MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

All-in-One Transformer: Unifying Speech Recognition, Audio Tagging, and Event Detection

Moritz, Niko; Wichern, Gordon; Hori, Takaaki; Le Roux, Jonathan

TR2020-138 October 24, 2020

Abstract

Automatic speech recognition (ASR), audio tagging (AT), and acoustic event detection (AED) are typically treated as separate problems, where each task is tackled using specialized system architectures. This is in contrast with the way the human auditory system uses a single (binaural) pathway to process sound signals from different sources. In addition, an acoustic model trained to recognize speech as well as sound events could leverage multi-task learning to alleviate data scarcity problems in individual tasks. In this work, an all-in-one (AIO) acoustic model based on the Transformer architecture is trained to solve ASR, AT, and AED tasks simultaneously, where model parameters are shared across all tasks. For the ASR and AED tasks, the Transformer model is combined with the connectionist temporal classification (CTC) objective to enforce a monotonic ordering and to utilize timing information. Our experiments demonstrate that the AIO Transformer achieves better performance compared to all baseline systems of various recent DCASE challenge tasks and is suitable for the total transcription of an acoustic scene, i.e., to simultaneously transcribe speech and recognize the acoustic events occurring in it.

Annual Conference of the International Speech Communication Association (Interspeech)

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

All-in-One Transformer: Unifying Speech Recognition, Audio Tagging, and Event Detection

Niko Moritz, Gordon Wichern, Takaaki Hori, Jonathan Le Roux

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

{moritz, wichern, thori, leroux}@merl.com

Abstract

Automatic speech recognition (ASR), audio tagging (AT), and acoustic event detection (AED) are typically treated as separate problems, where each task is tackled using specialized system architectures. This is in contrast with the way the human auditory system uses a single (binaural) pathway to process sound signals from different sources. In addition, an acoustic model trained to recognize speech as well as sound events could leverage multi-task learning to alleviate data scarcity problems in individual tasks. In this work, an all-in-one (AIO) acoustic model based on the Transformer architecture is trained to solve ASR, AT, and AED tasks simultaneously, where model parameters are shared across all tasks. For the ASR and AED tasks, the Transformer model is combined with the connectionist temporal classification (CTC) objective to enforce a monotonic ordering and to utilize timing information. Our experiments demonstrate that the AIO Transformer achieves better performance compared to all baseline systems of various recent DCASE challenge tasks and is suitable for the *total transcription* of an acoustic scene, i.e., to simultaneously transcribe speech and recognize the acoustic events occurring in it.

Index Terms: automatic speech recognition, DCASE challenge, acoustic event detection, audio tagging, Transformer

1. Introduction

In recent years, researchers have reported that machines have reached human speech recognition performance on some well-defined tasks [1, 2]. However, humans are still unmatched in recognizing speech in difficult acoustic conditions [3] and for complex or mismatched tasks [4]. More remarkably, the human auditory system can detect and recognize acoustic signals irrespective of their nature using a single (binaural) pathway, whereas state-of-the-art systems used for audio tagging (AT), acoustic event detection (AED), and automatic speech recognition (ASR) mostly use task-specific architectures and are treated as separate problems. This is true even among different AED and AT domains, as can be noticed from the past Detection and Classification of Acoustic Scenes and Events (DCASE) challenges [5–10]. The human auditory system can thus still be regarded as a role model due to its strong detection and classification performance as well as its ability to better generalize to unknown sounds, tasks, and acoustic conditions.

Along the auditory pathway, an acoustic signal passes several processing stages, whereby early stages mainly extract and analyze different acoustic cues, while the final stages, in the auditory cortex, are responsible for perception [11]. Such processing is in many ways analogous to encoder-decoder neural network architectures, where the encoder extracts the important acoustic cues for a given task, the attention mechanism acts as the relay, and the decoder performs the perception, detecting and recognizing acoustic events.

