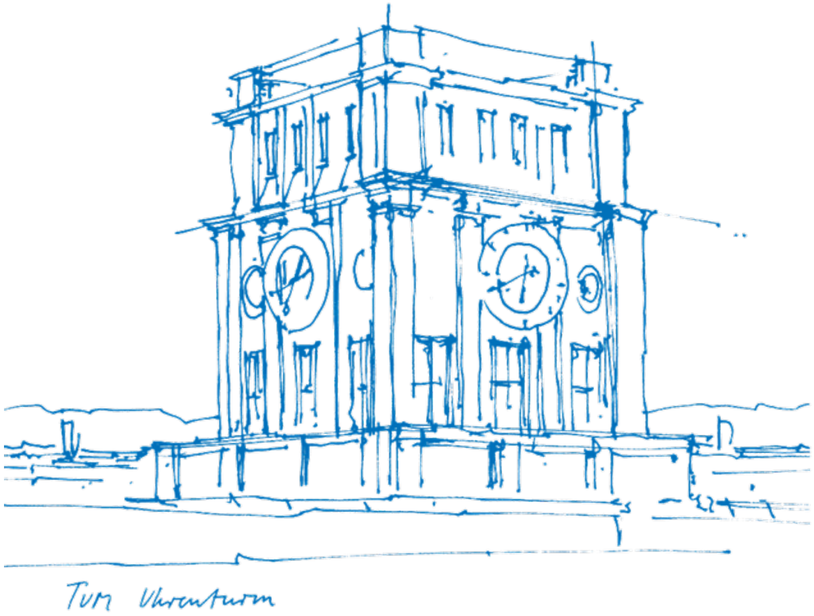
Exploring Hybrid CTC/Attention End-to-End Speech Recognition with Gaussian Processes

Dipl.-Ing. Ludwig Kürzinger
Technische Universität München
Department of Electrical and Computer Engineering
Chair for Human Machine Interaction
Istanbul, 22. August 2019



Contributions

1. Sequential Gaussian Process hyperparameter optimization for the hybrid CTC/Attention end-to-end speech recognition

2. Distinct parameter groups found in architecture exploration

3. We revisit the *hybrid CTC/Attention hypothesis*:

HYP: CTC primarily regularizes alignments of the attention mechanism

Contributions

1. Sequential Gaussian Process hyperparameter optimization
   for the hybrid CTC/Attention end-to-end speech recognition

2. Distinct parameter groups found in architecture exploration

3. We revisit the *hybrid CTC/Attention hypothesis*:

> **HYP:** CTC primarily regularizes alignments of the attention mechanism

Contributions

1. Sequential Gaussian Process hyperparameter optimization
   for the hybrid CTC/Attention end-to-end speech recognition

2. Distinct parameter groups found in architecture exploration

3. We revisit the *hybrid CTC/Attention hypothesis*:

HYP: CTC primarily regularizes alignments of the attention mechanism

PART 1

Preliminaries

- Gaussian Process Optimization
- Hybrid CTC/Attention ASR

Gaussian Process Optimization

- Many parameters, but few of them are more influential?

  $\Rightarrow$ GP optimization is better than *grid* or *random search*

- Black box function $f(x)$ approximated by a kernel

$$k_{\text{Matérn}}(r^{(n)}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r^{(n)}}{l}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}r^{(n)}}{l}\right), \quad \text{with} \quad r^{(n)} = ||X^{(n)} - X'^{(n)}||. \tag{1}$$

- Sequential optimization
- Next point is chosen by maximizing the *Expected Improvement*

$$f_{\text{EI}}(X^{(n+1)}) = \mathbb{E}[max(0, f_{min} - f_{GP}(X^{(n+1)}))|X^{(n+1)}, D]. \tag{2}$$

Gaussian Process Optimization

- Many parameters, but few of them are more influential?

