CONTEXT-AWARE TRANSFORMER TRANSDUCER FOR SPEECH RECOGNITION

Feng-Ju Chang[*]*, Jing Liu*[*]*, Martin Radfar, Athanasios Mouchtaris,*
Maurizio Omologo, Ariya Rastrow, Siegfried Kunzmann

Amazon Alexa

{fengjc,jlmk,radfarmr,mouchta,omologo,arastrow,kunzman}@amazon.com

ABSTRACT

End-to-end (E2E) automatic speech recognition (ASR) systems often have difficulty recognizing uncommon words, that appear infrequently in the training data. One promising method, to improve the recognition accuracy on such rare words, is to latch onto personalized/contextual information at inference. In this work, we present a novel context-aware transformer transducer (CATT) network that improves the state-of-the-art transformer-based ASR system by taking advantage of such contextual signals. Specifically, we propose a multi-head attention-based context-biasing network, which is jointly trained with the rest of the ASR sub-networks. We explore different techniques to encode contextual data and to create the final attention context vectors. We also leverage both BLSTM and pretrained BERT based models to encode contextual data and guide the network training. Using an in-house far-field dataset, we show that CATT, using a BERT based context encoder, improves the word error rate of the baseline transformer transducer and outperforms an existing deep contextual model by 24.2% and 19.4% respectively.

***Index Terms*—** speech recognition, context-aware training, attention, transformer-transducers, BERT

1. INTRODUCTION

E2E ASR systems such as connectionist temporal classification (CTC) [1], listen-attend-spell (LAS) [2], recurrent neural network transducer (RNN-T) [3] and transformer [4] have gained significant interest due to their superior performance over hybrid HMM-DNN systems, as sufficient training data is available [5]. While hybrid models optimize the acoustic model (AM), pronunciation model (PM) and language model (LM) independently, E2E models implicitly subsume these modules and jointly optimize them to output word sequences [6, 7] directly given an input sequence. In addition, E2E models simplify the inference pipeline without external alignment modules and LMs, which make them more capable for on-device deployment [8].

However, one major limitation of an E2E ASR system is that it cannot accurately recognize words that appear rarely in the paired audio-to-text training data, such as entity names or personalized words [9–11]. To address this issue, previous work has leveraged context where rare words would appear more frequently or are associated with weights, e.g. the weighted finite state transducer (WFST) [12] constructed from the speaker's context [13], domain [14], text metadata of video [15, 16], dialogue state, location, or personalized information about the speaker (e.g., personalized device names or contact names) [9, 17], and so on.

In general, the methods that integrate the above context into E2E ASR systems can be divided into two categories: post-training integration [13, 17–24] and during-training integration [9, 15, 25]. The former is only applied in inference time while the latter occurs in both training and inference. Shallow fusion [18] is one of the dominant paradigms used to incorporate WFST via an independently-trained on-the-fly rescoring framework to adjust the LM weights of n-grams relevant to the particular context [19]. Another post-training integration method is deep fusion [20], which integrates the external LM into the E2E ASR by fusing together the hidden states of the external LM and the decoder. A major drawback of post-training integration, however, is that it requires external LMs to rescore the ASR model's outputs and it is sensitive to rescoring weights.

The most relevant work to ours in the during-training integration category is contextual LAS (C-LAS) [9], which proposes an additional bias encoder with a location-aware attention mechanism [26] on top of LAS [2] in order to rescore the personalized words at training and inference, with label embeddings. Similarly, contextual RNN-T (C-RNN-T) [15] applies the same attention mechanism but using RNN-T [3]. Phonetic information is explored in [11] and [27] to further improve C-LAS.

Witnessing the transformer network and its variant, transformer transducer [28–31], have become state-of-the-art ASR models, we propose a novel Context-Aware Transformer Transducer (CATT) network, enabling transformer transducer models to leverage contextual information both in training and

*Joint First Authors

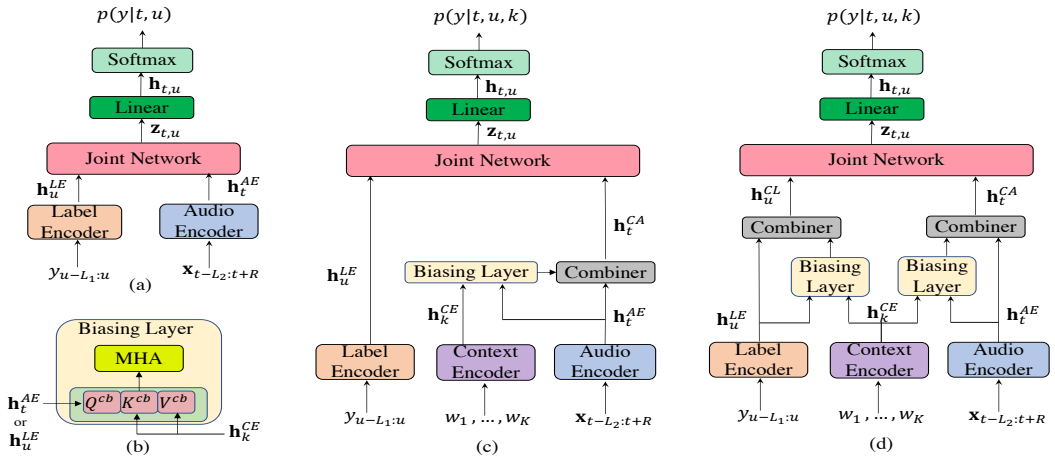

Fig. 1. (a) Transformer Transducer (b) The proposed context biasing layer via multi-head attention (MHA) between audio/label and context embeddings. The proposed context-aware transformer transducer (CATT) with (c) audio embeddings or (d) audio+label embeddings to attend the context embeddings

inference to improve ASR performance.

