Stochastic Transformer Networks with Linear Competing Units: Application to end-to-end SL Translation

-

Supplementary Material

Andreas Voskou*[1], Konstantinos P. Panousis[1], Dimitrios Kosmopoulos[2],
Dimitris N. Metaxas[3], and Sotirios Chatzis[1]

[1]Cyprus University of Technology, [2]University of Patras, [3]Rutgers University, New Jersey

1. Evidence Lower-Bound expression

The expression of the model Evidence Lower-Bound (ELBO) takes the form:

$$\mathcal{L}(q) = -\mathbb{E}_{q(\cdot)}[CE] - (\mathrm{KL}[q(\boldsymbol{\xi})||p(\boldsymbol{\xi})] + \mathrm{KL}[q(\boldsymbol{w})||p(\boldsymbol{w})]) \tag{1}$$

In this expression, CE represents the categorical cross-entropy loss obtained from the translation process viewed as a classification (selection of one-out-of-many) task. In the context of our model, this is a (stochastic) function of the model latent variables, $\boldsymbol{\xi}$ and $\boldsymbol{w}$. We have

$$\mathbb{E}_{q(\cdot)}[CE] = -\mathbb{E}_{q(\cdot)}\big[\log p(\mathcal{D}|\{\boldsymbol{w}, \boldsymbol{\xi}\})\big] \tag{2}$$

On the other hand, KL stands for the Kullback-Leibler divergences pertaining to the model latent variables, i.e. the winner unit indicator latent variables, $\boldsymbol{\xi}$, and the connection weights, $\boldsymbol{w}$. We have:

$$\mathrm{KL}[q(\boldsymbol{\xi})||p(\boldsymbol{\xi})] = \sum_{\forall \boldsymbol{\xi}} \beta_\xi \sum_{i=1}^{U} q(\xi_i) \log\left(q(\xi_i) U\right) \tag{3}$$

$$\mathrm{KL}[q(\boldsymbol{w})||p(\boldsymbol{w})] = \sum_{\forall \boldsymbol{w}} \beta_w [\frac{\boldsymbol{\mu}^2 + \boldsymbol{\sigma}^2}{2} - \log \boldsymbol{\sigma} - \frac{1}{2}] \tag{4}$$

In these expressions, the factors β_ξ and β_w are heuristic hyperparameters needed to scale appropriately the KL contribution of each layer.

As we observe, the model ELBO entails posterior expectations over the network latent variables, $\boldsymbol{\xi}$ and $\boldsymbol{w}$. These cannot be computed analytically. Therefore, we approximate them via Monte-Carlo (MC) samples. To allow for low-variance ELBO gradients, we resort to the reparameterization trick.

*ai.voskou@edu.cut.ac.cy

Specifically, we express the samples of the connection weights, $\boldsymbol{w}$, in the form:

$$\boldsymbol{w} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{5}$$

For the winner latent indicator variables, $\boldsymbol{\xi}$, the Gumbel-Softmax relaxation trick [1] yields the following expression for the drawn samples:

$$\begin{aligned} \boldsymbol{\xi} &= \frac{\exp\left((\log \boldsymbol{\eta} + \boldsymbol{g})/T\right)}{\sum_{i=1}^{U} exp((\log \eta_i + g_i)/T)} \\ \boldsymbol{g} &= -log(-\log \boldsymbol{z}),\ \boldsymbol{z} \sim \mathrm{U}(\boldsymbol{0}, \boldsymbol{1}) \end{aligned} \tag{6}$$

where $\boldsymbol{\eta} = q(\boldsymbol{\xi})$, and T is a positive temperature hyperparameter.

2. Effect of sample size S at inference time

As we explained in the main text, inference is performed through Bayesian averaging with a sample size of $S = 4$. It is useful to examine how this selection affects BLEU-4 scores. Thus, we repeat our experiments with the (2-2) configuration for different S values, and report the outcomes in Table 1. In addition, we exploit this opportunity to examine how the model behaves if we do not sample winners and connection weights; instead we pick the unit the yields the maximum posterior probability $q(\boldsymbol{\xi})$, and use the connection weight means, $\boldsymbol{\mu}$, to perform feedforward computations ("Deterministic" scenario).

