Multiple Choice Learning of Low-Rank Adapters for Language Modeling

Victor Letzelter*[1,2], Hugo Malard*[2], Mathieu Fontaine[2]
Gaël Richard[2], Slim Essid[2], Andrei Bursuc[1], Patrick Pérez[3]

*Equal Contribution.

[1] Valeo.ai
[2] LTCI, Telecom Paris, Institut Polytechnique de Paris
[3] Kyutai

Ambiguous tasks?

Image Captioning.

Ambiguous tasks?

{Pedestrians crossing a busy downtown street.}

Image Captioning.

Ambiguous tasks?

{Pedestrians crossing a busy downtown street.}

{An ambulance drives through an intersection with its lights on.}

Image Captioning.

Ambiguous tasks?

{Pedestrians crossing a busy downtown street.}

{An ambulance drives through an intersection with its lights on.}

{Taxis and cars move through a crowded city street.}

Image Captioning.

Ambiguous tasks?

{Pedestrians crossing a busy downtown street.}

{An ambulance drives through an intersection with its lights on.}

{Taxis and cars move through a crowded city street.}

{A foggy afternoon in a large city with tall buildings lining the street.}

Image Captioning.

Ambiguous tasks?

{Pedestrians crossing a busy downtown street.}

{An ambulance drives through an intersection with its lights on.}

{Taxis and cars move through a crowded city street.}

{A foggy afternoon in a large city with tall buildings lining the street.}

Image Captioning.

☞How to capture this ambiguity in Language models ?

(Causal) Language Modeling

- Sequence of tokens $x = (x_t)_{t=1}^{T} \in \mathcal{V}^T$
- Context vector embeddings $c = (c_t)_{t=1}^{\tau} \in \mathbb{R}^{\tau \times d}$

Training. "Next-token-prediction" with "teacher-forcing"

$$\mathcal{L}(\theta) = \mathbb{E}_{c,x}\left[-\sum_{t=1}^{T} \log p_\theta\left(x_t \mid x_{<t}, c\right)\right]$$

(Causal) Language Modeling

Decoding. Inference is performed in an autoregressive manner.

1. Compute $p_\theta(x_1 \mid c)$ and select $\hat{x}_1$.
2. For $t \geq 2$, compute $p_\theta(x_t \mid \hat{x}_{<t}, c)$ and select $\hat{x}_t$.

How to select the tokens ?

- Maximum A Posteriori (greedy, beam search[1], diverse beam search[2]).
- Sampling (top-k[3], nucleus[4]).

---

[1] B. T. Lowerre (1976). *The harpy speech recognition system.* CMU.

[2] A. Vijayakumar et al. (2018). "Diverse beam search for improved description of complex scenes". In: *AAAI*.

[3] A. Fan et al. (2018). "Hierarchical neural story generation". In: *ACL*.

[4] A. Holtzman et al. (2020). "The curious case of neural text degeneration". In: *ICLR*.

Motivations with a synthetic data example

Synthetic data example with Markov Chain generated data.

- Transition matrix $P_{i,j} = p(x_{t+1} = j \mid x_t = i)$
- GPT-like[1] model trained on such sequences

☞ Does the transformer recover the true transition matrix?

[1] A. Radford et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog*.

Motivations with a synthetic data example

Synthetic data example with Markov Chain generated data.

- Transition matrix $P_{i,j} = p(x_{t+1} = j \mid x_t = i)$
- GPT-like[1] model trained on such sequences

☞ Does the transformer recover the true transition matrix?

