NUANCE | WHAT'S NEXT BLOG

Enterprise R&D, Innovation & Research

Prediction Network Architecture in RNN-T for ASR

Innovation at Nuance | Employee Guest Blogger

October 5, 2022

Recurrent Neural Network Transducer (RNN-T) models have gained popularity in commercial systems because of their competitiveness and capability of operating in online streaming mode. In this work we focus on the Prediction Network, an important module in a RNN-T, by comparing four different architectures. We also propose a new simpler Prediction Network, N-Concat, that outperforms the others in our online streaming benchmark.

About the author: Dario Albesano is a Principal Research Scientist at Nuance Communications. He received his master degree in Computer Science and Artificial Intelligence from Università di Torino – Italy in 1990. His research interests include neural networks, speech recognition, artificial intelligence, acoustics and music reproduction.

Acknowledgements: Jesus Andrés-Ferrer contributed equally to this work; Nicola Ferri and Puming Zhan also contributed to this work. The paper was presented at the Interspeech 2022 conference, 18 -22 September 2022.

Recurrent Neural Network Transducer (RNN-T), Connectionist Temporal Classification (CTC) and Attention

Encoder Decoder (AED) are among the most promising End-2-End (E2E) models presently used for Automatic Speech Recognition (ASR). E2E approach has gradually replaced the traditional ASR methods based on totally separated components, where each component demanded for specific know-how and expertise for its design, development, optimization and integration into the target ASR system. On the contrary, E2E models natively integrate all the needed steps for converting speech into the corresponding text transcription, simplifying the whole ASR system. Although RNN-T, CTC and AED offer very good accuracy in recognizing speech, RNN-T typically outperform the others and while being naturally suitable for online streaming mode, it enables the development and deployment of real-time speech recognition applications.

A RNN-T model (Fig. 1) is composed of 3 components: the acoustic Encoder, that receives in input the speech segments to be recognized and generates a corresponding high-level representation; the Prediction Network that autoregressively incorporates previously emitted symbols into the model; the Joiner, that mixes both acoustic and autoregressive label representations via a monotonic alignment process. Despite the Prediction Network being a core RNN-T differentiator, there is a lack of deep understanding of its role. Some works show that the Prediction Network plays a language modeling role while others challenge this interpretation.

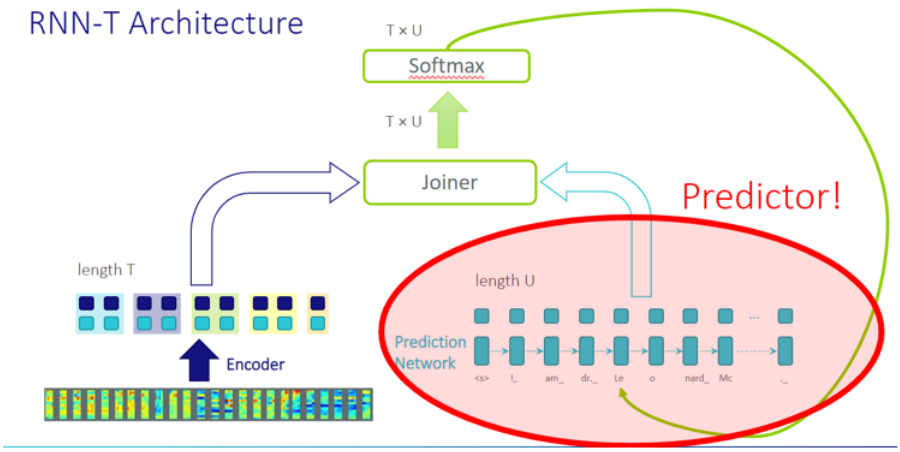

Figure 1 – RNN-T architecture. Bottom left: speech signal is encoded by the acoustic Encoder; bottom right: the Prediction Network, focus of this work - highlighted in the red ellipse, receives in input the symbol predicted by the model at the previous step. In the middle, the Joiner fuses the Encoder and Prediction Network contributions. The Softmax layer transforms the Joiner output into a probability distribution.

In our paper entitled On the Prediction Network Architecture in RNN-T for ASR, we present a systematic study of several Prediction Networks under various conditions. While some partial explorations have been done in the literature, we conduct a clean comparison of several Prediction Networks, namely: LSTM, Transformer, Conformer and Tied-Reduced, together with a common state-of-the-art Conformer encoder. Inspired by Tied-Reduced approach we propose N-Concat, a new n-gram Prediction Network that outperforms all the other Prediction Networks in online configurations on both the Librispeech 100 hours subset and an internal medical data set consisting of 1000 hours speech data from doctor-patient conversations. We assess the Prediction Networks in two different regimes, streaming (On-line) and batch recognition (Off-line), as well as with characters and word-pieces vocabularies. Our paper is accepted for presentation at Interspeech 2022 conference.