  $\Rightarrow$ GP optimization is better than *grid* or *random search*

- Black box function $f(x)$ approximated by a kernel

$$k_{\text{Matérn}}(r^{(n)}) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\frac{\sqrt{2\nu}r^{(n)}}{l})^{\nu}K_{\nu}(\frac{\sqrt{2\nu}r^{(n)}}{l}), \quad \text{with} \quad r^{(n)} = ||X^{(n)} - X'^{(n)}||. \tag{1}$$

- Sequential optimization
- Next point is chosen by maximizing the *Expected Improvement*

$$f_{\text{EI}}(X^{(n+1)}) = \mathbb{E}[max(0, f_{min} - f_{GP}(X^{(n+1)}))|X^{(n+1)}, D]. \tag{2}$$

Gaussian Process Optimization

- Many parameters, but few of them are more influential?
  $\Rightarrow$ GP optimization is better than *grid* or *random search*

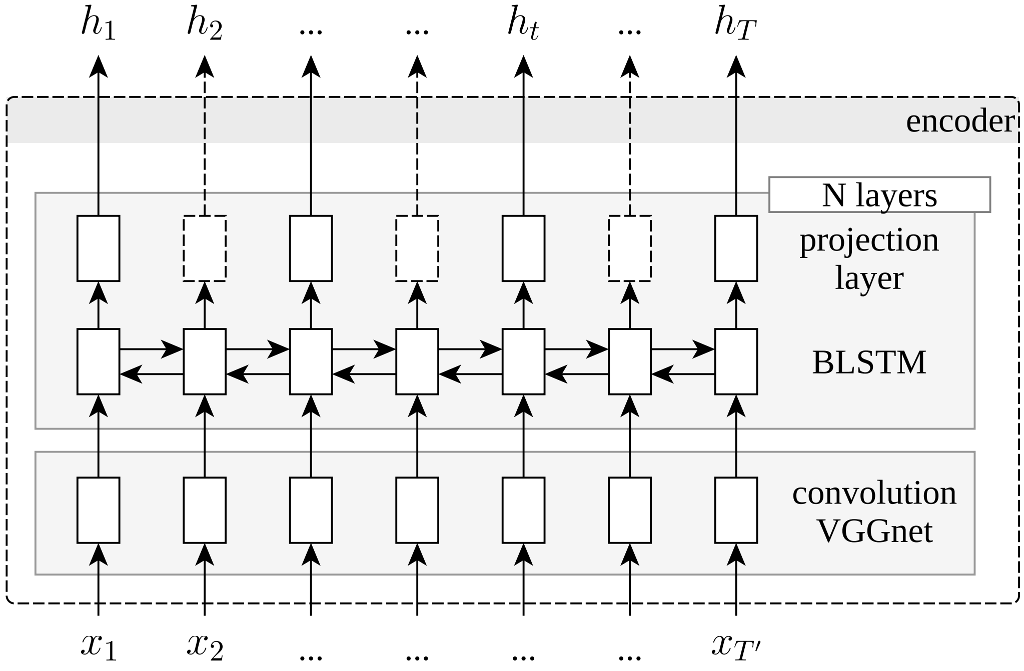
- Black box function $f(x)$ approximated by a kernel

$$k_{\text{Matérn}}(r^{(n)}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r^{(n)}}{l}\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}r^{(n)}}{l}\right), \quad \text{with} \quad r^{(n)} = ||X^{(n)} - X'^{(n)}||. \tag{1}$$

