



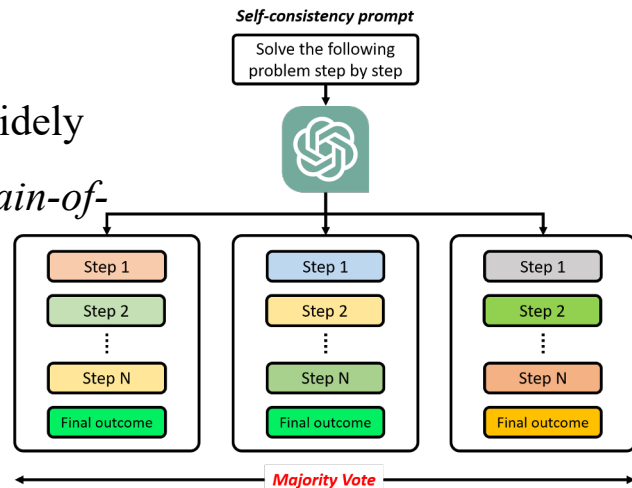
# DC-MBR: Distributional Cooling for Minimum Bayesian Risk Decoding

Jianhao Yan, Jin Xu, Fandong Meng, Jie Zhou, Yue Zhang

*elliottyan37@gmail.com*

# Research Background

- Minimum Bayesian Risk Decoding (MBR) emerges as a promising decoding technique to *best leverage the potential of diverse decoding paths*.
- The core idea is *majority voting*, which takes multiple hypotheses and combines them together.
- It is originally introduced in machine translation, but is now widely adopted in large language models, e.g., *self-consistency for chain-of-thought prompting*.



- Label smoothing is shown to be beneficial for various tasks, by improving the generalization.

$$y_{ls} = (1 - \lambda)y_{hot} + \frac{\lambda}{|V|}$$

Network	Top-1 Error	Top-5 Error	Cost Bn Ops
GoogLeNet [20]	29%	9.2%	1.5
BN-GoogLeNet	26.8%	-	<b>1.5</b>
BN-Inception [7]	25.2%	7.8	2.0
Inception-v2	23.4%	-	3.8
Inception-v2 RMSProp	23.1%	6.3	3.8
Inception-v2 Label Smoothing	22.8%	6.1	3.8
Inception-v2 Factorized 7 × 7	21.6%	5.8	4.8
Inception-v2 BN-auxiliary	<b>21.2%</b>	<b>5.6%</b>	4.8

ImageNet Performance, Table 3 of Szegedy et.al, 2016.

Model	RNMT+	Trans. Big
Baseline	41.00	40.73
- Label Smoothing	40.33	40.49
- Multi-head Attention	40.44	39.83
- Layer Norm.	*	*
- Sync. Training	39.68	*

Table 4: Ablation results of RNMT+ and the Transformer Big model on WMT'14 En → Fr. We report average BLEU scores on the test set. An asterisk '\*' indicates an unstable training run (training halts due to non-finite elements).

MT Performance with beam search, Table 4 of Chen et.al, 2018.

1. Szegedy, Christian, et al. "Rethinking the Inception Architecture for Computer Vision." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 2016. IEEE, 2016.

2. Chen, Mia Xu, et al. "The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2018.

- **However, models trained with label smoothing works poorly with MBR!!**

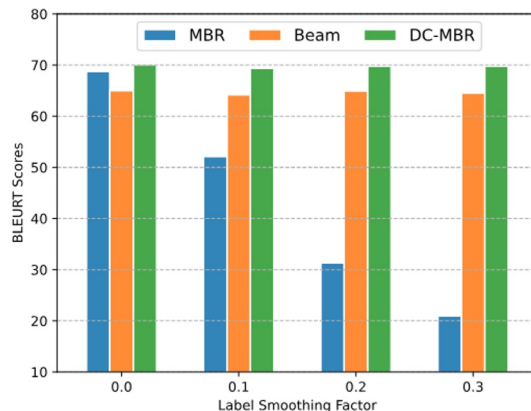
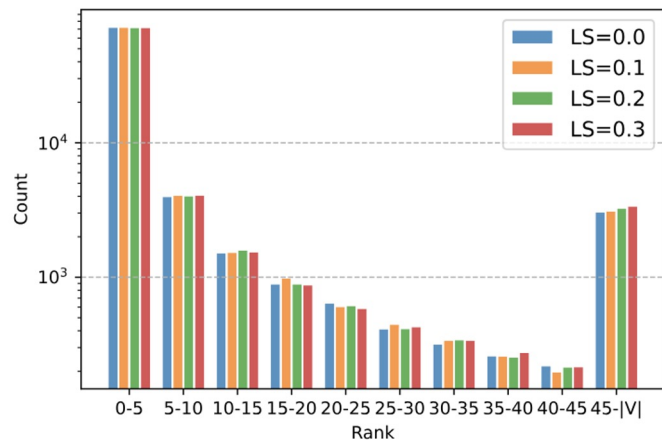


Figure 1: Translation quality against label smoothing factors. As the factor of label smoothing increases, beam search retains its performance while that of MBR drops drastically.

## - Why?



Label smoothing only introduces slight changes in token-level distribution, as intended.