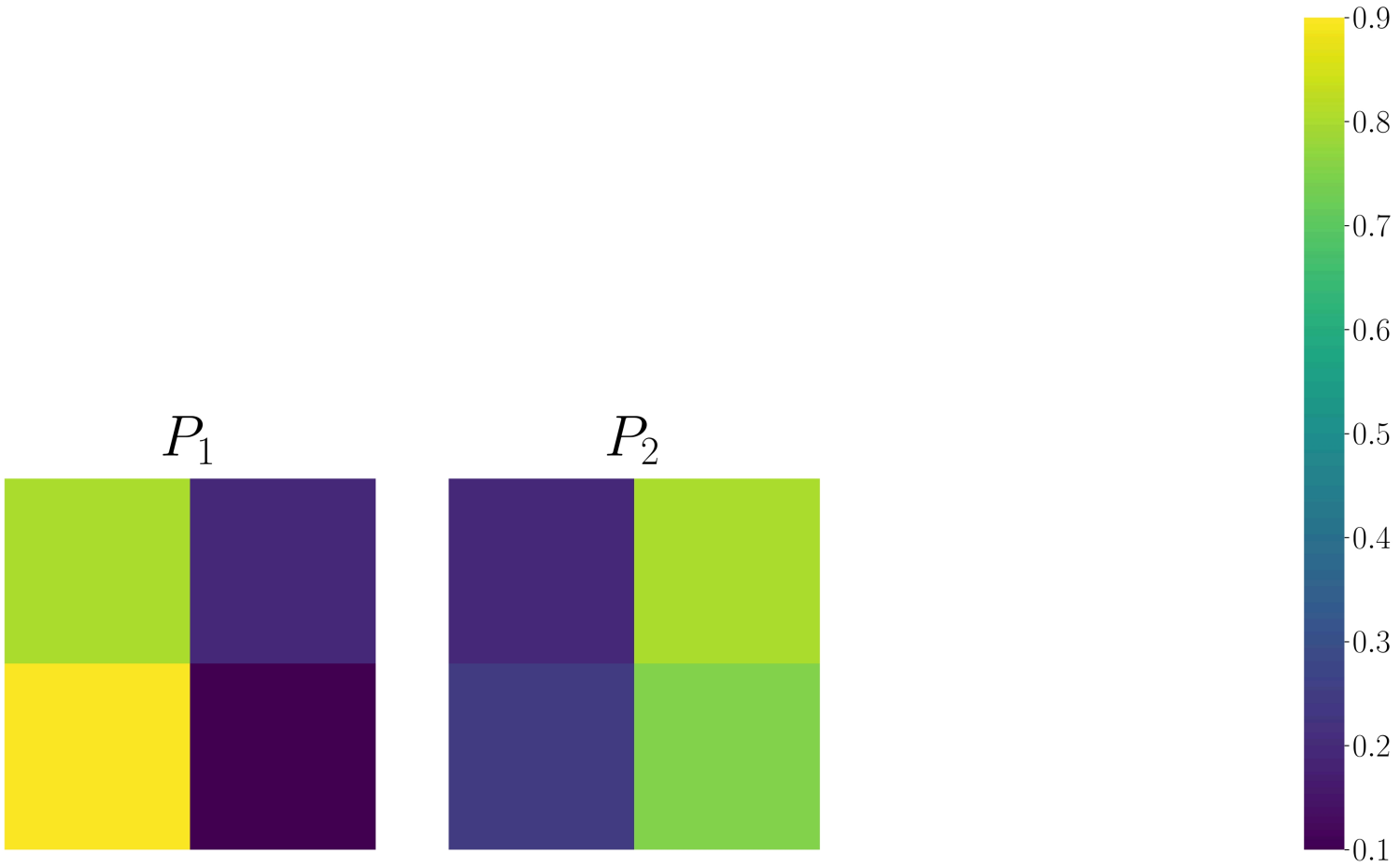
[1] A. Radford et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog*.

Motivations with synthetic data

Let the data be sampled from a **mixture** of Markov Chains: an index is sampled uniformly $k \sim \mathcal{U}\{1, 2\}$, then a sequence using P_k ($\mathcal{V} = \{1, 2\}$)