In this work, we propose a unified ASR, AT, and AED system based on an encoder-decoder model. More specifically, we develop our system around the Transformer architecture [12], which has demonstrated improved end-to-end ASR results compared to recurrent neural network (RNN) based systems by leveraging self-attention to analyze the temporal information of an acoustic signal [13]. However, the full Transformer architecture has not yet been applied to AT or AED, where the use of self-attention has so far been limited to encoder-only architectures [14, 15]. Besides the analogy to the auditory system, another motivation for using an encoder-decoder architecture in AT and AED problems is that the decoder directly outputs symbols, i.e., class labels, thus avoiding the cumbersome process of setting detection thresholds for each class during inference [14]. Moreover, encoder-decoder based systems do not require a monotonic ordering of labels, and can thus easily make use of weakly labeled audio recordings, annotated without temporal or sequential information, which is often the case for acoustic events. However, this can also be a disadvantage when temporal information is needed, such as for AED and ASR. We thus train our Transformer model jointly with the connectionist temporal classification (CTC) objective for the AED and ASR tasks using a multi-objective CTC-attention type of architecture [16, 17] to leverage the monotonic alignment properties of CTC.

The present work aims at investigating the following questions: 1) Can we develop a system that moves closer to the versatility of the human auditory system? 2) Can training on multiple heterogeneous tasks lead to a single system with performance similar to or better than systems developed independently for each task? 3) Can a single system successfully handle multiple tasks with widely varying characteristics, large length discrepancies, and with or without monotonicity?

2. System Architecture

Figure 1 shows the proposed unified ASR, AT, and AED system, which is based on a joint CTC-attention architecture [16, 17]. The encoder and decoder neural network weights are shared for all tasks, and the CTC objective is emphasized solely for ASR and AED, for which sequential label information is available. The decoder is initialized with a task-specific start symbol, which indicates the task of interest and controls the set of labels that are recognized by the decoder, as shown in Fig. 1. Label symbols are tagged using a task identifier, i.e., similar labels are not shared across tasks for simplicity reasons.

The Transformer model leverages two different attention types: encoder-decoder attention and self-attention [12] that are both based on the scaled dot-product attention mechanism,

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^{\intercal}}{\sqrt{d_k}}\right)V, \qquad (1)$$

where $Q \in \mathbb{R}^{n_q \times d_q}$, $K \in \mathbb{R}^{n_k \times d_k}$, and $V \in \mathbb{R}^{n_v \times d_v}$ are the queries, keys, and values, where the d_* denote dimensions,

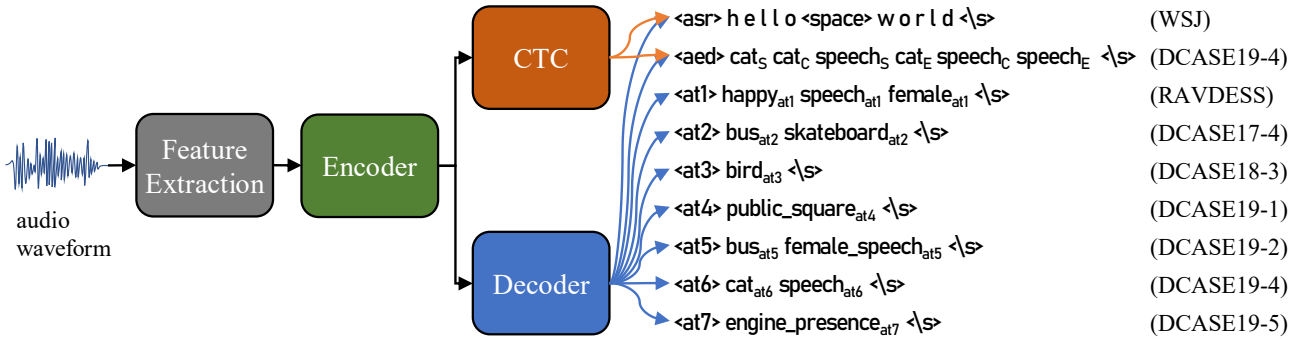

Figure 1: *The AIO Transformer system with an example output for each task that is switched by initially feeding a different start-of-task token to the decoder, shown in angle brackets ($\langle asr\rangle, \langle aed\rangle, \langle at1\rangle, \ldots, \langle at7\rangle$). $\langle\backslash s\rangle$ denotes the stop symbol for decoding, and the label suffixes $_S$, $_E$, and $_C$ denote start and end boundaries as well as continuation of an event. The ASR and AED tasks are trained and decoded jointly with CTC, whereas AT tasks use only the decoder output.*

