traditional Metrics: BLEU, chrF, etc.

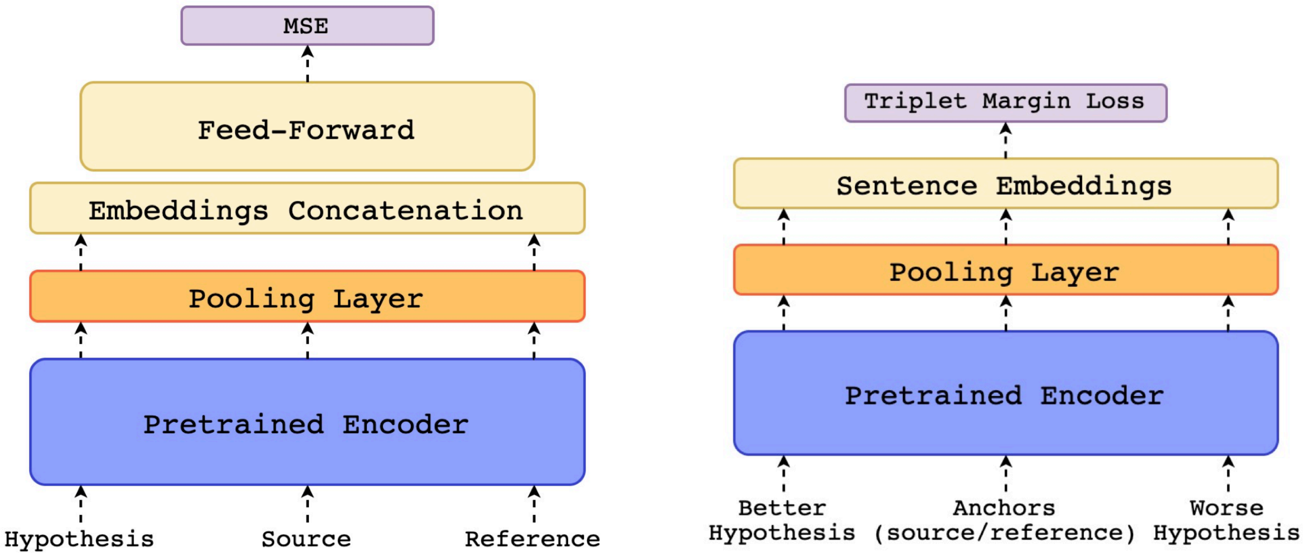


Neural Metrics: COMET, BLEURT, etc.

Introduction

Related Work

- **Representative Metric - COMET:** Fine-tune XLM-R pre-trained model with human rating data.
- Publicly available rating data solely come from WMT-News domain.



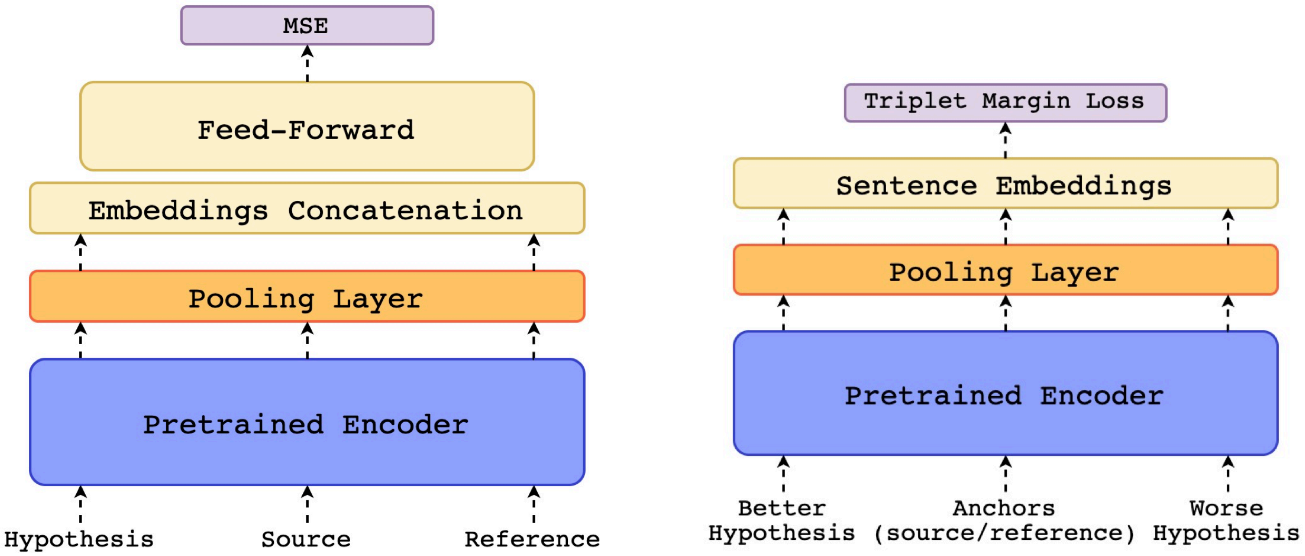
COMET Architectures.

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In *Proceedings of the EMNLP 2020* (pp. 2685-2702).

Introduction

Related Work

- **Representative Metric - COMET:** Fine-tune XLM-R pre-trained model with human rating data.
- Publicly available rating data solely come from WMT-News domain.
- **Problem:** Neural metrics were trained on WMT-News human rating data.



COMET Architectures.

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In *Proceedings of the EMNLP 2020* (pp. 2685-2702).

Introduction

Background

- **Problem:** Neural metrics were trained on WMT-News human rating data.
- **Potential Risk:** Robustness problem when evaluating out-of-distribution (OOD) samples.

Introduction

Background

- **Problem:** Neural metrics were trained on WMT-News human rating data.
- **Potential Risk:** Robustness problem when evaluating out-of-distribution (OOD) samples.
- **Results:** Neural metrics sometimes underperform traditional metrics.

	Traditional metrics			Neural metrics		
	BLEU	**TER**	**ChRF**	**COMET-DA**	**COMET-QE**	**COMET-MQM**
Chinese-English (TED Domain)	32.4	42.1	36.3	25.1	-20.9	26.6

System-level Pearson correlations (%) of metrics with human scores on WMT21 Metrics Shared Task - MQM.

Introduction

Dilemma

- **Direct Solution:** Collect human scores for out-of-distribution (OOD) samples.
  - ✘ **Cost:** Expensive to collect annotated data!

Introduction

Dilemma

- **Direct Solution:** Collect human scores for out-of-distribution (OOD) samples.
  - ✘ **Cost:** Expensive to collect annotated data!
- **Challenge:** Can we alleviate OOD problem without collecting annotated data?
  - **OOD vs. In-domain:** Performance degradation means more prediction errors.

Introduction

Dilemma

- **Direct Solution:** Collect human scores for out-of-distribution (OOD) samples.
  - ✘ **Cost:** Expensive to collect annotated data!
- **Challenge:** Can we alleviate OOD problem without collecting annotated data?
  - **OOD vs. In-domain:** Performance degradation means more prediction errors.

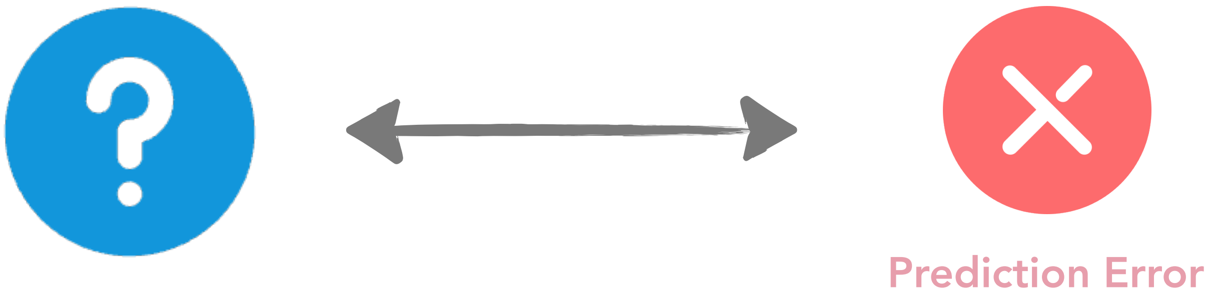
🤔 **What factors are related to the prediction errors from model perspective?**

Introduction

Dilemma

- **Direct Solution:** Collect human scores for out-of-distribution (OOD) samples.
  - ✘ **Cost:** Expensive to collect annotated data!
- **Challenge:** Can we alleviate OOD problem without collecting annotated data?
  - **OOD vs. In-domain:** Performance degradation means more prediction errors.

