DEFORMER: COUPLING DEFORMED LOCALIZED PATTERNS WITH GLOBAL CONTEXT FOR ROBUST END-TO-END SPEECH RECOGNITION

INTERSPEECH 2022
September 18 - 22 • Incheon Korea
International Speech Communication Association

Jiamin Xie and John H. L. Hansen
Jiamin.Xie@utdallas.edu, John.Hansen@utdallas.edu
Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas
Richardson, Texas 75080-3021, U.S.A.

INTERSPEECH 2022 Sept. 18-22, 2022 Incheon, Korea

UTD CRSS
UT DALLAS

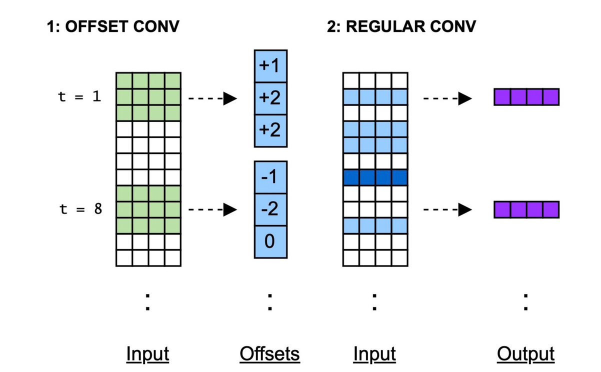
1 Introduction

- **Background**
  - Convolutional neural networks (CNN) have improved speech recognition greatly by exploiting localized T-F patterns
  - Patterns are assumed to exist in a **rigid** and **symmetric** kernel
  - What about **asymmetric** kernels? How will the localized pattern learned in this way interact with each other in a global context?
- **Proposition**
  - Use the **Deformable CNN (DCNN)** to replace regular CNNs
  - Analyze localized patterns obtained and its global interaction by modifying the popular Conformer [1] architecture
  - Experiment different initialization methods for the DCNN

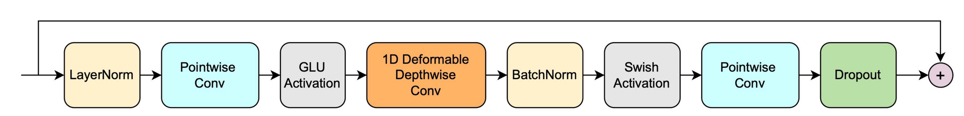
[1] A.Gulati, J.Qin, C.-C.Chiu, N.Parmar, Y.Zhang, J.Yu, W.Han, S. Wang, Z. Zhang, Y. Wu et al., "Conformer: Convolution- augmented transformer for speech recognition," *Interspeech 2020*.

2 Motivation

- **Learnable kernel by the DCNN**
  - The DCNN consists of two regular convolution steps to compute
  - One to compute the kernel offsets from the input, another for the output
  - **The learned positions can overlap and focus on certain input region**



3 Proposed method



- **Encoder using 1-D DCNN**
  - We replace the 1D depth-wise convolution in the convolution module of the Conformer by its deformable variant
- **Formulation**
  - 1-D deformable convolution [2]

$$\Delta p = Conv1D_{offset}(X,\ p_0) \qquad X(p') = X(\lfloor p' \rfloor) * (\lfloor p' \rfloor - p' + 1) + X(\lfloor p' \rfloor + 1) * (p' - \lfloor p' \rfloor)$$

$$p' = p_0 + \Delta p \qquad Y = Conv1D_{output}(X,\ p')$$

[2] K. An, Y. Zhang, and Z. Ou, "Deformable tdnn with adaptive receptive fields for speech recognition," *Interspeech 2021*.

4 Proposed method

- **Formulation cont.**
  - 1-D deformable depth-wise convolution

$$X \in \mathbb{R}^{T\times F} \xrightarrow{\text{Slice}} X_g \in \mathbb{R}^{T\times F/g} \xrightarrow[\{Conv1D_{offset}(X_{g1}),\ Conv1D_{output}(X_{g2})\}]{\text{DCNN}} Y_g \in \mathbb{R}^{T\times N/g} \xrightarrow{\text{Concatenate}} Y \in \mathbb{R}^{T\times N}$$

  - Which convolution step to be made depth-wise or both?
    - *Intuitive to make $Conv1D_{output}$ depth-wise*
    - *Debatable to make $Conv1D_{offset}$ depth-wise*