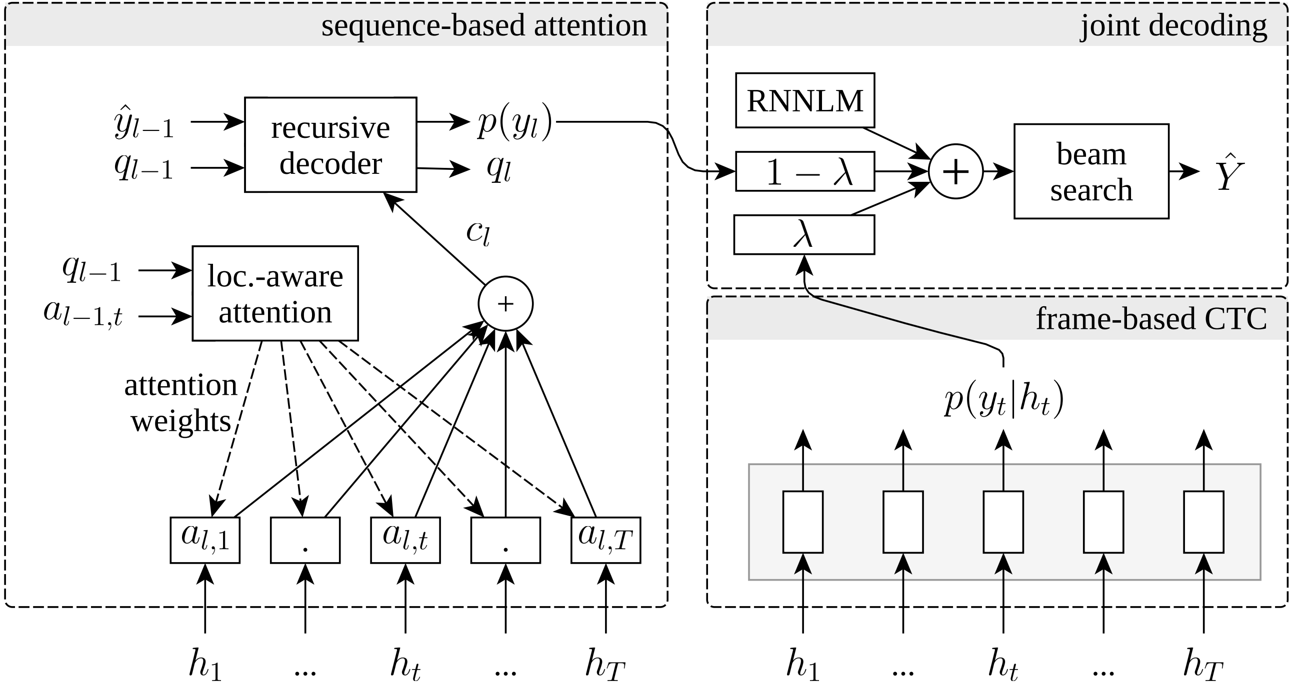
- Sequential optimization
- Next point is chosen by maximizing the *Expected Improvement*

$$f_{\text{EI}}(X^{(n+1)}) = \mathbb{E}[max(0, f_{min} - f_{GP}(X^{(n+1)}))|X^{(n+1)}, D]. \tag{2}$$

Hybrid CTC/Attention ASR - Encoder (1/2)



Hybrid CTC/Attention ASR - Decoder (2/2)



PART 2

Experiment Setup

Gaussion Process Optimization in Two Stages

Stage 1: Network Training

- 20 initial CTC/Attention models
- Lower and upper bounds on model parameters, e.g. CTC vs. attention $\lambda \in [0.0; 1.0]$
- $\Rightarrow$ in total **70** models

Stage 2: Beam Search

- Started with networks from stage 1 decoded with and without RNNLM
- Optimized parameters:
  (1) Weight of CTC activations
  (2) weight of the LM
- $\Rightarrow$ in total **590** beam search results

Gaussion Process Optimization in Two Stages

Stage 1: Network Training

- 20 initial CTC/Attention models
- Lower and upper bounds on model parameters, e.g. CTC vs. attention $\lambda \in [0.0; 1.0]$
- $\Rightarrow$ in total **70** models

Stage 2: Beam Search

- Started with networks from stage 1 decoded with and without RNNLM
- Optimized parameters:
  (1) Weight of CTC activations
  (2) weight of the LM
- $\Rightarrow$ in total **590** beam search results