Moving to a more accurate description, the Prediction Networks architectures we have evaluated are the following:

LSTM: it consists of a single layer of unidirectional LSTM, to model the full left context of the predicted symbols, thanks to the internal LSTM state.

Transformer: the main contribution comes from the Multi Head Self Attention (MHSA), that makes the Prediction Network able to keep into account the global context information, focusing on the most important past segments. Please see Fig.2

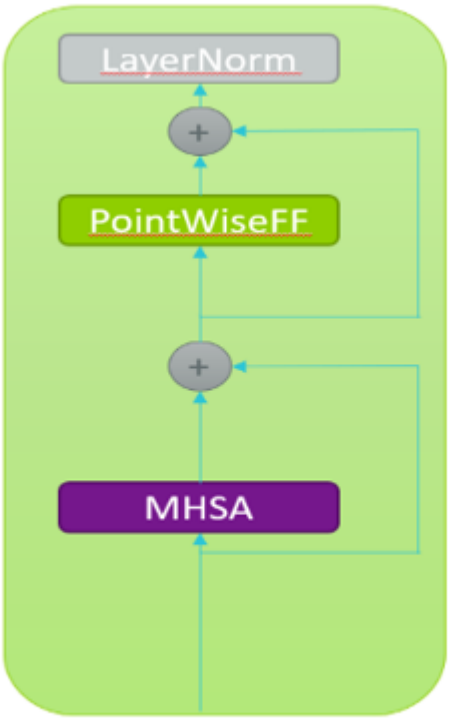

Figure 2 – Transformer Prediction Network.

Conformer: compared to the Transformer Prediction Network, it adds a Convolutional layer to better catch and model the local correlations. Please see Fig. 3.

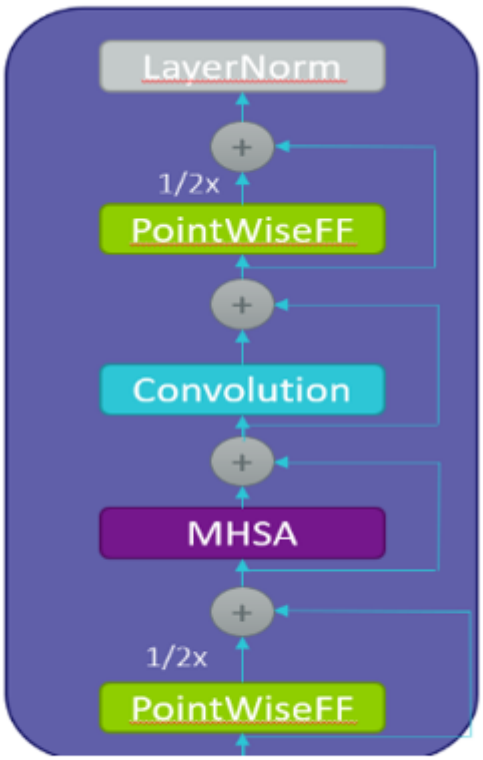

Figure 3 – Conformer Prediction Network.

Tied-Reduced (N-Avg): simplifies the autoregressive dependency to an n-gram dependency. The previous N-1 predicted labels are encoded by the shared embedding matrix, and then weighted with a weight achieved by using devoted positional encodings, trained together with the Prediction Network training. Then, the weighted embeddings are averaged across all the N-1 positions as well as across all the heads. Please see Fig. 4.