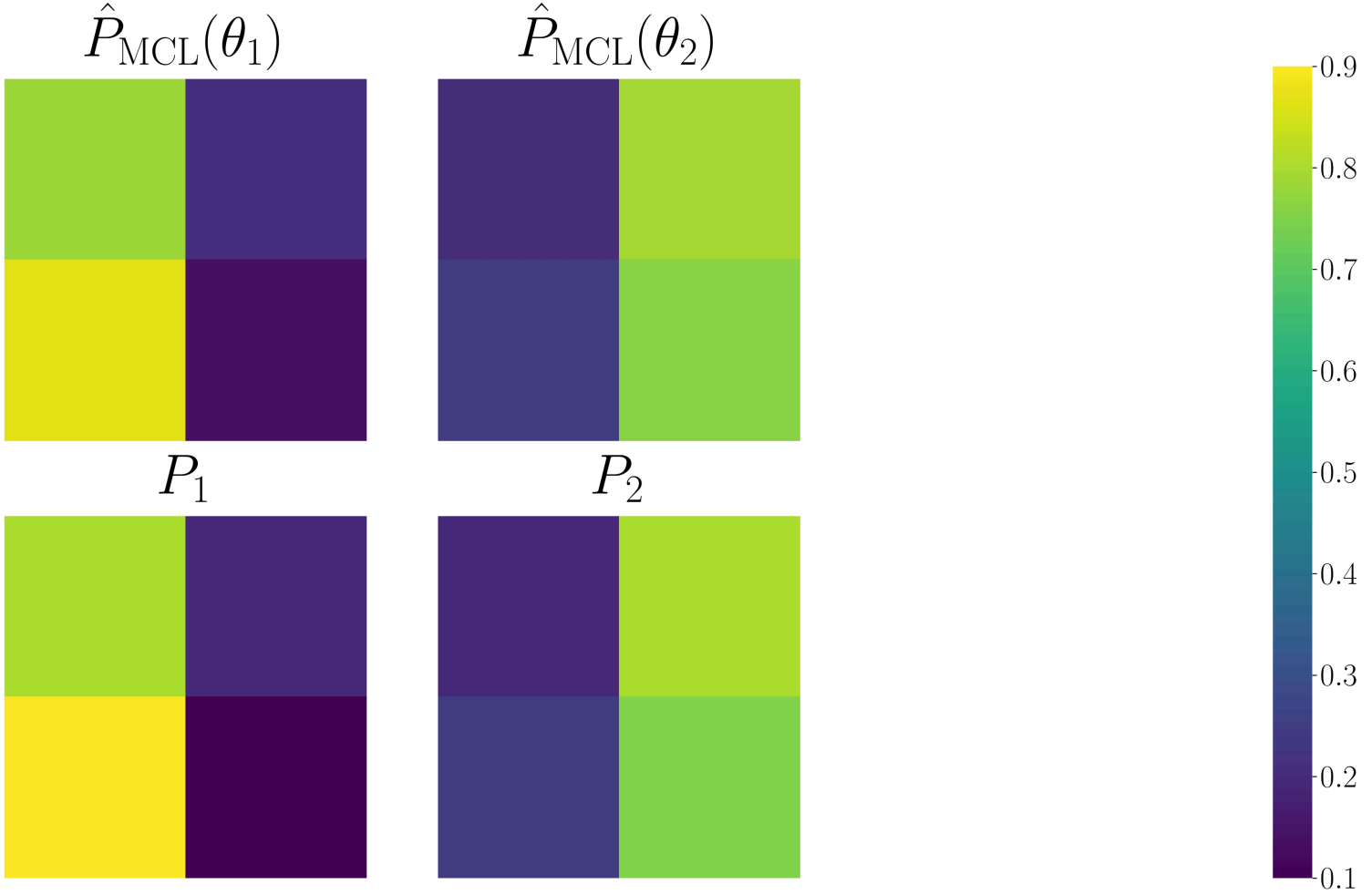
Motivations with synthetic data

Let the data be sampled from a **mixture** of Markov Chains: an index is sampled uniformly $k \sim \mathcal{U}\{1, 2\}$, then a sequence using P_k ($\mathcal{V} = \{1, 2\}$)



