2020「中技社科技獎學金」
2020 CTCI Foundation Science and Technology Scholarship
研究獎學金
Research Scholarship

DARTS-ASR: Differentiable Architecture Search for Multilingual Speech Recognition and Adaptation

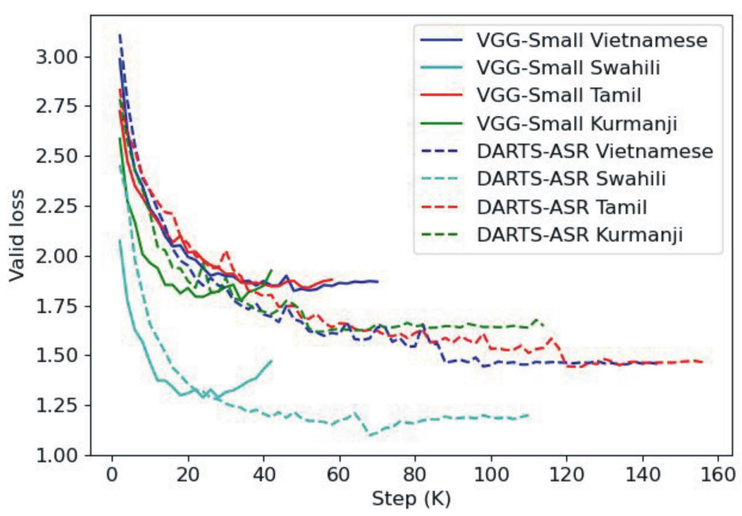
國立臺灣大學 電信工程研究所 博二 陳奕禎
指導教授: 李宏毅 教授

Motivation & Background

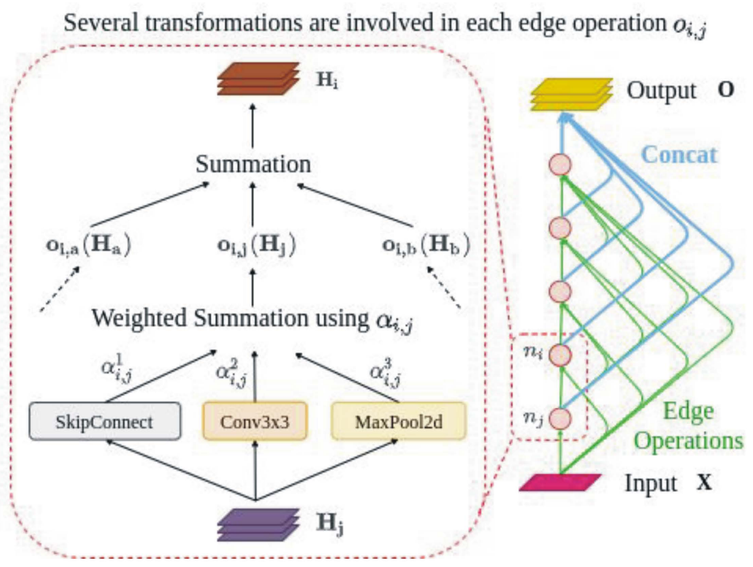
In previous works, only parameter weights of **automatic speech recognition (ASR)** models are optimized under fixed-topology architecture. However, the design of successful model architecture has always relied on human experience and intuition. Automatic **neural architecture search (NAS)** for ASR, aiming to optimize not only parameter weights but also the design of architecture itself, is intriguing and valuable.

We propose an ASR approach with efficient gradient-based architecture search, **DARTS-ASR**. To examine the generalizability, we apply DARTS-ASR not only on many languages, but also on a multilingual transfer learning setting.

Validation Loss vs Training Step



Approach: DARTS-ASR



The search space is a directed acyclic graph consisting of K nodes $\{n_0, \ldots, n_K\}$, where n_0 is the input feature X and other nodes represent latent features $H_1, \ldots, H_K$. (Here X is a segment of acoustic features such as Mel filterbanks and H_i have the shape like CNN features.)