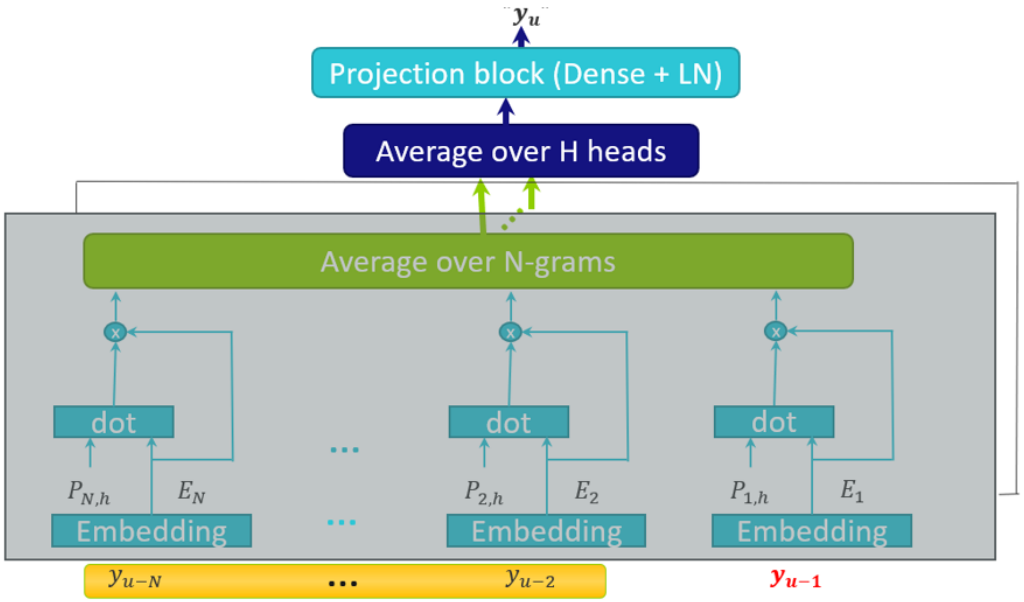

Figure 4 – Tied – Reduced Prediction Network

N-Concat: derived from Tied-Reduced, N-Concat introduces a specific bias towards attending different context labels, by splitting the embedded labels across the heads. Then, concatenation through the heads is performed instead of averaging. Please see Fig. 5.

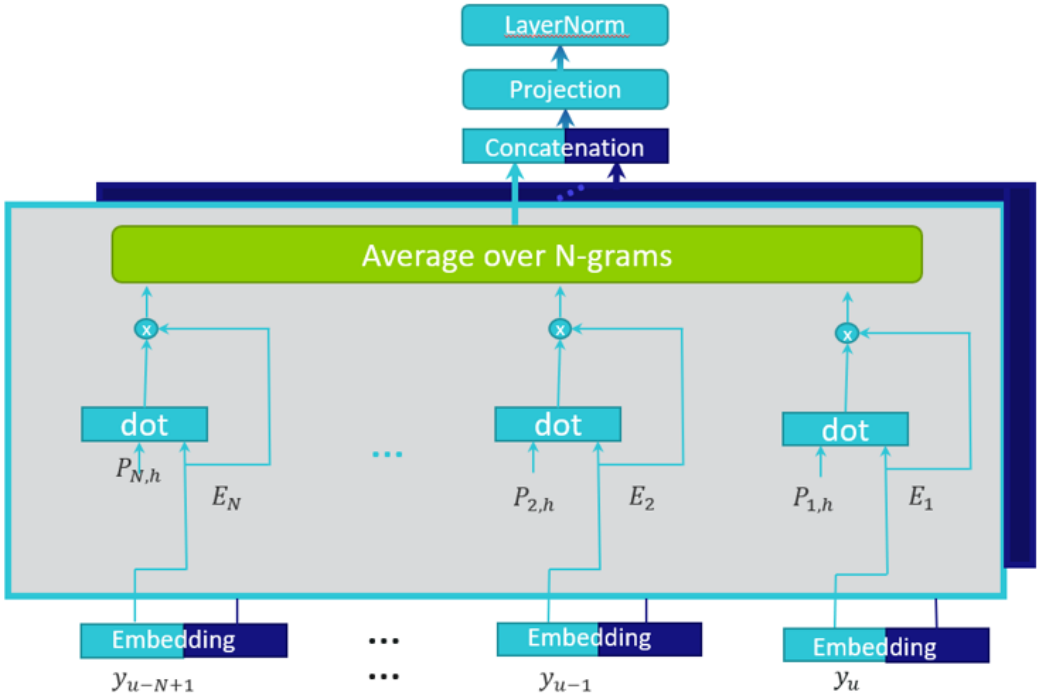

Figure 5 – N-Concat Prediction Network

To evaluate the Prediction Networks we used 100 hours of transcribed speech data from Librispeech corpus (LS100), with 30 characters and 300 word-pieces output vocabularies. Besides, we also used an Internal, speaker independent, medical speech transcription task (D2P1K) of doctor-patient conversations

across multiple specialties; the training set contains 1000 hours speech data and the test set contains 263,000 words. The vocabulary has 2500 word-pieces.

We investigated the Prediction Networks performance, in both off-line and on-line scenarios (1 second induced latency). The LS100 models have 30 million parameters, while the D2P1K models have 67 million parameters.

The results on on-line LS100 show that N-Concat is significatively better than the other Prediction Networks on word Pieces (up to 4.1% relative Word Error Rate Reduction (WERR)) and it reduces 8.5 times the number of the Prediction Network parameters w.r.t LSTM Prediction Network and 22.9 times w.r.t. Conformer Prediction Network. Please see table 1 for more details. On Off-line LS100, N-Concat is competitive with Transformer and Conformer Prediction Networks and slightly worse than LSTM Prediction Network.