the n_* denote sequence lengths, $d_q = d_k$, and $n_k = n_v$ [12]. Instead of using a single attention head, multiple attention heads are used by each layer of the Transformer model with

$$\mathrm{MHA}(\hat{Q}, \hat{K}, \hat{V}) = \mathrm{Concat}(\mathrm{Head}_1, \ldots, \mathrm{Head}_{d_h})W^H \quad (2)$$

$$\text{and } \mathrm{Head}_i = \mathrm{Attention}(\hat{Q}W_i^Q, \hat{K}W_i^K, \hat{V}W_i^V), \quad (3)$$

where $\hat{Q}$, $\hat{K}$, and $\hat{V}$ are inputs to the multi-head attention (MHA) layer, Head_i represents the output of the i-th attention head for a total number of d_h heads, and $W_i^Q \in \mathbb{R}^{d_{\mathrm{model}} \times d_q}$, $W_i^K \in \mathbb{R}^{d_{\mathrm{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\mathrm{model}} \times d_v}$ as well as $W^H \in \mathbb{R}^{d_h d_v \times d_{\mathrm{model}}}$ are trainable weight matrices that typically satisfy $d_k = d_v = d_{\mathrm{model}}/d_h$.

The encoder of our Transformer architecture consists of a two-layer CNN module EncCNN and a stack of E Transformer encoder layers with self-attention EncSA:

$$X_0 = \textsc{EncCNN}(X), \quad (4)$$

$$X_E = \textsc{EncSA}(X_0), \quad (5)$$

where $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ denotes a sequence of acoustic input features, which are 80-dimensional log mel-spectral energies (LMSEs) plus 3 extra features for pitch information [18]. Both CNN layers of EncCNN use a stride of size 2, a kernel size of 3×3, and a ReLU activation function, which reduces the frame rate of output sequence X_0 by a factor of 4. The EncSA module of (5) consists of E layers, where the e-th layer, for $e = 1, \ldots, E$, is a composite of a multi-head self-attention layer and two ReLU-separated feed-forward neural networks of inner dimension d_{ff} and outer dimension d_{model}:

$$X'_e = X_{e-1} + \mathrm{MHA}_e(X_{e-1}, X_{e-1}, X_{e-1}), \quad (6)$$

$$X_e = X'_e + \mathrm{FF}_e(X'_e), \quad (7)$$

$$\mathrm{FF}_e(X'_e) = \mathrm{ReLU}(X'_e W_{e,1}^{\mathrm{ff}} + b_{e,1}^{\mathrm{ff}})W_{e,2}^{\mathrm{ff}} + b_{e,2}^{\mathrm{ff}}, \quad (8)$$

where $W_{e,1}^{\mathrm{ff}} \in \mathbb{R}^{d_{\mathrm{model}} \times d_{\mathrm{ff}}}$, $W_{e,2}^{\mathrm{ff}} \in \mathbb{R}^{d_{\mathrm{ff}} \times d_{\mathrm{model}}}$, $b_{e,1}^{\mathrm{ff}} \in \mathbb{R}^{d_{\mathrm{ff}}}$, and $b_{e,2}^{\mathrm{ff}} \in \mathbb{R}^{d_{\mathrm{model}}}$ are trainable weight matrices and bias vectors.