🤔 **What factors are related to the prediction errors <u>from model perspective</u>?**



💡 **Minimize [?] ➜ Minimize (model's) prediction error**

Introduction

Observation

Introduction

Observation

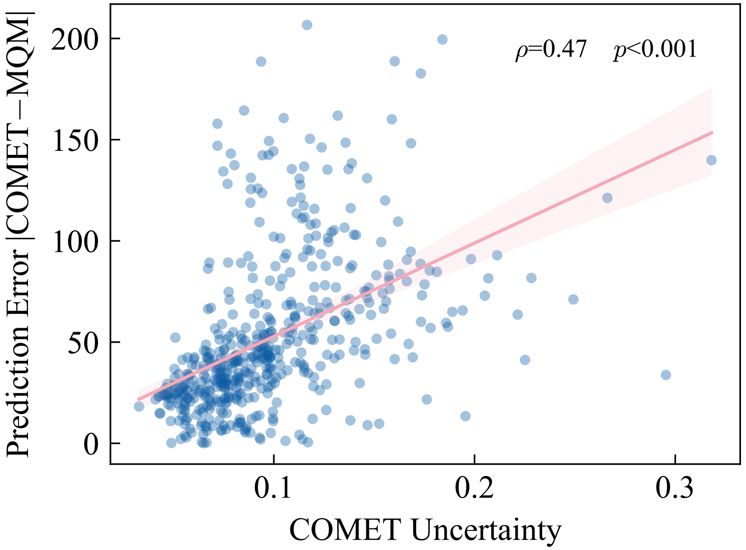
🤔 What factors are related to the prediction errors <u>from model perspective</u>?

Introduction

Observation

🤔 **What factors are related to the prediction errors from model perspective?**

- **Model uncertainty** reflects the risk of model's prediction.
- **Observation:** Model uncertainty **positively aligns** with its prediction errors.
  - Also observed by Glushkova et al. (2021).



👉

MQM: Human Ratings

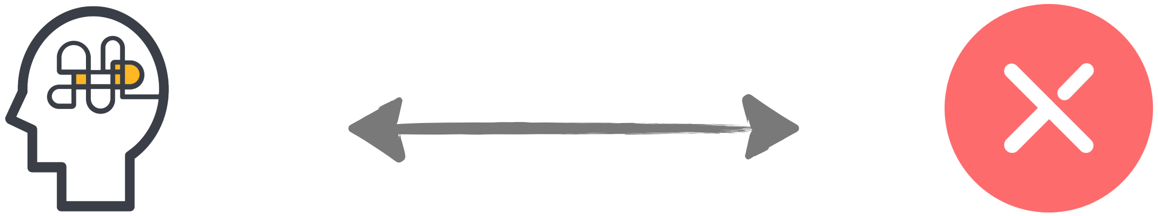
COMET: A Neural Metric

Prediction Error: Difference between metric score and human score.

Glushkova, T., Zerva, C., Rei, R., & Martins, A. F. (2021). Uncertainty-Aware Machine Translation Evaluation. In Findings of the EMNLP 2021 (pp. 3920-3938).

Introduction

Motivation



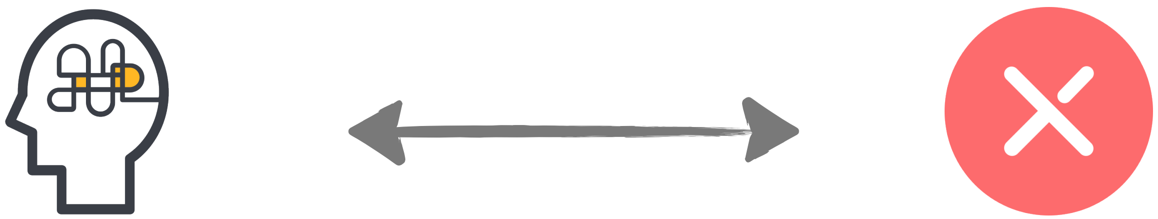
Model Uncertainty

Prediction Error

🤔 **Key Idea: Minimize (model's) <u>uncertainty</u> ➜ Minimize (model's) prediction error**

Introduction

Motivation



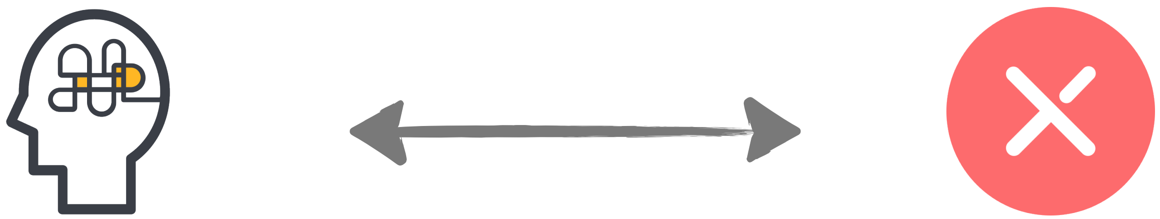
Model Uncertainty   Prediction Error

🤔 **Key Idea: Minimize (model's) <u>uncertainty</u> ➜ Minimize (model's) prediction error**

- ✓ Can be estimated **during test time**
- ✓ Can be estimated **without additional data**
- ✓ Make the model correct the predictions by itself

Introduction

Motivation



Model Uncertainty

Prediction Error

🤔 **Key Idea: Minimize (model's) uncertainty ➜ Minimize (model's) prediction error**

- ✓ Can be estimated **during test time**
- ✓ Can be estimated **without additional data**
- ✓ Make the model correct the predictions by itself

Test-time **A**daptation by **U**ncertainty Minimization **(TaU)**

Methodology

Proposal: TaU

🤔 **Key Research Questions:**

1) How can we **estimate** the uncertainty **for metrics' model** ?

2) How can we **reduce** the uncertainty by **test-time adaptation** ?

Methodology

Proposal: TaU

🤔 Key Research Questions:

1) How can we estimate the uncertainty for metrics' model ?

2) How can we reduce the uncertainty by test-time adaptation ?