PART 3

Results

- Observed Parameter Groups
- CTC-Only Networks
- Attention-Only Networks

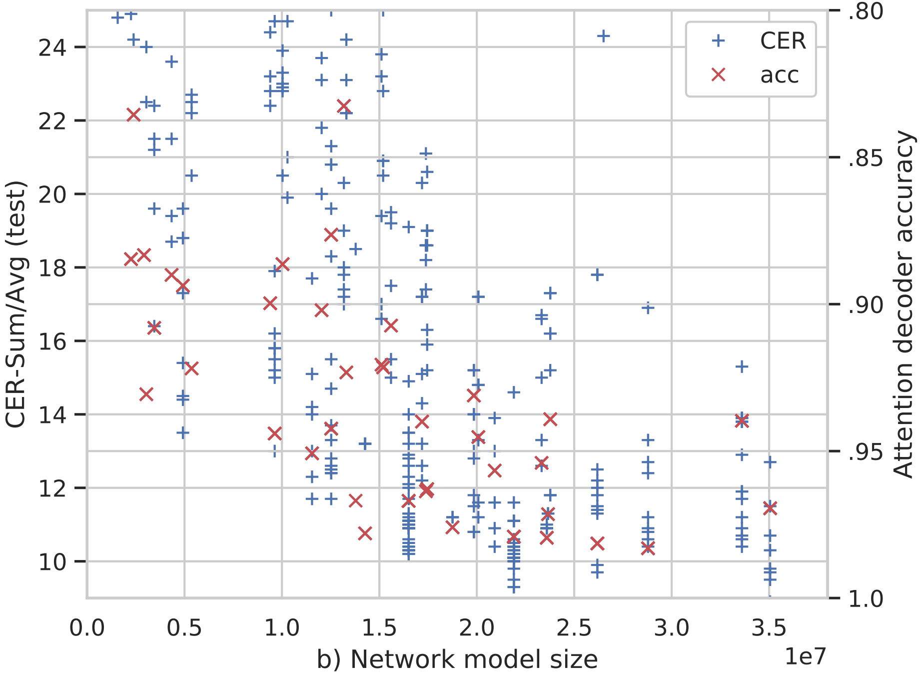
General Results

- Unsurprising:

  Deeper models are better

- Deeper attention decoders are better

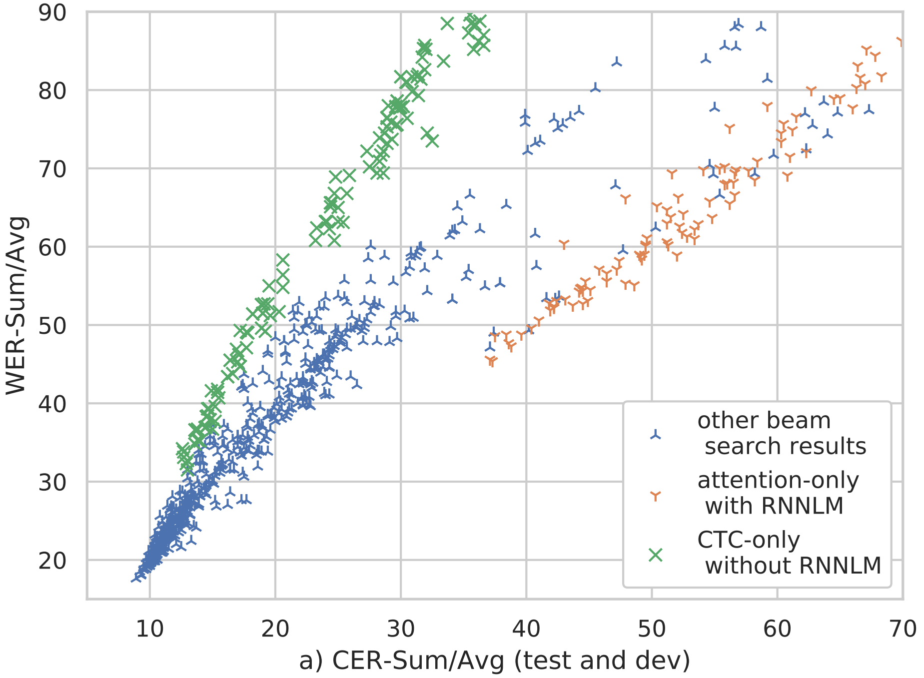
  they predict $p(y_t | y_{t-1})$, similar to a LM