The Transformer objective function is defined as

$$p_{\mathrm{att}}(Y|X_E) = \prod_{l=1}^{L} p(y_l|\boldsymbol{y}_{1:l-1}, X_E) \quad (9)$$

with label sequence $Y = (y_1, \ldots, y_L)$, label subsequence $\boldsymbol{y}_{1:l-1} = (y_1, \ldots, y_{l-1})$, and the encoder output sequence X_E. The term $p(y_l|\boldsymbol{y}_{1:l-1}, X_E)$ represents the Transformer decoder model, which can be written as

$$p(y_l|\boldsymbol{y}_{1:l-1}, X_E) = \textsc{Dec}(X_E, \boldsymbol{y}_{1:l-1}), \quad (10)$$

with

$$\boldsymbol{z}_{1:l}^0 = \textsc{Embed}(\langle s\rangle_\theta, y_1, \ldots, y_{l-1}), \quad (11)$$

$$\overline{\boldsymbol{z}}_l^d = \boldsymbol{z}_l^{d-1} + \mathrm{MHA}_d^{\mathrm{self}}(\boldsymbol{z}_l^{d-1}, \boldsymbol{z}_{1:l}^{d-1}, \boldsymbol{z}_{1:l}^{d-1}), \quad (12)$$

$$\overline{\overline{\boldsymbol{z}}}_l^d = \overline{\boldsymbol{z}}_l^d + \mathrm{MHA}_d^{\mathrm{dec}}(\overline{\boldsymbol{z}}_l^d, X_E, X_E), \quad (13)$$

$$\boldsymbol{z}_l^d = \overline{\overline{\boldsymbol{z}}}_l^d + \mathrm{FF}_d(\overline{\overline{\boldsymbol{z}}}_l^d), \quad (14)$$

for $d = 1, \ldots, D$, where D denotes the number of decoder layers. Function Embed converts the input label sequence $(\langle s\rangle_\theta, y_1, \ldots, y_{l-1})$ into a sequence of trainable embedding vectors $\boldsymbol{z}_{1:l}^0$, where $\langle s\rangle_\theta \in \Theta$ denotes a task specific start symbol using θ to index sequence $\Theta = (\langle \mathrm{asr}\rangle, \langle \mathrm{aed}\rangle, \langle \mathrm{at1}\rangle, \ldots, \langle \mathrm{at7}\rangle)$, as shown in Fig. 1. Function Dec finally predicts the posterior probability of label y_l by applying a fully-connected neural network to $\boldsymbol{z}_l^D$ and a softmax distribution over that output. Sinusoidal positional encodings are conventionally added to the sequences X_0 and Z_0 [12].

For the ASR and AED tasks, the Transformer model is trained jointly with the CTC objective function

$$p_{\mathrm{ctc}}(Y|X_E) = \sum_{\pi \in \mathcal{B}^{-1}(Y)} p(\pi|X_E), \quad (15)$$

where $\mathcal{B}^{-1}$ denotes a one-to-many map to expand the label sequence Y to a set of all possible frame-level label sequences using the CTC transition rules [19]. π represents a frame-level label sequence. The multi-objective loss function

$$\mathcal{L} = -\gamma \log p_{\mathrm{ctc}} - (1-\gamma) \log p_{\mathrm{att}} \quad (16)$$

is used for training, where hyperparameter γ is used to control the weighting between the two objective functions p_{ctc} and p_{att}.

3. Experimental Setup

Parameter settings of the Transformer model are $d_{\mathrm{model}} = 256$, $d_{\mathrm{ff}} = 2048$, $d_h = 4$, $E = 12$, and $D = 6$. The Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$ and learning rate scheduling similar to [12] is applied for training using 25000 warmup steps. The initial learning rate is set to 5.0 and the number of training epochs amounts to 80. Weight factor γ, which is used to balance the CTC and Transformer model objectives during training, is set to 0.3 for a batch of ASR samples, 0.4 for a batch of AED samples, and to 0.0 otherwise. The same weights are used for decoding as well. Layer normalization is applied before and dropout with a rate of 10 % after each MHA and FF layer. In addition, label smoothing with a penalty of 0.1 is used. For ASR inference, a word-level long short-term memory (LSTM) based language model (LM) [20] is applied via shallow

fusion using an LM weight of 1.0. For the AED task, temporal information for the recognized acoustic event sequence is obtained by using CTC-based forced alignment [21, 22].