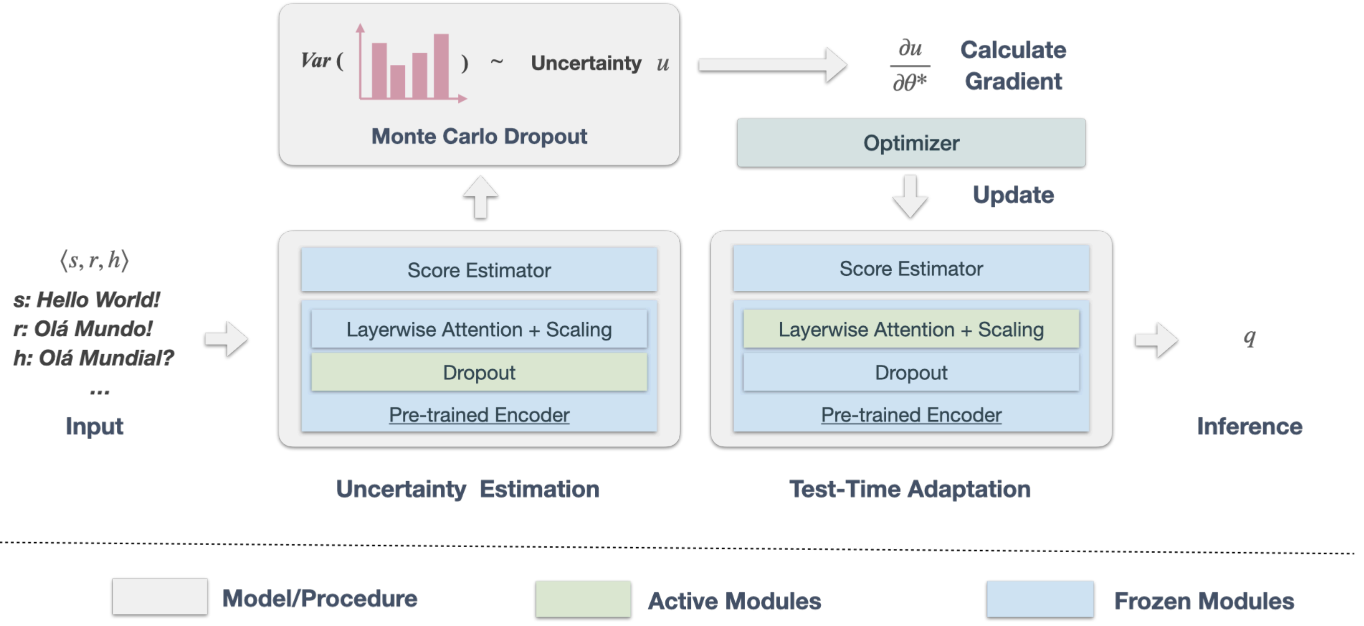


Illustration of the proposed method: Test-time Adaption by Uncertainty estimation (TaU)

Methodology

Proposal: TaU

🤔 **Key Research Questions:**

1) How can we **estimate** the uncertainty **for metrics' model** ?

2) How can we **reduce** the uncertainty by **test-time adaptation** ?

Uncertainty Estimation (RQ.1)

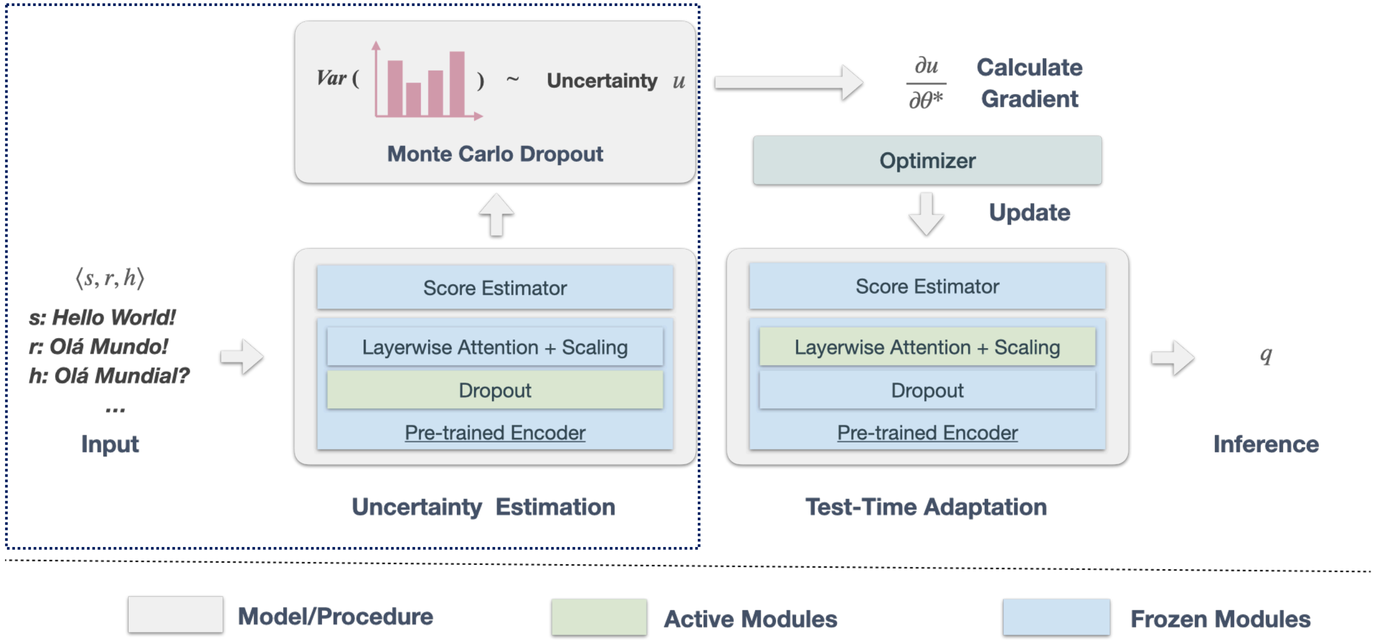


Illustration of the proposed method: **T**est-time **A**daption by **U**ncertainty estimation (**TaU**)

Methodology

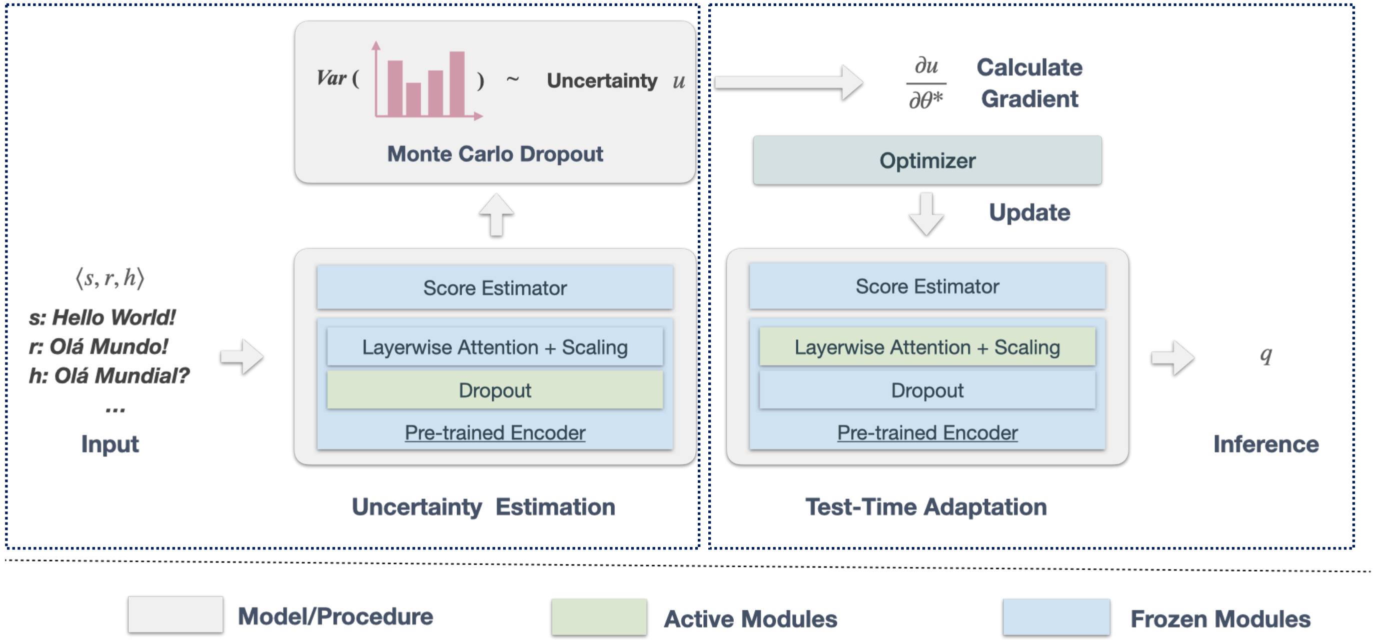
Proposal: TaU

🤔 **Key Research Questions:**

1) How can we estimate the uncertainty for metrics' model ?