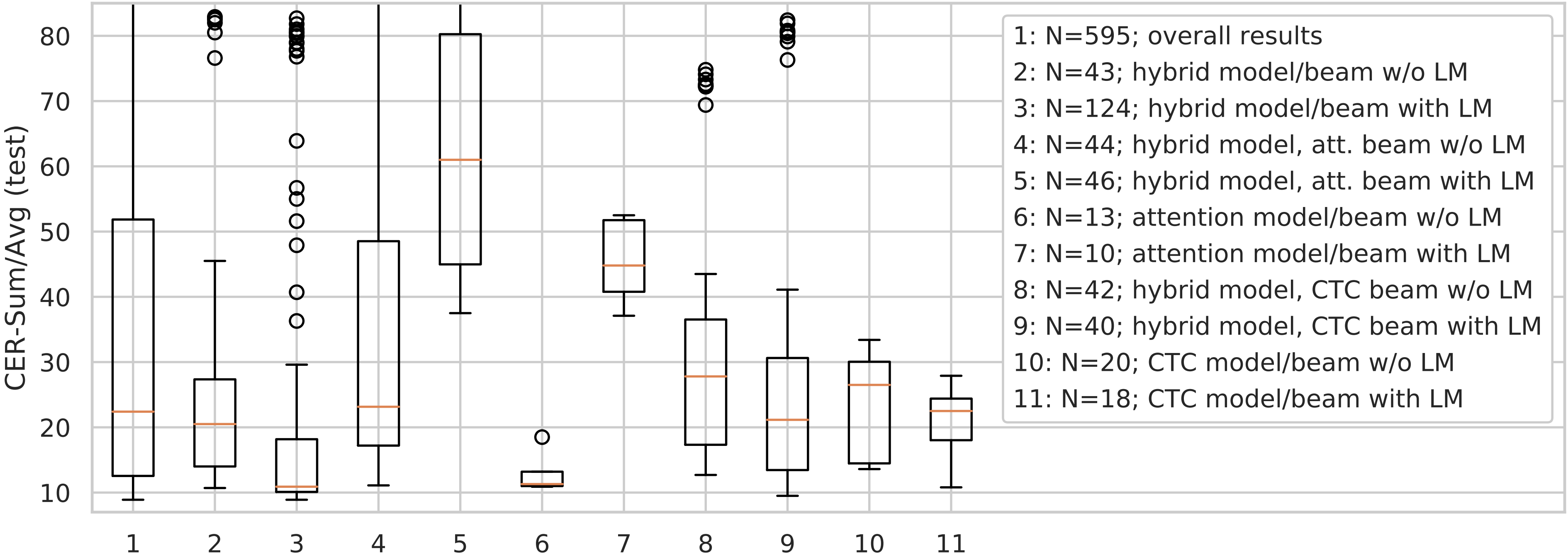
Observed Parameter Groups

- Deteriorated Results in some parameter configurations
  $\rightarrow$ CTC-only models without RNNLM
  $\rightarrow$ Attention-only models *with* RNNLM

- Optimization criterion based on CER



Parameter Groups Overview



c) Selected categories of results over the TEDlium v2 test set.