We observe that the increased sample size indeed improves model stability and performance. More specifically, a large sample size $S = 16$ can slightly increase BLEU-4, while $S = 1$ produces a much lower performance than the proposed $S = 4$ size. However, large sample size also increases the model's inference time, and it may be impractical, especially for $S >> 16$. Moreover, we can see

Table 1. Sample size effect (BLEU-4 scores).

Sample	32 bit		Reduced	
	Dev	Test	Dev	Test
Deterministic (1)	22.99	22.83	22.21	22.31
1	23.45	23.05	22.66	23.03
4	23.23	23.64	23.09	23.52
16	23.34	**23.95**	23.14	23.39

that a deterministic version of our model is clearly inferior to stochastic operation, even if we draw just one sample, $S = 1$. This corroborates our theoretical intuitions regarding the doubly stochastic formulation of our model.

3. Computational Complexity.

We emphasize we do MC (simple drawing of i.i.d. samples; augmented with the reparameterization trick during training for low-variance gradients), and *not* Markov chain Monte-Carlo (MCMC), which of course is extremely inefficient. In table 2, we compare the computation times of our approach with the baseline. The differences in training time are due to the increased number of trainable parameters, namely the trainable variances of the Gaussian weights, and *not* due to Bayes. We emphasize that these trainable variances are used for proper weight compression, and are eventually discarded after training. Crucially, *convergence took almost the same number of epochs* with the baseline. Figure 1 illustrates both the learning curves and proves the latter statement and the smooth convergence of our network. The computational efficiency of our approach becomes even more apparent when we check inference times, which are comparable with the baseline Transformer when using $S = 4$ samples (that's the best trade-off between accuracy and complexity).

Table 2. Computation Time on a signle Quadro P5000 16GB.

	Baseline	Ours
Training (per batch)	0.2s	0.5s
Inference (single video)	0.04s	0.05s

Note also that, contrary to large Transformers that dominate NLP, our Transformer-type model is *small*, as we describe in the paper; just 2 layers! Crucially, the times reported in Table 2 were obtained on a *single* Quadro P5000 16GB GPU, which is inexpensive. Even more importantly, the compressed version of our network imposes a memory footprint of less than 1 GB. Thus, loading the trained model is feasible even on a commercial smartphone device.

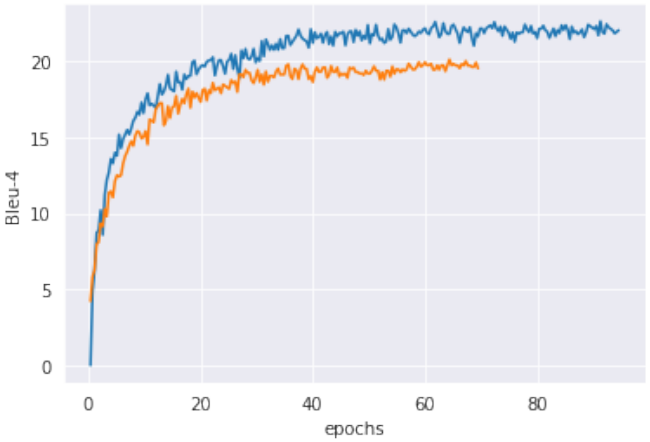

Figure 1. Convergence curves. Blue: proposed, Orange: baseline.

4. Translation examples

In Table 3, we present an extended sample of our outputs. We include both the German original and the corresponding translation in English.

References

[1] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.

Table 3. Output sentences with the corresponding translations in Engish: Reference (R), single model (S), and ensemble (E).

