Masked Prediction Task
a parameter identifiability view

Bingbin Liu, Daniel Hsu, Pradeep Ravikumar, Andrej Risteski

NEURAL INFORMATION PROCESSING SYSTEMS

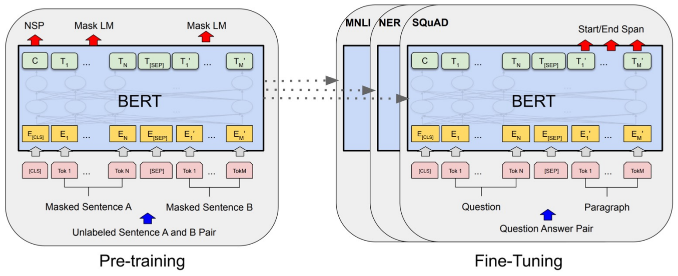
Carnegie Mellon University

COLUMBIA UNIVERSITY

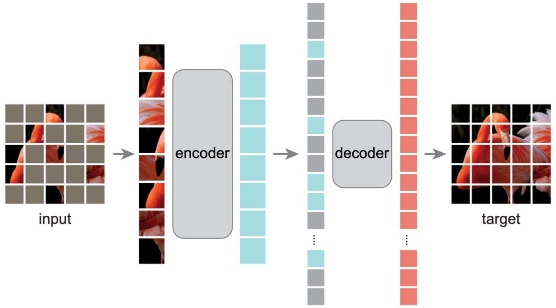
Masked Prediction

Task: train a model by predicting the missing part of the input.

- Empirically successful: good features for downstream tasks.
  - NLP: Word2vec [Milokov et al. 13], BERT [Devlin et al. 16]
  - Vision: Context Encoder [Pathak et al.16], MAE [He et al. 21]
- Theoretically studied [Lee et al. 21, Wei et al. 21]



BERT [Devlin et al. 16]



MAE [He et al. 21]

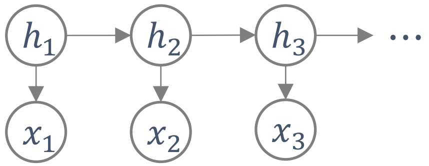
Evaluation metrics

- Downstream performance – unclear which downstream tasks to use.
- **Parameter identifiability**: a natural quality measure; e.g. common in graphical models.

Masked Prediction – parameter identifiability

Setup: sequential data generated by a latent-variable model with known parametric form.

HMM: discrete latents $\{h_t\}$, discrete or continuous observables $\{x_t\}$.



$$P(h_{t+1} = i | h_t = j) = T_{ij}, \quad \begin{cases} P(x_t = i | h_t = j) = O_{ij}, & \text{(discrete case)} \\ P(x_t = x | h_t = j) \propto \exp\left(-\frac{||x - \mu_i||^2}{2}\right). & \text{(continuous case)} \end{cases}$$

(parameters)

Identifiability: *when* can we *read off* the parameters from an optimal predictor with the *correct form*?

Identifiability

Are the HMM parameters identifiable from an optimal predictor?

A masked prediction task is identifiable, if for two HMMs with (O, T) and $(\tilde{O}, \tilde{T})$,

matching the predictor means $O = \tilde{O}\Pi, T = \Pi^\top \tilde{T} \Pi$ for some permutation matrix Π .

e.g. discrete case, pairwise prediction:

$$f^{2|1}(x) = \mathbb{E}[x_2 | x_1 = x] = \sum_{i \in [k]} \sum_{j \in [k]} O_i T_{ij} \overbrace{\frac{O_{x,j}}{\sum_{l \in [k]} O_{x,l}}}^{[\phi(x)]_j \text{ (posterior distr on } h \text{ given } x)}$$

$$= \sum_{i \in [k]} \sum_{j \in [k]} \tilde{O}_i \tilde{T}_{ij} \frac{\tilde{O}_{x,j}}{\sum_{l \in [k]} \tilde{O}_{x,l}} = \tilde{f}^{2|1}(x)$$

Results overview

Are the HMM parameters identifiable from an optimal predictor – Task & model dependent.

Identifiable: matching the predictor → $O = \tilde{O}\Pi, T = \Pi^\top \tilde{T} \Pi$ for some permutation Π .

	$x_i \mid x_j$	$x_i \otimes x_j \mid x_k$
HMM (discrete)	❌ rotation problem (matrix)	✅ Tensor decomposition
G-HMM (cond Gaussian)	more informative posterior ✅	✅

Discrete case

transition matrix $T \in \mathbb{R}^{k \times k}$
emission matrix $O \in \mathbb{R}^{d \times k}$

Pairwise prediction: non-identifiable due to rotation.

(Thm 4) There exist parameters $\tilde{O}, O$ such that $\tilde{O} \neq O$ (up to permutation), yet the predictors for $x_2|x_1, x_1|x_2$, $x_3|x_1, x_1|x_3$ are the same.

Triplet prediction: identifiable due to the uniqueness of tensor decomposition (**Kruskal's theorem**).

(Thm 5) O, T are identifiable from the predictor for $x_{t_2} \otimes x_{t_3} \mid x_{t_1}$, for t_1, t_2, t_3 being any permutation of $\{1,2,3\}$.

Continuous case (conditionally Gaussian)

transition matrix $T \in \mathbb{R}^{k \times k}$
means $M := [\mu_1, \dots, \mu_k] \in \mathbb{R}^{d \times k}$

Pairwise prediction: identifiable:

(Thm 3) M, T are identifiable from the predictor for $x_2 \mid x_1$.

Intuition: the nonlinearity gives a more informative posterior ϕ (over h, given x).

(Lem 1) For 2 parameters $M, \widetilde{M}$, if $\phi = \tilde{\phi}$, then

- $\widetilde{M} = M := [\mu_1, \dots, \mu_k]$,
- or $\widetilde{M} = HM$, where H is a Householder transformation.

Contributions

Q: *when* can we *read off* the parameters from an optimal predictor with the *correct form*?

A: highly specific to the task & model:

	$x_i \mid x_j$	$x_i \otimes x_j \mid x_k$
HMM (discrete)	❌ rotation problem (matrix)	✅ Tensor decomposition
G-HMM (cond Gaussian)	more informative posterior ✅	✅

- Open: condition on more tokens? robustness / sample complexity? More general families?