2) How can we reduce the uncertainty by test-time adaptation ?

Uncertainty Estimation (RQ.1)

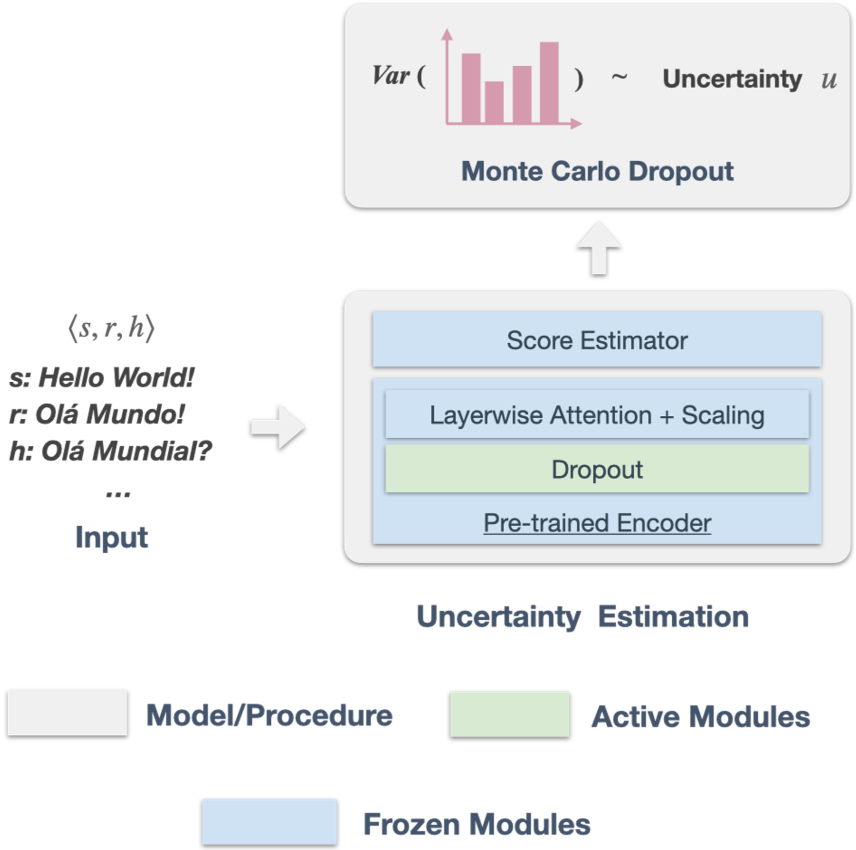


Test-time Adaptation (RQ.2)

Illustration of the proposed method: Test-time Adaption by Uncertainty estimation (TaU)

Methodology

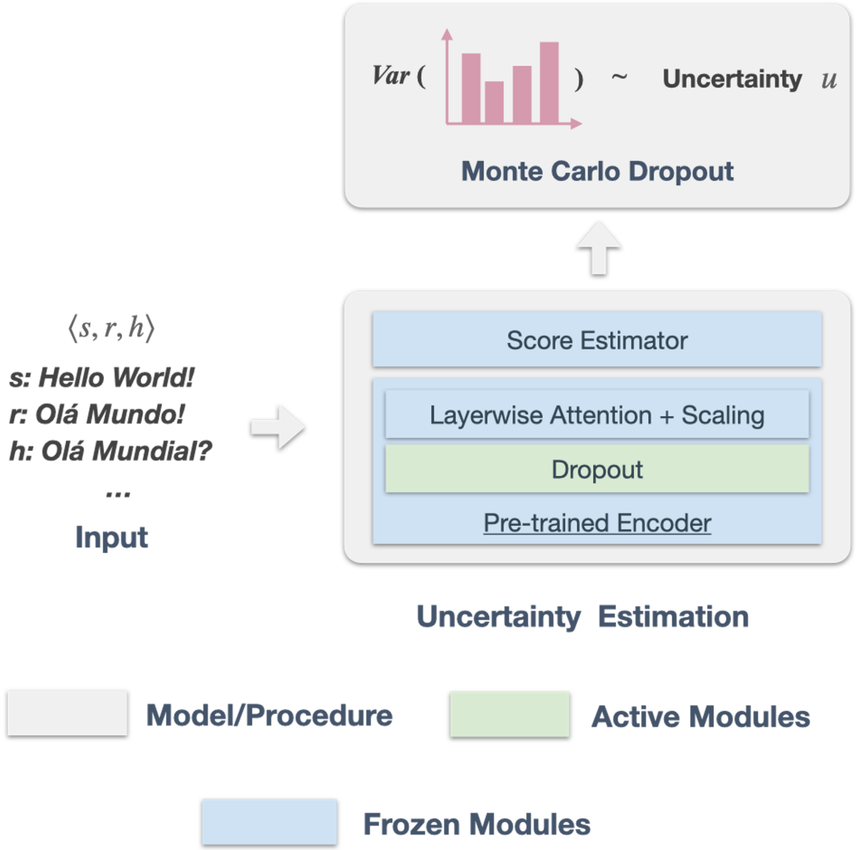
Proposal: TaU



- **Uncertainty Estimation**
  - Uncertainty = Variance of **scoring distribution**
  - Regression model only provides **a single score**

Methodology

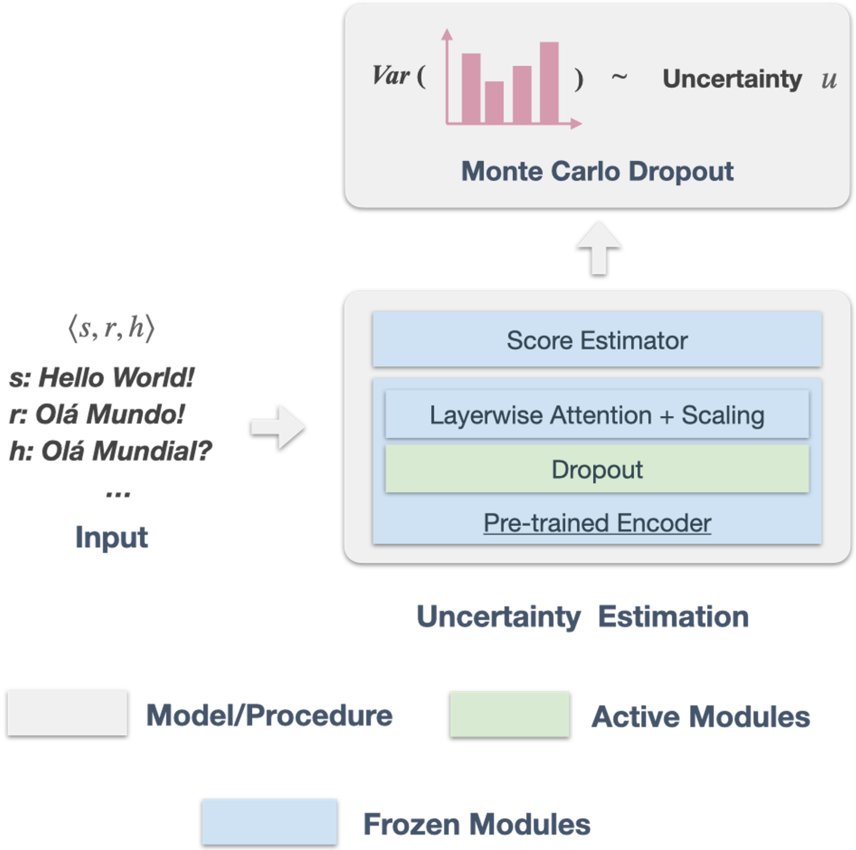
Proposal: TaU



- **Uncertainty Estimation**
  - Uncertainty = Variance of **scoring distribution**
  - Regression model only provides **a single score**
- Monte Carlo sampling method:
  - Randomly active some dropout layers and perform K-times forward-propagation.