		Characters		Word-pieces	
Prediction Network	**#params**	**Test-Clean**	**Test-Other**	**Test-Clean**	**Test-Other**
LSTM	0.59M	7.3	21.5	7.3	21.5
Transformer	0.9M	7.1	22.1	7.4	22.0
Conformer	1.6M	7.3	21.8	7.5	21.9
N-Avg	0.09M	7.1	21.9	7.3	21.3
N-Concat	0.07M	7.1	21.9	7.0	21.2

Table 1 – On-line results on LS100

On D2P1K On-line scenario, all the Prediction Networks achieve a similar accuracy, in terms of Word Error Rate (WER), but N-Concat achieves the best Real Time Factor (RTF) and the biggest Batch Size (computed at RTF=1) when Mono-RNNT greedy decoding is applied. Please see table 2 for more details. In off-line scenario, N-Concat performs best in terms of WER.

	LSTM	Transf.	Conf.	N-Avg	N-Conc
WER	14.2	14.1	14.1	14.3	14.2
RTF	1.22	1.18	1.23	1.18	1.17
BS	2.5K	2.4k	2.2k	2.5k	2.6k

Table 2 – On-line results on D2P1K

Finally, on LS100, varying the left context length at inference of a model trained with left context 127, figure 6 shows that N-Concat Prediction Network reaches maximal performance when using 4 left tokens at inference. On the contrary, for Transformer Prediction Network a matched left context length at

training and inference is needed for achieving maximal accuracy, though it still underperforms the N-Concat Prediction Network in such case and it is much more sensitive in short left context length conditions. Similar observations hold for D2P1K.

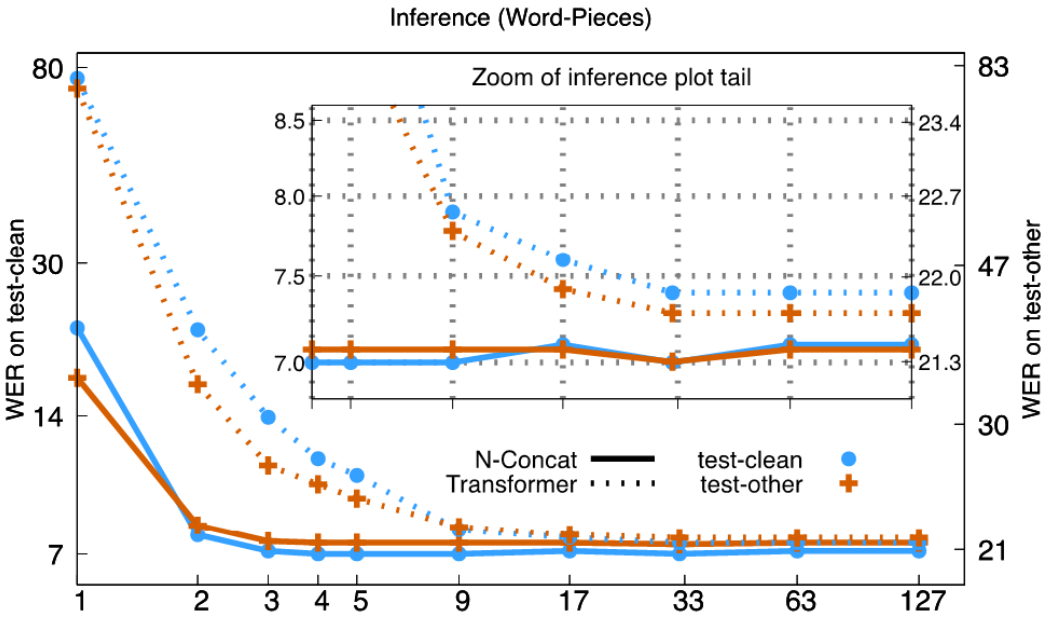

Figure 6: Inference left-context behavior for N-concat and Transformer Prediction Networks when limiting the left-context at inference time

In conclusion, we have observed that compared to the Off-line regime, in the On-line regime the Prediction Network is more effective. Interestingly, Transformer and N-Concat Prediction Networks showed resilient to limited left context at training/inference matching conditions. However, both Prediction Networks are unable to exploit long term output label dependencies in contrast to many advanced language models. In this context, N-Concat more efficiently exploits the left context than Transformer Prediction Networks while reducing the Prediction Network parameters significantly and outperforming the other Prediction Networks in On-line RNN-T models.

Tags: Automatic speech recognition, Research & development