R: vor allem im süden können die gewitter unwetterartig ausfallen . (*especially in the south, thunderstorms can be severe.*) **S:** vor allem im süden und süden muss mit sturmböen gerechnet werden . (*especially in the south and the south, gusts of wind must be expected.*) **E:** vor allem im süden und süden muss mit unwetterartigen gewittern gerechnet werden . (*especially in the south and south of the country, thunderstorms are to be expected.*)
R: am donnerstag vor allem in der nordwesthälfte hier und da sturmböen . (*on thursday, squalls here and there, especially in the northwestern half.*) **S:** am donnerstag vor allem im norden und westen teilweise kräftig . (*on thursday especially in the north and west partly strong.*) **E:** am donnerstag vor allem im norden und westen teilweise frischer wind . (*on thursday partly fresh wind, especially in the north and west.*)
R: morgen bei dauernebel im süden nur zwei am niederrhein bis elf grad . (*tomorrow with permanent fog in the south only two on the lower Rhine up to eleven degrees.*) **S:** morgen im süden bis zweiundzwanzig am oberrhein bis elf grad . (*tomorrow in the south up to twenty-two degrees on the upper rhine up to eleven degrees .*) **E:** morgen im süden noch zwei bis einundzwanzig am rhein bis elf grad . (*tomorrow in the south still two to twenty-one on the rhine to eleven degrees.*)
R: im erzgebirge morgen dreizehn sonst fünfzehn bis zweiundzwanzig grad . (*in the ore mountains tomorrow thirteen or fifteen to twenty-two degrees .*) **S:** an der saale morgen dreizehn in der eifel bis zweiundzwanzig grad . (*on the Saale tomorrow thirteen in the Eifel to twenty-two degrees .*) **E:** an der see morgen dreizehn sonst zwei bis zweiundzwanzig grad . (*at the lake tomorrow thirteen otherwise two to twenty-two degrees .*)
R: in der neuen woche kühlt es dann bei wechselhaftem wetter deutlich ab . (*in the new week it cools down considerably with changeable weather.*) **S:** in der neuen woche wechselhaft und deutlich kühler . (*in the new week changeable and clearly cooler.*) **E:** in der neuen woche unbeständig und noch kühler . (*in the new week unstable and even colder.*)
R: und so warm mit zwanzig grad wird es nicht mehr in den nächsten tagen . (*and it won't be that warm with twenty degrees in the next few days.*) **S:** und die zwanzig geht es in den nächsten tagen . (*and the twenty it goes in the next few days.*) **E:** und die zwanzig grad in den nächsten tagen . (*and the twenty degrees in the next few days.*)
R: mit föhnunterstützung klettern die temperaturen am alpenrand bis zehn grad . (*with foehn support, temperatures climb to ten degrees on the alpine fringe.*) **S:** es ist ein bisschen kühler temperaturen so ein bisschen verhalten zehn bis zehn grad . (*it is a bit cooler temperatures so a bit restrained ten to ten degrees.*) **E:** es kann ein bisschen geben bei temperaturen am alpenrand da sinken die temperaturen bis zehn grad . (*there can be a bit of a dip in temperatures at the edge of the alps, where temperatures drop to ten degrees.*)
R: mit der leicht kühleren luft dann bis vierundzwanzig grad in der nordhälfte . (*with the slightly cooler air then up to twenty-four degrees in the north half .*) **S:** in diesem bild von vierundzwanzig bis vierundzwanzig grad in der lausitz . (*in this picture from twenty-four to twenty-four degrees in Lusatia .*) **E:** in der zweiten nachthälfte die vierundzwanzig bis vierundzwanzig grad in der lausitz . (*in the second half of the night the twenty-four to twenty-four degrees in Lusatia .*)
R: auch in den folgenden tagen ändert sich an diesem wechselhaften wetter wenig . (*even in the following days there is little change in this changeable weather.*) **S:** in den folgenden tagen setzt sich das wechselhafte witterung fort . (*in the following days the changeable weather continues.*) **E:** in den folgenden tagen setzt sich das wechselhafte wetter fort . (*the weather continues to be changeable in the following days.*)