Methodology

Proposal: TaU



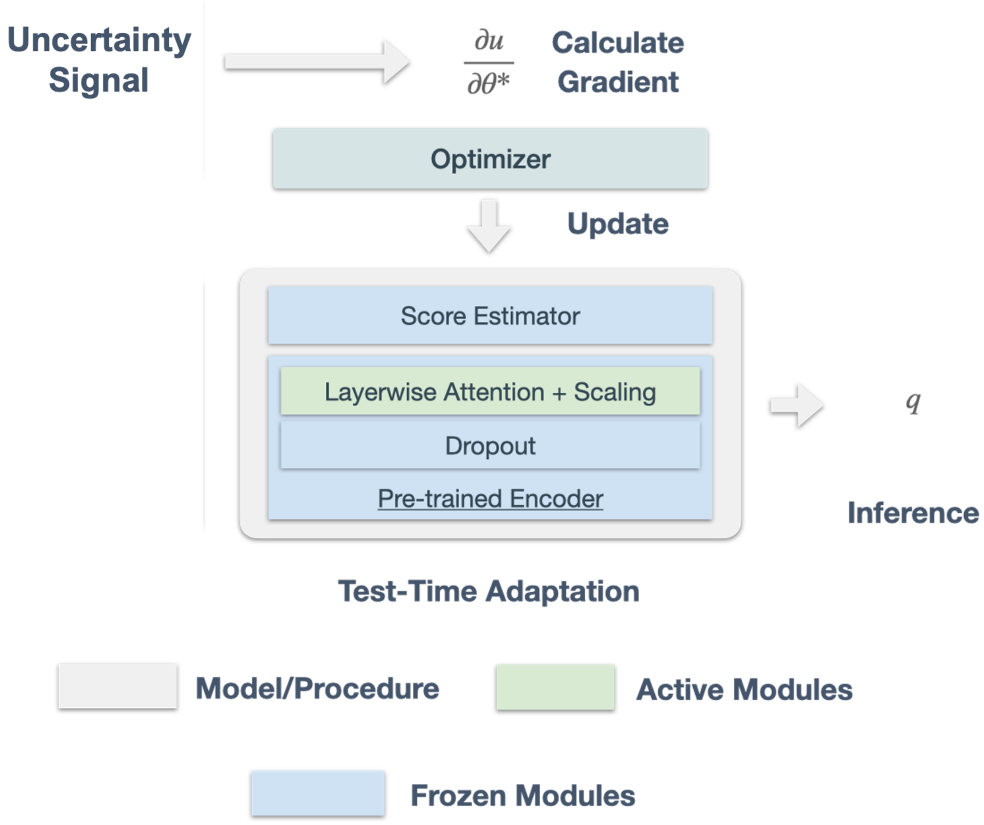
- **Uncertainty Estimation**
  - Uncertainty = Variance of **scoring distribution**
  - Regression model only provides **a single score**
- **Monte Carlo** sampling method:
  - Randomly active some **dropout layers** and perform K-times forward-propagation.
  - Uncertainty = **Var**iance of **K-times prediction**

$$u(\langle h, s, \cdot \rangle) = \mathbf{Var}(\{\mathrm{M}(\langle h, s, \cdot \rangle; \theta_k)\}_{k=1}^{K})$$

Input Data ↑ — Metric Model (w/ Dropout) ↑

Methodology

Proposal: TaU



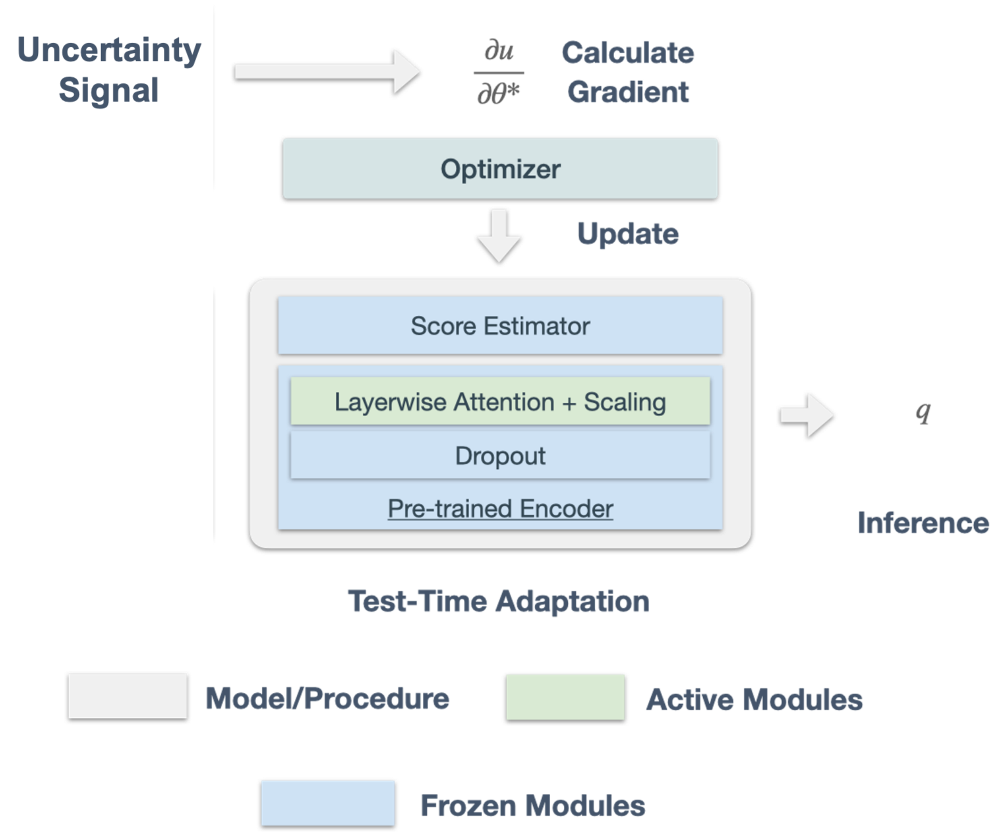
- **Test-time adaptation**
  - Online optimization
  - Objective function: minimize the uncertainty

$$\theta^* = \arg\min_{\theta^*} \mathbb{E}_{\langle h,s,\cdot\rangle \in \mathcal{D}} \left[u(\langle h, s, \cdot \rangle) \right]$$

Optimization of partial modules

Methodology

Proposal: TaU



- **Test-time adaptation**
  - Online optimization
  - Objective function: minimize the uncertainty