5 Proposed method

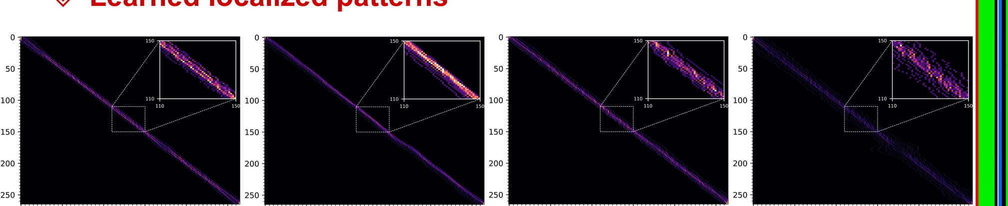
- **Deformer architecture**
  - 12-layer Deformer encoder and 6-layer Transformer decoder
  - A mix of deformable and non-deformable layers works the best
  - Deformation in the early layer can help as well

Configurations	Deformable Layers	Non-deformable Layers
Layer Index	{1, 6, 7, 10, 11}	{0, 2-5, 8, 9}
Layer Dimensions	256	256
Attention Heads	4	4
Kernel Size	15	15
Dilation	1	1
Stride	1	1
Convolution Groups	256	256
Deformable Groups	1	1

Table 1: *Deformer Encoder Configuration*

6 Pattern Analysis

- **Learned localized patterns**

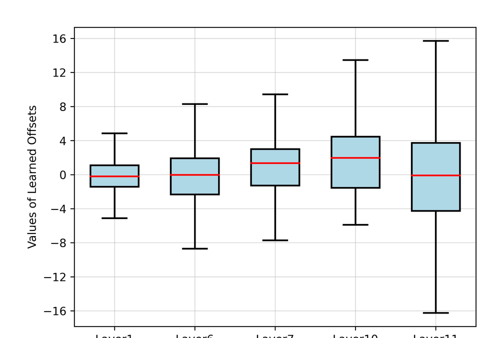


(a) *Deformer Layer 1* (b) *Deformer Layer 6* (c) *Deformer Layer 7* (d) *Deformer Layer 11*

- On the utterance level, kernels follow the diagonal pattern that was initialized
  - Confirms the importance of a monotonic alignment
- Deviation in the receptive field of each kernel is small, larger in deeper layers
  - Indicates the benefits of a localized structure
- Concentration-like pattern in the learned receptive field of each kernel
  - Several positions overlap and form a focus on a specific input region
  - Other positions scatter and form support for the local focus

7 Pattern Analysis

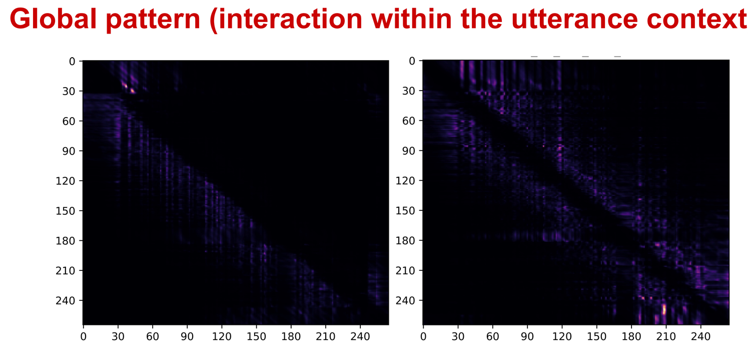
- **Localized offset statistics**



- Increasing spread of distribution over layers
  - Deeper layers learn both larger offsets and receptive fields for the kernel
- The distribution has a long tail and is symmetric around zero median
  - Verifies a persisting concentration-like pattern even in different utterances

8 Pattern Analysis

- **Global pattern (interaction within the utterance context)**



(a) *Conformer attention map* (b) *Deformer attention map*

- Compared both attention maps of a single head in the last layer
  - The Deformer attends additionally to the future above the diagonal
  - Indicates deformation brings more relevance among features, but is **structured** and kept in a small context

9 Pattern Analysis

- **Quantitative evaluation of pattern interaction**

Model	Globalness $(0, 7.48]$	Verticality $[-7.48, 0)$	Diagonality $[-0.75, 0]$
Conformer	4.57	-6.95	-0.18
Deformer	4.82	-6.97	-0.19

- Evaluate attention heads of the Deformer using three metrics from [3]
  - The globalness increased by +3.3% with a slight decrease of 0.3% in verticality and 1.3% in diagonality, within each respective range
  - Verifies a boost in global relevance among features
  - Indicates the boost retains the original attention structure, for example, a vertical or a diagonal structure