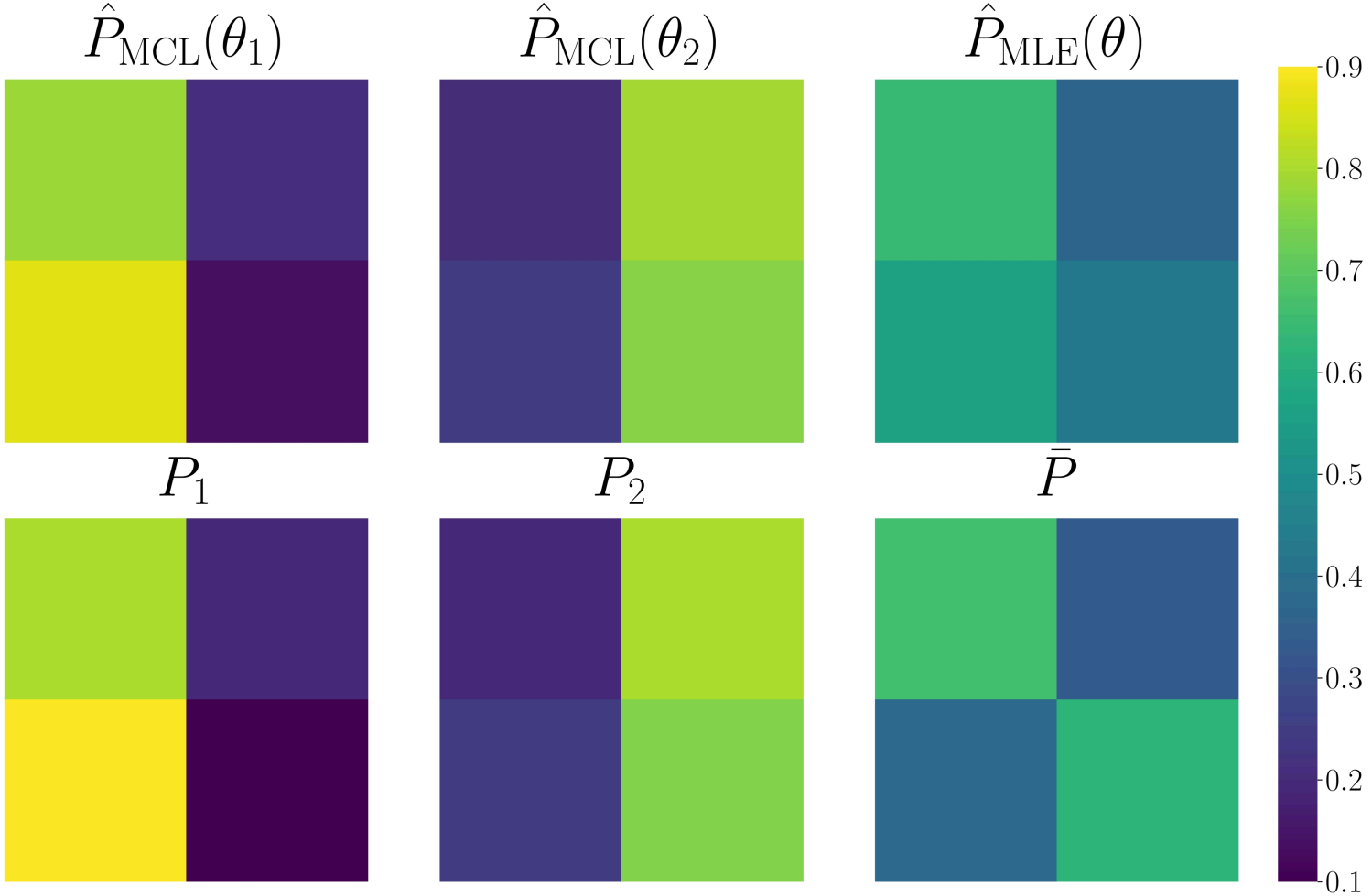
Motivations with synthetic data

Let the data be sampled from a **mixture** of Markov Chains: an index is sampled uniformly $k \sim \mathcal{U}\{1, 2\}$, then a sequence using P_k ($\mathcal{V} = \{1, 2\}$)



Motivations with synthetic data

Let the data be sampled from a **mixture** of Markov Chains: an index is sampled uniformly $k \sim \mathcal{U}\{1, 2\}$, then a sequence using P_k ($\mathcal{V} = \{1, 2\}$)



Multiple choice learning for language modeling

Consider several models $\theta_1, \ldots, \theta_K$.

Multiple choice learning for language modeling

Consider several models $\theta_1, \ldots, \theta_K$.