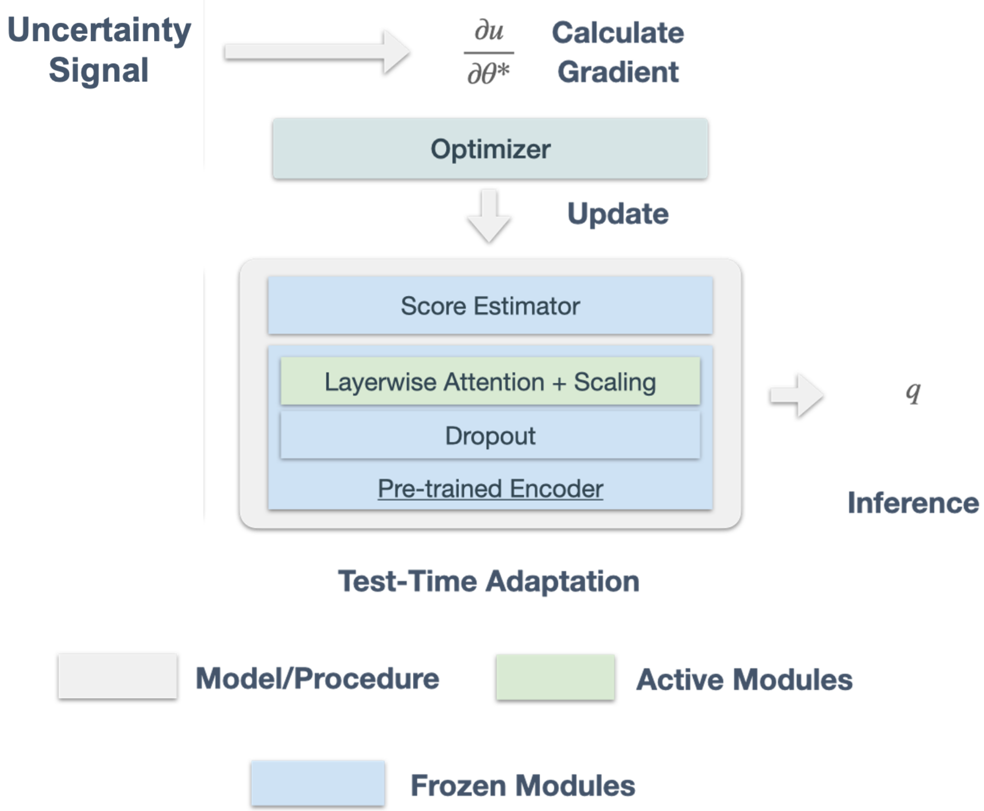
$$\theta^* = \arg\min_{\theta*} \mathbb{E}_{\langle h,s,\cdot\rangle \in \mathcal{D}} \left[u(\langle h, s, \cdot \rangle) \right]$$

Optimization of partial modules

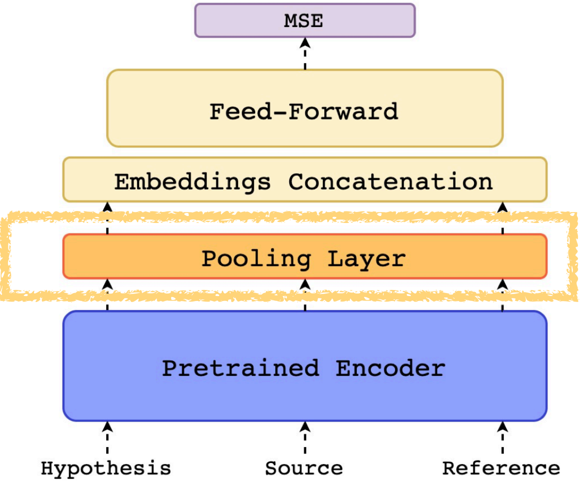
- **Choice of Optimization Parameters**
  - Do not deviate far from original parameters.
  - Only optimize partial parameters.

Methodology

Proposal: TaU



- **Choice of Optimization Parameters**

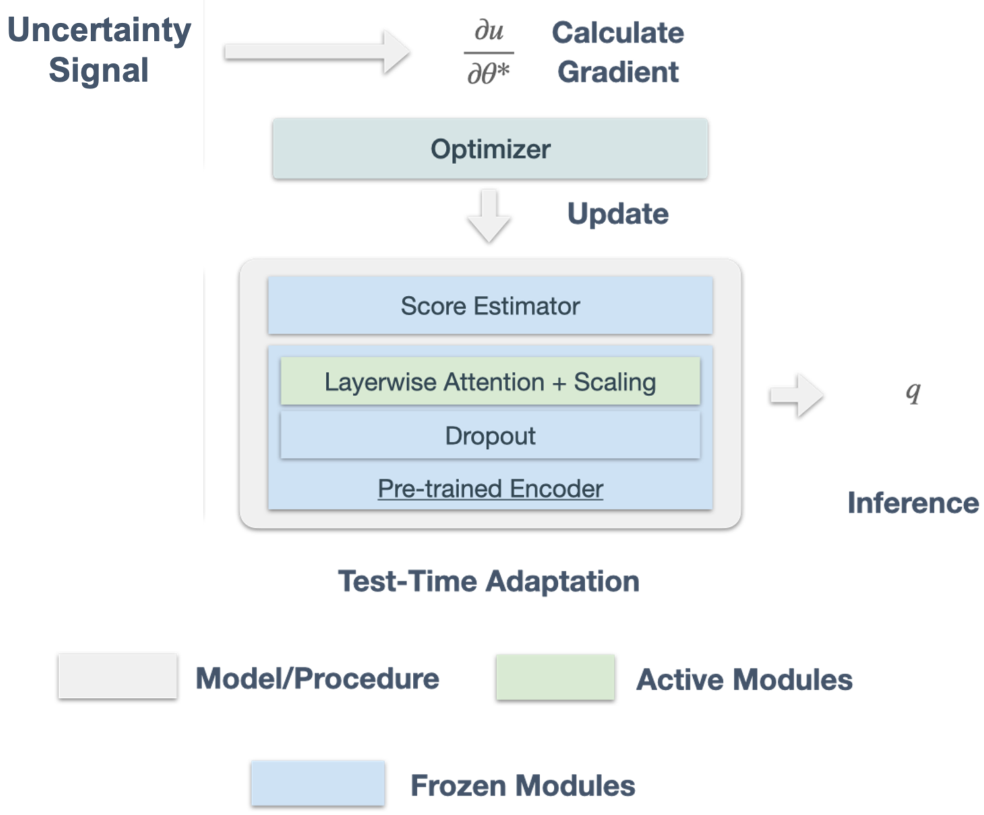


$$\mathbf{O}_{\text{embed}} = \boxed{\gamma} \cdot \sum_{l=1}^{L} \boxed{w_i} \cdot \text{LayerNorm}(\mathbf{h}_i)$$

Output of each layer

Methodology

Proposal: TaU



- **Choice of Optimization Parameters**

Domain	LAtt.	LN.	Estim.	ρ	Acc.
News	✓	✗	✗	**85.7**	**89.7**
	✗	✓	✗	79.5	76.9
	✗	✗	✓	78.7	80.8
	✓	✓	✗	79.6	76.9
	✓	✗	✓	78.6	80.8
	✓	✓	✓	78.0	79.4
TED	✓	✗	✗	**85.9**	**85.9**
	✗	✓	✗	79.4	82.1
	✗	✗	✓	77.2	76.9
	✓	✓	✗	79.4	82.1
	✓	✗	✓	77.1	76.9
	✓	✓	✓	76.9	76.9

Ablation Results.
LAtt. = Layerwise Attention | LN. = Layer Normalization
Estim. = Score Estimator

Methodology

Algorithm: TaU

Require: Model θ, System-level evaluation tuple $\mathcal{D} = \{\langle \mathbf{h}, \mathbf{s}, \cdot \rangle\}$, Adaptation rate α, Adaptation times J.

Backup original model $\theta' \leftarrow \theta$

Select parameters for adaptation $|\theta^*| \ll |\theta|$

for adaptation iteration $j = 1, .., .J$ **do**

  Score set $\mathbf{q} = \{\phi\}$

  for mini-batch $\{\langle h, s, \cdot \rangle\}_{i=1}^{N} \in \mathcal{D}$ **do**

    Estimate uncertainty u by Equation 3

    Optimize $\theta^* \leftarrow \theta^* - \alpha \nabla_{\theta^*} \frac{1}{N} \sum_{i=1}^{N} u_i$

  end for

  Infer score $[q]$ by Equation 7

  $\mathbf{q} \Leftarrow [q]$

end for

Restore to original model $\theta \leftarrow \theta'$

return q

- **Three steps: Estimate, Adapt, Predict**

1. Estimate the model uncertainty

$$u(\langle h, s, \cdot \rangle) = \mathbf{Var}(\{\mathrm{M}(\langle h, s, \cdot \rangle; \theta_k)\}_{k=1}^{K})$$

$$\mathbf{Var}(P) = \sqrt{\mathbb{E}\left[(P - \mu_P)^2\right]}$$

Methodology

Algorithm: TaU

Require: Model θ, System-level evaluation tuple $\mathcal{D} = \{\langle \mathbf{h}, \mathbf{s}, \cdot \rangle\}$, Adaptation rate α, Adaptation times J.