3.1. Data Sets

In this work, 8 different data sets are employed, see Table 1. For ASR, the Wall Street Journal (WSJ) corpus of read English newspapers is utilized. For multi-condition ASR training (*indicated by superscript "a"*) and testing, a noisy training and test set is generated by mixing the WSJ training data with the DCASE training data sets of Table 1 and the eval92 test set with DEMAND [23] and NOISEX-92 [24] noise data using a signal-to-noise ratio (SNR) of 5 dB for training and 10 dB for testing.

For AT, we use various data sets of the recent DCASE challenges as well as the RAVDESS corpus of emotional speech and song [25]. The DCASE corpora used are DCASE 2017 task 4 (DCASE17-4) "large-scale weakly supervised sound event detection for smart cars" [5], DCASE 2018 task 3 (DCASE18-3) "bird audio detection" [6], DCASE 2019 task 1 (DCASE19-1) "acoustic scene classification" [7], DCASE 2019 task 2 (DCASE19-2) "audio tagging with noisy labels and minimal supervision" [8], DCASE 2019 task 4 (DCASE19-4) "sound event detection in domestic environments" [9], and DCASE 2019 task 5 (DCASE19-5) "urban sound tagging" [10]. In this work, we only use the coarse-level label information of DCASE19-5. The DCASE18-3 and RAVDESS corpora do not provide development and test data sets with ground truth annotations. We have therefore created DCASE18-3 development and test data sets, each consisting of 5 % of data randomly sampled from the DCASE18-3 training data set, which results in training, development, test data sets of size 89h, 5h, and 5h, respectively. The RAVDESS corpus features 24 actors speaking and singing; we use the audio recordings of the first 20 actors for training, those of actors #21 and #22 for development, and those of actors #23 and #24 for testing. Labels used for the RAVDESS data sets are those from the "emotion", "vocal channel", and "gender" information, e.g., "calm song male" or "disgust speech female". Since the DCASE19-1 data does not provide ground truth annotations for the official test data, we use only 10 % of the official development data for validation and the full set for testing. For multi-condition AT training (*indicated by superscript "b"*) and testing, we mixed the DCASE19-2 training and development data with speech recordings from the WSJ training and evaluation data sets using an SNR of approximately 0 dB.

For AED, the synthetic training data set of DCASE19-4 is used, which is the only AT training data set with strong annotations, i.e., where timing information is available. However, we are not utilizing the exact timing information for training but instead only the sequential label information. In order to handle overlapping events and for evaluating the timing information, each acoustic event is split into three event labels indicating the start position, continuation, and end position, indicated by subscripts $_S$, $_C$, and $_E$, respectively. Continuation labels are repeated every second depending on the duration of an event.

For the experiments with our Transformer model, all data sets are resampled to 16 kHz.

3.2. Baseline Systems

The presented baseline results are generated using the official baseline systems provided for the respective DCASE challenge tasks. The DCASE19-1 baseline system first extracts 40-dimensional LMSE features from 48 kHz sampled audio data and applies two CNN layers followed by a fully connected neural network layer and a softmax layer [7]; inference is based on the maximum output of the softmax layer. The DCASE19-2 baseline system uses a MobileNet v1 type of neural network architecture, which consists of a CNN layer followed by 13 separable CNN layers including a pooling layer for each and finally an 80-way logistic classifier layer [8]; 96-dimensional LMSE features extracted from 44.1 kHz sampled audio recordings are used as input to the network. The DCASE19-4 baseline system is based on the winning system of the previous year DCASE challenge, which is a mean-teacher model with context-gating CNN and RNN to maximize the use of unlabeled and weakly labeled data [26]; as an input to the neural network, 128-dimensional LMSE features are extracted from 22,050 Hz sampled audio data. The baseline system of DCASE19-5 is based on a VGGish neural network setting to extract 128-dimensional embeddings for classification [10, 27]; the network is fed with 64-dimensional LMSE features, which are extracted from 16 kHz sampled audio data. The DCASE17-4 baseline system consists of two 50-dimensional densely connected layers with 20 % dropout for each and a final output layer with sigmoid units; five consecutive frames of 40-dimensional LMSE features are used as input.