1. For each training sample (c, x): Compute $p(x \mid c; \theta_k)$ for $k \in \{1, \ldots, K\}$, and choose

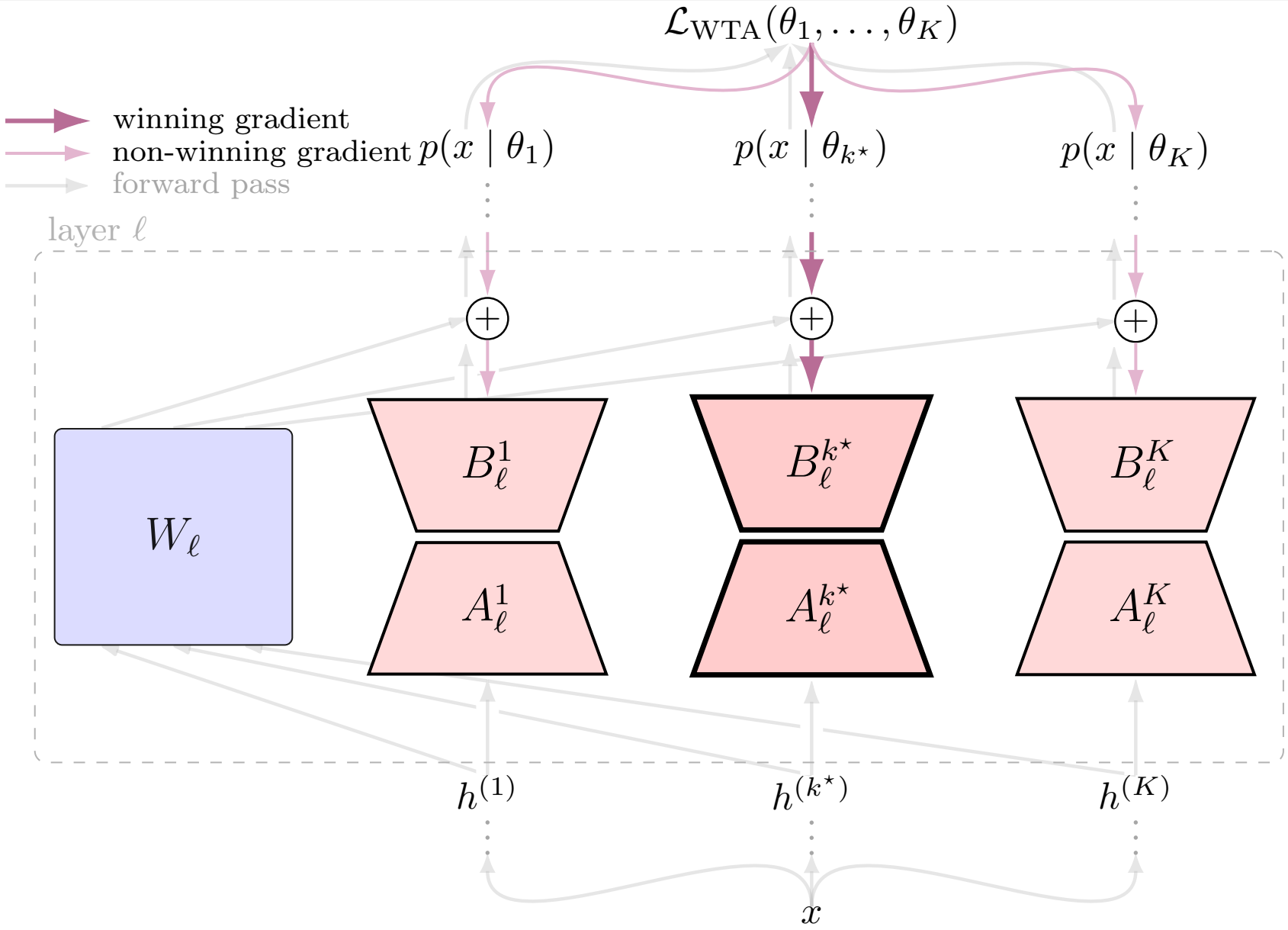
$$k^\star(x, c) = \text{argmax}_k \; p(x \mid c; \theta_k) \; .$$

2. Compute the *(relaxed)* winner-takes-all loss:

$$\mathcal{L}^{\text{WTA}}(\theta_1, \ldots, \theta_K) = -\mathbb{E}_{c,x}\Big[\max_{k=1,\ldots,K} \log p(x \mid c; \theta_k)\Big] \; ,$$

where $\log p(x \mid c; \theta_k) = \sum_{t=1}^{T} \log p(x_t \mid x_{<t}, c; \theta_k)$.