Backup original model $\theta' \leftarrow \theta$
Select parameters for adaptation $|\theta^*| \ll |\theta|$
for adaptation iteration $j = 1, .., .J$ **do**
  Score set $\mathbf{q} = \{\o\}$
  for mini-batch $\{\langle h, s, \cdot \rangle\}_{i=1}^{N} \in \mathcal{D}$ **do**
    Estimate uncertainty u by Equation 3
    Optimize $\theta^* \leftarrow \theta^* - \alpha \nabla_{\theta^*} \frac{1}{N} \sum_{i=1}^{N} u_i$
  end for
  Infer score $[q]$ by Equation 7
  $\mathbf{q} \Leftarrow [q]$
end for
Restore to original model $\theta \leftarrow \theta'$
return q

- **Three steps: Estimate, Adapt, Predict**

1. Estimate the model uncertainty

$$u(\langle h, s, \cdot \rangle) = \mathbf{Var}(\{\mathrm{M}(\langle h, s, \cdot \rangle; \theta_k)\}_{k=1}^{K})$$

$$\mathbf{Var}(P) = \sqrt{\mathbb{E}\left[(P - \mu_P)^2\right]}$$

2. Adapt by minimizing the uncertainty

$$\theta^* = \arg\min_{\theta^*} \mathbb{E}_{\langle h,s,\cdot \rangle \in \mathcal{D}} \left[u(\langle h, s, \cdot \rangle)\right]$$

Methodology

Algorithm: TaU

Require: Model θ, System-level evaluation tuple $\mathcal{D} = \{\langle \mathbf{h}, \mathbf{s}, \cdot \rangle\}$, Adaptation rate α, Adaptation times J.

```
Backup original model θ' ← θ
Select parameters for adaptation |θ*| ≪ |θ|
for adaptation iteration j = 1, .., .J do
    Score set q = {ø}
    for mini-batch {⟨h, s, ·⟩}_{i=1}^N ∈ D do
        Estimate uncertainty u by Equation 3
        Optimize θ* ← θ* − α∇_{θ*} (1/N) Σ_{i=1}^N u_i
    end for
    Infer score [q] by Equation 7
    q ⇐ [q]
end for
Restore to original model θ ← θ'
return q
```

- **Three steps: Estimate, Adapt, Predict**

1. Estimate the model uncertainty

$$u(\langle h, s, \cdot \rangle) = \mathbf{Var}(\{\mathrm{M}(\langle h, s, \cdot \rangle; \theta_k)\}_{k=1}^{K})$$

$$\mathbf{Var}(P) = \sqrt{\mathbb{E}\left[(P - \mu_P)^2\right]}$$

2. Adapt by minimizing the uncertainty

$$\theta^* = \arg\min_{\theta^*} \mathbb{E}_{\langle h,s,\cdot \rangle \in \mathcal{D}} \left[u(\langle h, s, \cdot \rangle)\right]$$

3. Predict with the adapted parameters

$$q = \mathrm{M}_{\theta + \Delta\theta^*}(\{\langle h, s, \cdot \rangle\})$$

Experiments

TaU: Performance

- **Testbed:** COMET models, Developmental data: WMT20
- Improved system-level Pearson's correlation on partial multi-domain evaluation tasks.

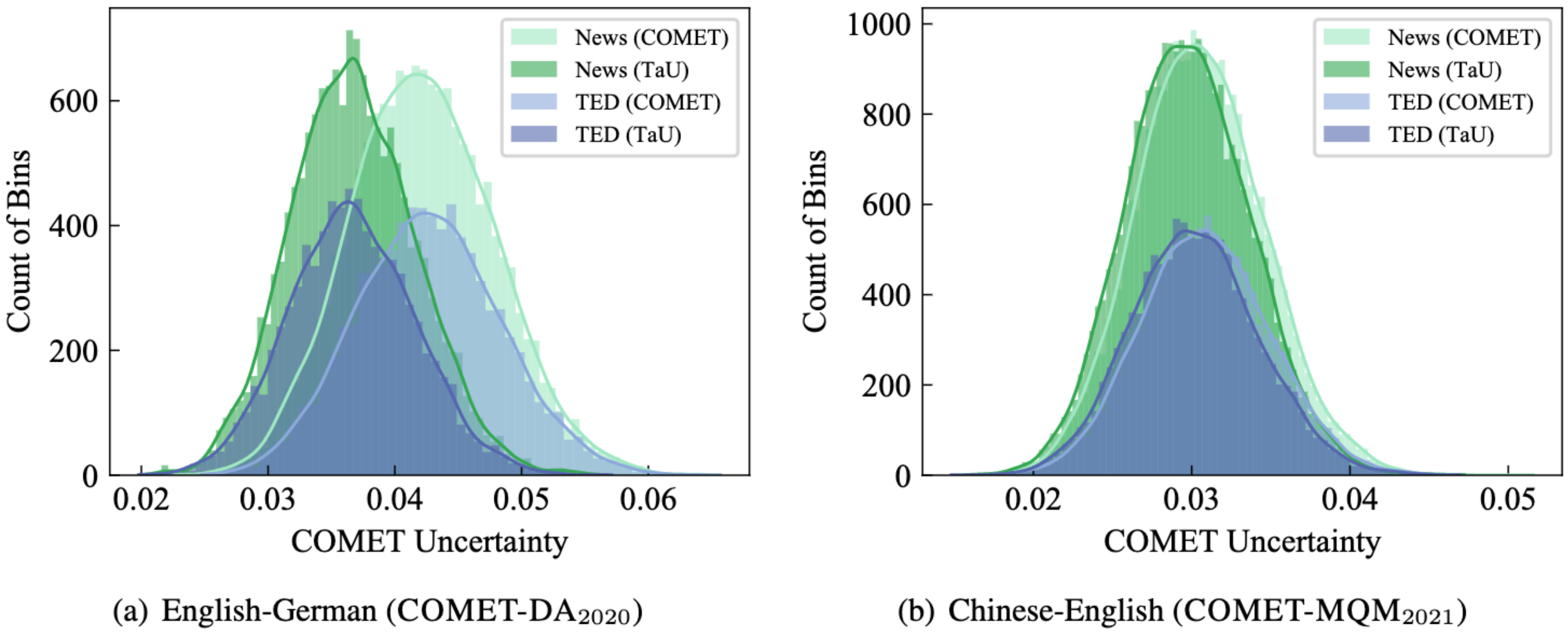
Metrics	News w/o HT			News w/ HT			TED			Avg.
	En-De	Zh-En	En-Ru	En-De	Zh-En	En-Ru	En-De	Zh-En	En-Ru	
Baselines										
TER	93.0	41.6	-4.1	7.4	-8.5	-28.9	50.6	42.1	69.7	29.2
BLEU	93.7	31.0	50.7	13.2	-15.2	-4.3	62.0	32.4	82.8	38.5
CHRF	89.8	30.2	78.3	1.7	-14.3	12.3	47.1	36.3	82.5	40.4
BERTSCORE	93.0	54.2	62.9	7.4	9.5	-12.3	50.6	30.6	83.1	42.1
COMET-DA$_{2020}$	81.4	51.1	67.6	65.8	22.1	55.6	78.8	25.1	85.9	59.3
COMET-MQM-QE$_{2021}$	71.1	52.9	63.2	79.2	61.9	68.1	69.4	-20.9	88.4	59.3
COMET-MQM$_{2021}$	77.1	62.8	65.9	72.0	33.6	68.5	81.8	26.6	84.1	63.6
Reproduced Results and Our Methods										
◇ COMET-DA$_{2020}$	81.5	51.1	67.5	58.0	26.4	56.8	78.8	25.0	85.9	59.0
+TAU	**85.7**	**53.5**	**71.0**	48.0	**27.4**	54.5	**85.9**	**28.3**	**87.3**	**60.2**
◇ COMET-MQM-QE$_{2021}$	71.2	53.0	68.8	79.2	61.9	68.1	69.4	-20.8	81.7	59.2
+TAU	62.8	**57.4**	**70.3**	72.0	**65.2**	**78.1**	**82.9**	**25.7**	80.7	**66.1**
◇ COMET-MQM$_{2021}$	77.2	62.8	65.9	69.8	48.7	69.7	81.8	26.6	84.1	65.2
+TAU	76.5	**69.2**	**67.2**	**75.4**	**67.8**	**71.5**	**87.5**	24.5	**84.9**	**69.4**