[3] S.-w. Yang, A. T. Liu, and H.-y. Lee, "Understanding self-attention of self-supervised audio transformers," *Interspeech 2020.*

10 Experimental Setup

- **Front-end**
  - Fbank+Pitch features
  - SpecAugument [4]
- **Data split**
  - WSJ (train: *si284,* dev: *dev93*, eval: *eval92*)
- **Systems**
  - Baseline: 12-layer Conformer encoder + 6-layer Transformer decoder
  - Model: 12-layer Deformer encoder + 6-layer Transformer decoder
  - The training setup follows the exact recipe in the Espnet [5]
- **Experiments**
  - Initializing DCNN with Xavier random weights or weights of zeros
  - Varying 0.5x or 1x learning rates for the DCNN compared to the rest of the network
  - Changing the number of deformable groups to verify the depthwise computation

[4] D.S.Park, W.Chan, Y.Zhang, C.-C.Chiu, B.Zoph, E.D.Cubuk, and Q.V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv:1904.08779, 2019*
[5] S.Watanabe, T.Hori, S.Karita, T.Hayashi, J.Nishitoba, Y.Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen et al., "Espnet: End-to-end speech processing toolkit," *arXiv:1804.00015, 2018.*

11 Results

- **Parameter initialization**

Model	#Params	Initialization	WER (%) dev	WER (%) eval	CER (%) dev	CER (%) eval
Conformer Base	43.05M	Xavier	11.2	8.9	3.9	3.0
Deformer (*mult=0.5*)	43.34M	Xavier	10.7	**8.4**	3.7	**2.8**
Deformer (*mult=1.0*)	43.34M	Xavier	11.1	9.0	3.9	3.0
Deformer (*mult=0.5*)	43.34M	Zero	10.8	8.8	3.7	3.0
Deformer (*mult=1.0*)	43.34M	Zero	10.5	**8.4**	3.7	2.9

- Effectively initializing offsets from zero performs the best overall and gives a more calibrated system
  - Improves the Conformer baseline by a +5.6% relative
  - Different learning rate multiplier yield results differently, but the fluctuation is smaller with zero intialization

12 Results

- **Deformable groups**

Model	Deformable Groups	WER (%) dev	WER (%) eval	CER (%) dev	CER (%) eval
Deformer (*mult=0.5, zero init.*)	256	11.2	9.1	3.9	3.0
Deformer (*mult=1.0, zero init.*)	256	11.1	8.9	3.9	3.0
Deformer (*mult=0.5, zero init.*)	2	11.1	8.6	3.8	2.9
Deformer (*mult=1.0, zero init.*)	2	10.9	8.6	3.9	**2.8**
Deformer (*mult=0.5, zero init.*)	1	10.8	8.8	3.7	3.0
Deformer (*mult=1.0, zero init.*)	1	10.5	**8.4**	3.7	2.9

- Lower number of deformable groups performs better
  - Unnecessary to make the offset convolution depth-wise
  - A small number of deformable groups probably find the mapping to offsets that are more generalizable

13 Results

- **Integration with a language model (LM)**

Model	WER (%) dev	WER (%) eval	CER (%) dev	CER (%) eval
Conformer Base	7.0	4.7	3.1	2.1
Deformer (*mult=1.0, zero init.*)	6.7	**4.4**	2.9	**2.0**

- Consistent improvement found in the proposed Deformer encoder vs. the Conformer encoder
  - Achieves +6.7% relative WER improvement
  - Performance persists after adding an external LM

14 Discussion

- **Sensitivity towards the learning rate multiplier**
  - We note the learning rate multiplier can influence the training and the subsequent performance of the Deformer
  - Further investigation is needed, e.g. whether to release the tying of learning rate multiplier in different layers
- **Towards the full capacity of DCNN**
  - The DCNN can behave same as the regular CNN when offsets result in zero
  - We did not observe a benefit from replacing all CNN layers by the DCNN, which needs a future investigation

15 Conclusions

- **Outcomes**
  - A novel encoder design based on data-driven CNN kernels is presented for end-to-end speech recognition
  - It is shown DCNN improves results by focusing on local content and enhancing feature interaction in the global context
  - It is shown initializing the offset convolution with weights of zero is robust against changes in a hyper-parameter and a small number of deformable groups are beneficial
- **Future work**
  - Further alleviate the tuning requirement of learning rate multiplier by introducing regularization or better initialization
  - Investigate stacking more DCNNs to unlock the full capacity of asymmetric CNN kernels which are obtained by learning