For each node n_i, there are i directed input edges, where each edge transforms H_j with some operation $g_{i,j}$.

$$\mathbf{H_i} = \Sigma_{j<i}\ g_{i,j}(\mathbf{H_j}), \quad (1)$$

$$where\ g_{i,j}(\mathbf{H_j}) = \Sigma_{f\in\mathbb{F}} \frac{\exp(\alpha^f_{i,j})}{\Sigma_{f'\in\mathbb{F}}\exp(\alpha^{f'}_{i,j})} f(\mathbf{H_j}). \quad (2)$$

The transformation candidates we used were {3x3 conv, 5x5 conv, 3x3 dilated conv, 5x5 dilated conv, 3x3 average pool, 3x3 max pool, skip connection}.

Variables $\alpha_{i,j}$ is jointly trained with parameter weights directly by gradient descent. if they are sparse, equation (2) can be regarded as the selection of transformations used to connect node n_i and n_j, so it can be considered as controlling the architecture.

Experiments (Data: IARPA BABEL)

Table 1: Monolingual ASR with different CNNs (CER, %)

Language	CNN Module			
	VGG-Small	VGG-Large	DARTS-ASR	
			Full	Only Conv3x3
Vietnamese	46.0	48.3	**40.9**	45.7
Swahili	39.6	38.3	**35.9**	36.8
Tamil	57.9	60.1	**48.0**	51.6
Kurmanji	57.2	56.8	**55.5**	56.5

**Pretraining lang.: Bengali, Tagalog, Zulu*

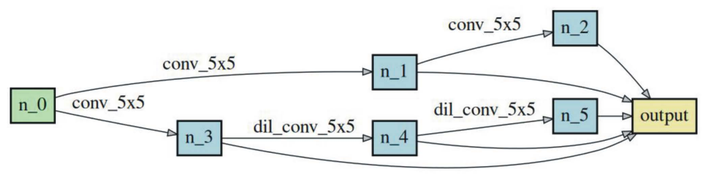
Table 2: Multilingual ASR with different finetuning approaches (CER, %)

Language	Fine-tuning of DARTS-ASR		
	Adapt only param.	Adapt arch.+param.	Adapt pruned arch.+param.
Vietnamese	**40.9**	**40.9**	41.1
Swahili	33.2	**32.3**	35.3
Tamil	46.4	**45.9**	47.5
Kurmanji	53.6	53.5	**53.2**

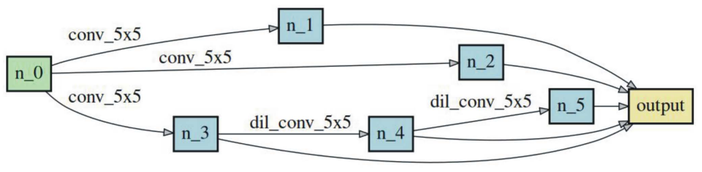
Table 3: Multilingual ASR with different CNNs (CER, %)

Language	CNN Module		
	VGG-Small	VGG-Large	DARTS-ASR
Vietnamese	45.3	43.2	**40.9**
Swahili	36.3	36.1	**32.3**
Tamil	55.7	55.0	**45.9**
Kurmanji	54.5	55.1	**53.5**

Found architectures for different languages (multi):



(Vietnamese, Kurmanji)



(Swahili, Tamil)

研究生活與心得

加入台大機器學習與語音處理實驗室已經邁入第四年，陸續學習、研究各種最新機器學習技術如self-/semi-supervised learning, feature disentanglement, architecture search等，並應用於語音技術上。研究成果也多次被刊登於國際頂尖期刊/會議上。這次研究更為本實驗室與NVIDIA產學合作的結果，並被刊登於國際語音頂會INTERSPEECH 2020上。非常感謝李琳山及李宏毅教授還有優秀同儕們一直以來的教導與討論，也非常感謝中技社的鼓勵，未來我也會繼續投注學習研究的熱忱，為產學界貢獻一份心力。