Table 1: *Summary of all the data sets used, where the number of hours per data set is shown in brackets. #C denotes the number of classes per task.*

Corpus	Task	#C	Train data	Dev. data	Test data
WSJ	ASR	49	train_si284 (81h)	dev93 (1.1h)	eval92 (0.7h)
DCASE17-4	AT	19	train (140h)	dev-test (1.3h)	eval (3h)
DCASE18-3	AT	2	train (99h)	n/a	n/a
DCASE19-1	AT	10	train (25.5h)	dev-eval (11.6h)	n/a
DCASE19-2	AT	80	train (10.5h), noisy (80.3h)	public (3.1h)	private (9.8h)
DCASE19-4	AT	10	synthetic (5.7h), weak (4.1h)	public (2.9h)	validate (1.9h)
DCASE19-4	AED	10*3	synthetic (5.7h)	public (2.9h)	validate (1.9h)
DCASE19-5	AT	8	train (4.4h)	validate (1.2h)	test (0.7h)
RAVDESS	AT	12	train (2.8h)	n/a	n/a

3.3. Evaluation Metrics

The ASR performance is measured using word error rates (WERs). For the AT tasks, we use micro-averaged F1-scores to determine the systems' accuracy. The AED systems' performance is assessed by the macro-averaged event-based F1-score measure using a 200 ms collar for both onsets and offsets, as well as by the macro-averaged segment-based F1-score measure using a segment length of 1 second [9, 28].

4. Results

Table 2 shows the F1-scores for the AT experiments using the DCASE challenge baseline systems and our proposed Transformer architecture with different training configurations. The "single" and "multi" AT training configuration denote that the Transformer is trained separately for each individual AT task using the respective training data only or for all AT tasks combined. Check marks for "ASR" and "AED" indicate that the model is trained with ASR and AED data, respectively. The AIO Transformer is jointly trained for all tasks, whereby superscript "a" denotes use of multi-condition ASR and superscript "b" the use of multi-condition AT training data, which are described in Section 3.1. The results show that multi-task training improves F1-scores on average for each task. Only small improvements can be seen for DCASE19-1 and RAVDESS, while for all other task, scores are considerably increased by more than 5 %. The proposed Transformer model clearly outperforms the baseline results for all tasks, except for DCASE19-1, where results are about the same.

Table 2: *Micro-averaged F1-scores [%] for the different audio tagging tasks. Baseline results are obtained using the baseline systems provided for the corresponding DCASE challenge tasks. For AT training data, "single" indicates the data for the corresponding single task, while "multi" indicates the data for all tasks. Superscripts "a" and "b" denote the usage of the multi-condition ASR and DCASE19-2 training data set, respectively.*

				DCASE19							DCASE18		DCASE17		RAVDESS	
	Training data			Task 1	Task 2		Task 4		Task 5		Task 3		Task 4			
System	AT	AED	ASR	dev	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Baseline systems	single			62.5	39.8	38.8	71.4	66.8	73.0	68.9	n/a	n/a	19.0	29.3	n/a	n/a
Transformer	single			59.2	45.3	46.0	71.9	71.0	73.7	70.9	83.6	84.1	45.4	51.6	**89.4**	86.1
Transformer	single		✓	60.4	48.2	47.4	74.2	72.7	74.9	69.9	89.1	89.2	50.4	55.2	83.0	85.7
Transformer	multi			**63.7**	45.0	46.5	74.7	71.8	77.2	73.3	88.3	88.2	46.7	52.9	84.6	**87.5**
Transformer	multi	✓		62.3	47.4	45.2	78.5	73.8	76.9	73.4	88.3	89.4	45.8	52.6	84.8	86.5
Transformer	multi		✓	62.4	48.8	49.3	78.7	77.4	77.5	74.6	87.7	88.1	49.1	56.6	82.4	83.5
AIO Transformer	multi	✓	✓	62.9	46.7	48.8	79.6	76.0	**79.0**	76.6	88.2	89.1	49.6	56.9	87.7	86.1
AIO Transformer	multi	✓	✓[a]	61.3	50.8	51.5	**81.1**	**78.7**	76.2	**77.7**	89.0	**89.9**	**51.0**	**58.2**	82.4	84.3
AIO Transformer	multi[b]	✓	✓[a]	61.6	**52.7**	**53.8**	79.8	78.2	74.9	74.2	**89.5**	89.5	50.7	56.0	85.3	87.3