CTC-Only Networks

CTC-only example transcription without RNNLM

REF: BUT IN FACT we ARE CHANGED we ARE MARKED OF COURSE by a CHALLENGE whether PHYSICALLY EMOTIONALLY or BOTH AND i AM GOING TO SUGGEST THAT this IS A GOOD THING

HYP: UT AN VACT we AR CHANSD we AR MARK TOF CORTS by a CHALENE whether FISICALLY IMOSNOLY or BOS AN i ** ***** ** MMNOSUGJEST T this ** IC COD TING

- Many spelling errors
- comparatively high WER-to-CER
- $\rightarrow$ needs a language model

Attention-Only Networks

Attention-only example transcription with RNNLM

REF: `serves to be more DISABLING TO the individual than the pathology itself by not treating the ***** WHOLENESS of *** ****** *** *** ****** *** *** ****** *** *** ****** *** *** ****** *** *** ****** *** *** ****** *** A person *** *** ********`

HYP: `serves to be more ********* THAN the individual than the pathology itself by not treating the WHOLE NUCLEUS of THE PERSON AND THE PERSON AND THE PERSON AND THE PERSON AND THE PERSON AND THE PERSON AND THE PERSON AND THE person AND THE PERSONAL`

- Feedback loops
- Also: dropped sentence parts
- comparatively low WER-to-CER
- $\rightarrow$ Misplaced attention focus

Best results in selected categories.

Parameter	baseline [?]	with LM hybrid	with LM att.-only	with LM CTC-only	without LM hybrid	without LM att.-only	without LM CTC-only
training parameters							
Encoder Layers	6	6	6	6	4	6	6
Decoder Layers	1	5	2	2	3	2	2
Attention neurons	320	379	100	350	172	100	350
Multi-obj. (training) κ	0.5	0.69	0.00	1.00	0.15	0.00	1.00
Model size ($1e6$)	18.7	35.1	23.6	26.6	28.8	23.6	26.6
beam search parameters							
RNNLM weight β	1.0	0.73	0.41	1.00	0.00	0.00	0.00
Multi-obj. (beam) λ	0.3	0.62	*0.08*	1.00	0.15	0.00	1.00
results							
TEDlium 2 test/CER	10.1	**8.9**	40.2	11.3	10.6	10.9	14.4
TEDlium 2 test/WER	18.6	**17.6**	49.3	22.6	22.1	22.4	36.9

PART 4

Discussion

- Discussion: Feedback Loops
- Concluding Remarks

Discussion: Attention Feedback Loops

But other publications with an Attention-based model + LM had a better performance!
Why Attention-only models + LM performed so bad?

Hybrid CTC/Attention

- Teacher Forcing
- Location-aware attention
  depends on s_{l-1} and a_{l-2}
  (*feedback delay of up to two steps*)

e.g. Listen-Attend-Spell

- Scheduled Sampling
- LSTM transducer attention
  based on s_l
  (*no delay for feedback on attention*)

Discussion: Attention Feedback Loops

But other publications with an Attention-based model + LM had a better performance!
Why Attention-only models + LM performed so bad?

Hybrid CTC/Attention

- Teacher Forcing
- Location-aware attention
  depends on s_{l-1} and a_{l-2}
  (*feedback delay of up to two steps*)

e.g. Listen-Attend-Spell

- Scheduled Sampling
- LSTM transducer attention
  based on s_l
  (*no delay for feedback on attention*)

Discussion: Attention Feedback Loops

But other publications with an Attention-based model + LM had a better performance!
Why Attention-only models + LM performed so bad?

Hybrid CTC/Attention

- Teacher Forcing
- Location-aware attention
  depends on s_{l-1} and a_{l-2}
  (*feedback delay of up to two steps*)

e.g. Listen-Attend-Spell

- Scheduled Sampling
- LSTM transducer attention
  based on s_l
  (*no delay for feedback on attention*)

Concluding Remarks

Previously as hypothesis of the *hybrid CTC/Attention model*:

CTC primarily regularizes alignments of the attention mechanism

Results indicate that:

CTC instead regularizes the impact of LM feedback in the attention mechanism

Dipl.-Ing. Ludwig Kürzinger
Technische Universität München
Department of Electrical and Computer Engineering
Chair for Human Machine Interaction
Istanbul, 22. August 2019