System-level Pearson correlations (%) of metrics with human scores on WMT21 Metrics Shared Task - MQM

Analysis

TaU: Effectiveness

- **Research Goal**: Reduce the uncertainty of OOD samples
- **Validity:** improve the correlation, also reduce the uncertainty



(a) English-German (COMET-DA2020)

(b) Chinese-English (COMET-MQM2021)

Uncertainty distribution of COMET baselines and corresponding models optimized by TaU

Conclusions

Conclusions

Conclusions

Conclusions

- OOD samples may hinder the application scope of neural metrics.

Conclusions

Conclusions

- OOD samples may hinder the application scope of neural metrics.
- We confirmed and also observed the uncertainty-error relationship for metric models.
- We propose a test-time adaptation method to reduce inference uncertainty.

Conclusions

Conclusions

- OOD samples may hinder the application scope of neural metrics.
- We confirmed and also observed the uncertainty-error relationship for metric models.
- We propose a test-time adaptation method to reduce inference uncertainty.
- OOD evaluation / adaptation are potential topics for the Large Language Models.

Conclusions

Conclusions

- OOD samples may hinder the application scope of neural metrics.
- We confirmed and also observed the uncertainty-error relationship for metric models.
- We propose a test-time adaptation method to reduce inference uncertainty.
- OOD evaluation / adaptation are potential topics for the Large Language Models.

Limitations

- Segment-level correlation performance is not satisfactory.
- Hyper-parameter searching is still time-consuming.
- Cannot fix the errors related to unseen knowledge.

Conclusions

Conclusions

- OOD samples may hinder the application scope of neural metrics.
- We confirmed and also observed the uncertainty-error relationship for metric models.
- We propose a test-time adaptation method to reduce inference uncertainty.
- OOD evaluation / adaptation are potential topics for the Large Language Models.

Limitations

- Segment-level correlation performance is not satisfactory.
- Hyper-parameter searching is still time-consuming.
- Cannot fix the errors related to unseen knowledge.

Poster

Thank you for listening!

Supplementary

Low-Resource Languages (In-domain)

- **Multi-domain benchmark for MT evaluation is scarce.**
- Experimental results on previous WMT-News benchmark with the same learning rate (did not tune on developmental data).

	Pl-En	Ru-En	Ta-En	Zh-En	En-Pl	En-Ru	En-Ta	En-Zh
COMET-DA	34.5	83.6	0.764	93.1	80.0	92.5	79.8	0.7
+TaU	**34.6**	**84.0**	**0.774**	**93.4**	79.0	91.6	75.3	**1.2**

System-level Pearson correlations (%) of metrics with human scores on WMT20 Metrics Shared Task (News).

Supplementary

Low-Resource Languages (In-domain)

- **Multi-domain benchmark for MT evaluation is scarce.**
- Experimental results on previous WMT-News benchmark with the same learning rate (did not tune on developmental data).

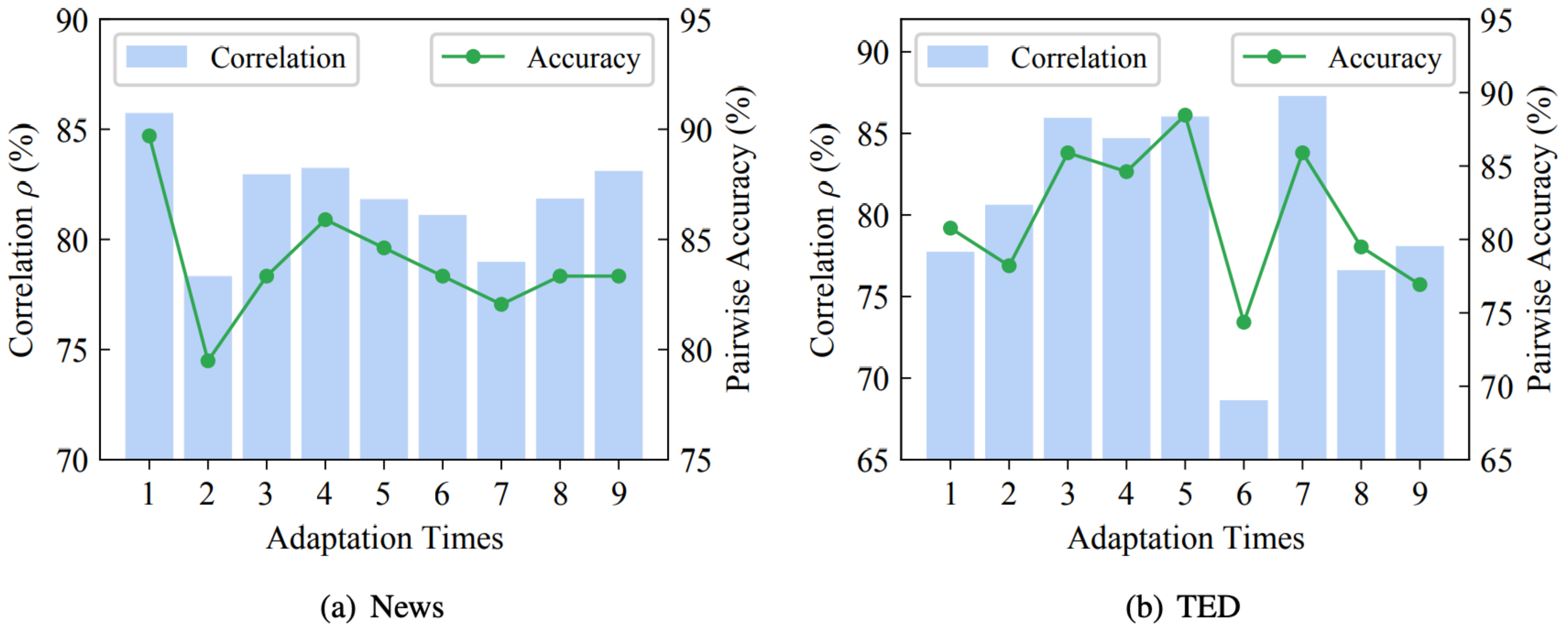
	into English				from English			
	Pl-En	**Ru-En**	**Ta-En**	**Zh-En**	**En-Pl**	**En-Ru**	**En-Ta**	**En-Zh**
COMET-DA	34.5	83.6	0.764	93.1	80.0	92.5	79.8	0.7
+TaU	**34.6**	**84.0**	**0.774**	**93.4**	79.0	91.6	75.3	**1.2**

System-level Pearson correlations (%) of metrics with human scores on WMT20 Metrics Shared Task (News).

Supplementary

Multi-turn Adaptation

- The out-of-distribution data requires more adaptation times than in-domain data, and both of them would suffer from extreme settings.



Performance of TAU with different settings of adaptation times.