Table 3: *WSJ-based ASR results for a CTC-Transformer based baseline system as well as for our proposed multi-task Transformer models.*

	Training data			WER [%]		
System	AT	AED	ASR	dev	test	10 dB
Transformer (Baseline)			✓	7.7	5.0	10.9
Transformer (Baseline)			✓[a]	7.6	4.8	5.4
Transformer		✓	✓	7.8	5.0	11.3
Transformer		✓	✓[a]	7.9	4.7	5.5
Transformer	multi		✓	8.0	5.3	14.4
AIO Transformer	multi	✓	✓	7.5	5.1	12.5
AIO Transformer	multi	✓	✓[a]	7.7	5.2	6.3
AIO Transformer	multi[b]	✓	✓[a]	7.8	5.3	5.8

Table 4: *AED results for the DCASE 2019 task 4 baseline system as well as for our proposed multi-task Transformer systems.*

				F1-scores [%]			
	Training data			Event-based		Segment-based	
System	AT	AED	ASR	dev	test	dev	test
Baseline system		✓		29.0	24.0	58.5	54.8
Transformer		✓		16.0	10.6	43.8	34.8
Transformer		✓	✓	26.3	18.8	48.9	38.2
Transformer		✓	✓[a]	21.4	15.4	44.6	34.2
Transformer	multi	✓		11.0	8.4	51.9	44.3
AIO Transformer	multi	✓	✓	21.2	12.5	60.7	49.6
AIO Transformer	multi	✓	✓[a]	26.3	16.7	61.2	50.7
AIO Transformer	multi[b]	✓	✓[a]	23.8	14.9	62.0	51.4

Table 3 shows the ASR results of our multi-task trained Transformer models. The ASR baseline system is based on a CTC-Transformer architecture as well, using the same model parameters [16]. It can be noticed that WERs of the AIO Transformer are similar or only slightly higher compared to the baseline. For the noisy test set, the AIO model proves to be less noise robust compared to the baseline system, if both systems are trained using clean speech only. We suppose the reason for this is that the AIO Transformer has learned to maintain features for both "noise" and speech, while an ASR model would learn to extract speech features only and to ignore other cues. Hence, the AIO Transformer can easily be confused by other sound events, which can be avoided by multi-condition training as shown by the results in Table 3.

AED results for the DCASE19-4 task are shown in Table 4. Without the ASR task, the Transformer model did not learn the AED task well. We suppose the reason is that it is arduous to learn how to estimate the correct temporal ordering of events given only a small amount of AED training data but it can be learned from the larger ASR task by transfer learning. In addition, the AT data does not help to improve the event-based F1-scores, since the CTC objective is only applied to the AED and ASR tasks, and thus the CTC projection layer is not updated for every batch, which can lead to mismatches, especially since the AT data is larger than the ASR and AED data sets. Compared to the baseline, the event-based F1-scores obtained by the AIO Transformer are lower, which is likely partly due to the evaluation metric that uses a collar of only 200 ms. Hence, segment-based F1-scores obtained by the AIO Transformer, which are also improved by using AT data unlike the event-based scores, show that the Transformer system is competitive to the baseline system but the learned temporal alignment of events may be less accurate for the above mentioned reasons.