Figure 2: Ranking statistics for tokens in the ground-truth sentence within the token-level distribution  $P(y_t | y_{<t}, x)$ .

## - Why?

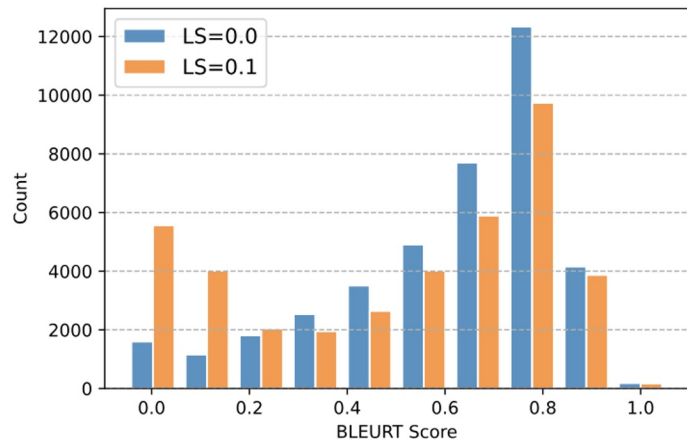


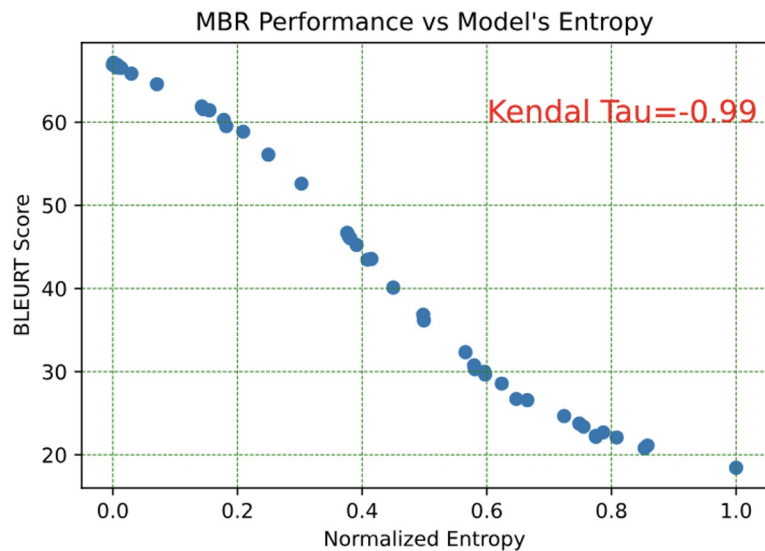
Figure 3: Translation quality statistics for sequences within the top-20 candidates of the sequence-level distribution  $P(y|x)$ .

However, in sequence-level, it causes much probability placed on low-quality candidates!!!!

$$\lambda = 0.1, L = 30 \Rightarrow P(\text{golden}) = 0.9^{30} = 0.4$$

Auto-regressive modeling and label smoothing together cause this phenomenon.

- Entropy correlates perfectly with the performance.



- Thus, we provide a principled post-training approach to mitigate this problem.
  - By theoretical derivation, temperature rescaling on MBR candidate generation can *reverse* the operation of label smoothing.

$$P(y_t | y_{<t}, x; \theta, T) = \frac{\exp \frac{o_t}{T}}{\sum_j |V| \exp \frac{o_j}{T}},$$

$$\hat{\mu}_u(h; x, \theta) := \frac{1}{N} \sum_r \mathcal{Y}_{\text{model}}^{T_r} u(h, r),$$

$$\mathcal{Y}_{\text{model}}^T \sim \prod_t P(y_t | y_{<t}, x; \theta, T).$$

$$\hat{y}^{\text{DC-MBR}} = \operatorname{argmax}_{h \in \mathcal{Y}_{\text{model}}^{T_h}} \hat{\mu}_u(h; x, \theta).$$

Models	Models	BS	MBR	Ours	$\Delta$
$N = 10$	Transformer	65.0	63.8	68.8	+5.0
	+ LS 0.1	64.2	41.0	67.9	+26.9
	+ LS 0.2	64.9	24.0	68.3	+44.3
	+ LS 0.3	64.6	17.1	68.3	+51.2
$N = 50$	Transformer	65.0	68.7	70.1	+1.4
	+ LS 0.1	64.2	52.1	69.4	+17.3
	+ LS 0.2	64.9	31.3	69.8	+38.5
	+ LS 0.3	64.6	21.0	69.8	+48.8

Table 1: BLEURT scores for En-De. Gray: Models perform poorly with original MBR. We investigate two settings: Low cost,  $N=10$ , 100 BLEURT calls per sentence; High cost,  $N=50$ , 2500 BLEURT calls per sentence. Our results are significantly better than “MBR” ( $p < 0.01$ ).

Extensive experiments show that our approach can effectively restore the results for label smoothing models!

- We first notice a curious phenomenon that label smoothing, which performs well for various tasks and beam search, performs poorly with minimum bayesian risk decoding.
- We analyze the effect from both token-level and sequence-level distribution, revealing a small perturbation in token-level results in a huge disparity on sequence-level.
- We propose a principled post-training approach called *DC-MBR* to effectively mitigate this issue, and empirically prove its effectiveness.