MCL with Multiple Low Rank Adapters



Experiments

Audio Captioning. Qwen-2-Audio.[1]
Image Captioning. LLaVA[2]
Machine Translation. ALMA[3]
Metrics. Quality (Oracle) / Diversity.

[1] Y. Chu et al. (2024). "Qwen2-audio technical report". In: *arXiv*.
[2] H. Liu et al. (2023). "Visual instruction tuning". In: *NeurIPS*.
[3] H. Xu et al. (2024). "A paradigm shift in machine translation: Boosting translation performance of large language models". In: *ICLR*.

Baselines

Training:

- Next-token-prediction (LoRA-MLE)
- Mixture of LoRA-Experts (LoRA-MoE)
- Test-time augmentation (TTA)

Decoding: Greedy, Beam Search, Diverse Beam Search.

Each baseline generate K hypotheses at inference.

Baselines

Training:

- Next-token-prediction (LoRA-MLE)
- Mixture of LoRA-Experts (LoRA-MoE)
- Test-time augmentation (TTA)

Decoding: Greedy, Beam Search, Diverse Beam Search.

Each baseline generate K hypotheses at inference.

☞Our method achieves *natively* an excellent trade-off between quality and diversity.

Hypotheses specialization in bilingual image description

LoRA-MLE.
{A bottle of Cerveza is on a table.}
{Une bouteille de vin de cidre de cidre de cidre [...]}
LoRA-MCL.
{A bottle of beer with a label that says "Sel Maguet"}
{Une bouteille de vin est étiquetée avec le mot « Maguay ».}

LoRA-MLE.
{A book titled Papa Told Me is being held by a person.}
{A book called Papa told me is being held by a person.}
LoRA-MCL.
{A book titled Papa Told Me is being held by a person}
{Un livre papier intitulé Papa Told Me.}

Observing specialization in bilingual image description. Quantitative (*Left*) and Qualitative *(Right)* analysis for `LoRA-MLE` and `LoRA-MCL`.

Hypotheses specialization in bilingual image description

LoRA-MLE.
{A bottle of Cerveza is on a table.}
{Une bouteille de vin de cidre de cidre de cidre [...]}

LoRA-MCL.
{A bottle of beer with a label that says "Sel Maguet"}
{Une bouteille de vin est étiquetée avec le mot « Maguay ».}

LoRA-MLE.
{A book titled Papa Told Me is being held by a person.}
{A book called Papa told me is being held by a person.}

LoRA-MCL.
{A book titled Papa Told Me is being held by a person}
{Un livre papier intitulé Papa Told Me.}

Observing specialization in bilingual image description. Quantitative (*Left*) and Qualitative *(Right)* analysis for `LoRA-MLE` and `LoRA-MCL`.

Hypotheses specialization in bilingual image description

LoRA-MLE.
{A bottle of Cerveza is on a table.}
{Une bouteille de vin de cidre de cidre de cidre [...]}
LoRA-MCL.
{A bottle of beer with a label that says "Sel Maguet"}
{Une bouteille de vin est étiquetée avec le mot « Maguay ».}

LoRA-MLE.
{A book titled Papa Told Me is being held by a person.}
{A book called Papa told me is being held by a person.}
LoRA-MCL.
{A book titled Papa Told Me is being held by a person}
{Un livre papier intitulé Papa Told Me.}

Observing specialization in bilingual image description. Quantitative (*Left*) and Qualitative *(Right)* analysis for `LoRA-MLE` and `LoRA-MCL`.

Limitations

- Specialization ?

Limitations

- Specialization ?
- MCL during pretraining ?

Thank you!