We showed above that the AIO transformer was able to robustly transcribe speech in the presence of interfering sound events. To show that the AIO Transformer can provide a total transcription of a complex acoustic scene, we now test the robustness of AT to interfering speakers as well. The multi-condition trained AIO Transformer achieves an F1-score of 63.1 % for the multi-condition DCASE19-2 test data, which is described in Section 3.1. Note that this score is higher than the scores reported in Table 2 as the multi-condition data features additional AT speech labels that can be detected.

5. Conclusions

In this work, we show that ASR, AED, and multiple AT tasks can be unified under a single Transformer-based system, whereby multi-task learning has been shown to improve the system's performance for each individual task. The proposed AIO Transformer model achieves competitive or better recognition scores compared to all baseline systems of recent DCASE challenges, as well as compared to an end-to-end ASR baseline system of similar architecture. The system's capability to simultaneously recognize speech and acoustic events is evaluated, and results demonstrate that it can be used to perform the total transcription of an audio signal, whereby efficient and simultaneous ASR, AED, and AT decoding can be achieved by batch processing similar to [29].

6. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, Dec. 2017.

[2] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. Engel, L. Fan, C. Fougner, T. Han, A. Y. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Y. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in English and Mandarin," *arXiv preprint arXiv:abs/1512.02595*, 2015.

[3] C. Spille, B. Kollmeier, and B. Meyer, "Comparing human and automatic speech recognition in simple and complex acoustic scenes," *Computer Speech & Language*, vol. 52, pp. 123–140, Apr. 2018.

[4] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," in *Proc. ISCA Interspeech*, Aug. 2017, pp. 132–136.

[5] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proc. of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Nov. 2017.

[6] D. Stowell, M. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, Oct. 2018.

[7] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Nov. 2018, pp. 9–13.

[8] E. Fonseca, M. Plakal, F. Font, D. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," in *Proc. of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Oct. 2019.

[9] N. Turpault, R. Serizel, J. Salamon, and A. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Oct. 2019, pp. 253–257.

[10] J. Bello, C. Silva, O. Nov, R. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: a system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, pp. 68–77, Jan. 2019.

[11] D. Peterson, V. Reddy, and R. Hamel, "Neuroanatomy, auditory pathway," *StatPearls*, Apr. 2020.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, Dec. 2017, pp. 6000–6010.

[13] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on transformer vs RNN in speech applications," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2019.

[14] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with CNN-transformer and automatic threshold optimization," *arXiv preprint arXiv:abs/1912.04761*, 2019.

[15] W. Boes and H. Van Hamme, "Audiovisual transformer architectures for large-scale classification and synchronization of weakly labeled audio events," in *Proc. of the ACM International Conference on Multimedia*, 2019, p. 1961–1969.

[16] S. Karita, N. Yalta, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *Proc. ISCA Interspeech*, Sep. 2019, pp. 1408–1412.

[17] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *J. Sel. Topics Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[18] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *Proc. ISCA Interspeech*, Aug. 2017, pp. 949–953.

[19] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. International Conference on Machine Learning (ICML)*, vol. 148, Jun. 2006, pp. 369–376.

[20] T. Hori, J. Cho, and S. Watanabe, "End-to-end speech recognition with word-based RNN language models," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2018, pp. 389–396.

[21] N. Moritz, T. Hori, and J. Le Roux, "Triggered attention for end-to-end speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 5666–5670.

[22] N. Moritz, T. Hori, and J. Le Roux, "Streaming end-to-end speech recognition with joint CTC-attention based models," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2019, pp. 936–943.

[23] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. of International Congress on Acoustics (ICA)*, vol. 19, no. 1. Acoustical Society of America, Jun. 2013.

[24] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, p. 247–251, Jul. 1993.

[25] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. 1–35, May 2018.

[26] L. JiaKai, "Mean teacher convolution system for DCASE 2018 task 4," in *Proc. of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Jul. 2018.

[27] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 131–135.

[28] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[29] H. Seki, T. Hori, S. Watanabe, N. Moritz, and J. Le Roux, "Vectorized beam search for CTC-attention-based speech recognition," in *Proc. ISCA Interspeech*, Sep. 2019, pp. 3825–3829.