Different from C-LAS [9] and C-RNN-T [15], we encode the contextual data not only with BLSTM [32, 33], but also with a pre-trained BERT [34, 35] model, which brings strong semantic knowledge to guide the network at training and inference. In addition, we propose a multi-head attention based context biasing module in order to measure the importance of the contextual phrases. We employ audio embeddings alone or with label embeddings to measure the importance of context, and hence create the relevant context vectors that are fed into the ASR frame by frame to improve the alignment learning.

Using an in-house far-field dataset, we demonstrate that injecting context into audio embeddings in our CATT model outperforms the baseline methods with no context, shallow fusion [18], and C-LAS [9]. Moreover, employing a BERT based context encoder achieves the greatest improvements. Finally, we find that injecting context into both audio and label embeddings can further improve the CATT model over the one with context injected in audio embeddings only.

2. PROPOSED APPROACH

In this section, we describe the transformer transducer model and our design of the context-aware module.

2.1. Transformer Transducer

The transformer transducer model [29–31] outputs a probability distribution over output sequences $\mathbf{y}$ given input audio frames $\mathbf{x}$. The model contains three main modules (Fig. 1 (a)): an audio encoder, a label encoder, and a joint network.

The audio encoder, f^{enc}, is comprised of stacked self-attention transformer layers [28] that uses audio features $\mathbf{x}$ in a predefined window centered at frame t, [t-L_2:t+R], to produce the audio embedding $\mathbf{h}_t^{AE}$ at frame t,

$$\mathbf{h}_t^{AE} = f^{\text{enc}}(\mathbf{x}_{t-L_2:t+R}) \tag{1}$$

where $\mathbf{h}_t^{AE} \in \mathbb{R}^{d_a \times 1}$ and d_a is the embedding size. It is similar to the AM in a hybrid ASR system.

The label encoder, f^{pred}, is also a stacked transformer network, which uses the previous L_1 non-blank tokens y to generate the label embedding $\mathbf{h}_u^{LE}$. We use subwords as tokens.

$$\mathbf{h}_u^{LE} = f^{\text{pred}}(y_{u-L_1:u}) \tag{2}$$

where $\mathbf{h}_u^{LE} \in \mathbb{R}^{d_l \times 1}$ and d_l is the embedding size; f^{pred} acts similarly to the LM in a hybrid ASR system.

The joint network combines the audio encoder outputs and label encoder outputs to produce a new embedding:

$$\mathbf{z}_{t,u} = \phi(U\mathbf{h}_t^{AE} + V\mathbf{h}_u^{LE} + \mathbf{b}_1) \tag{3}$$

where U, V, $\mathbf{b}_1$ are learnable parameters to project the audio and label embeddings into the same dimension. ϕ is a non-linear function; we opt for $\tanh$ in this work. $\mathbf{z}_{t,u}$ is then fed into a linear layer and Softmax layer to produce a probability distribution over output labels plus a blank symbol.

$$\begin{aligned} \mathbf{h}_{t,u} &= W\mathbf{z}_{t,u} + \mathbf{b}_2 \\ p(y|t,u) &= \text{Softmax}(\mathbf{h}_{t,u}) \end{aligned} \tag{4}$$

where W and $\mathbf{b}_2$ are trainable parameters. When the joint network predicts a blank symbol, the model proceeds to the audio encoder ouput of the next time frame; when a non-blank symbol is predicted, the label encoder output will be updated. This results in various alignment paths, and the sum of probabilities of them provides the probability of an output sequence (with non-blank outputs) given an input sequence.

2.2. Context-Aware Transformer Transducer (CATT)

To inject context information, we modify the base transformer transducer described in Section 2.1 and add two additional components: (1) a context encoder, and (2) a multi-head attention-based context biasing layer, as shown in Fig. 1 (b).

Context Encoder: The context we employ in this work contains personalized information provided by the speakers, such as speaker-defined device names, device settings, and device locations as presented in Table 1. Each contextual word or phrase w_k is first represented as subwords and then fed into the context encoder f^{context}, to produce fixed dimensional vector representations.

$$\mathbf{h}_k^{CE} = f^{\text{context}}(w_k). \tag{5}$$

where $\mathbf{h}_k^{CE} \in \mathbb{R}^{d_c \times 1}$.
In particular, we investigate two types of context encoders, f^{context}:

- *BLSTM based context embedding*: The inputs for this embedding extractor are tokenized contextual phrases by a subword tokenizer [6]. Then, the last state of BLSTM is used as the embedding of a context word or phrase. Note that we train the BLSTM from scratch along with the rest of the networks.
- *BERT [34] based context embedding*: This context encoder is pre-trained with a substantial amount of text data. Using BERT, we bring strong semantic prior knowledge to guide the training of the rest of the networks in our model. Specifically, the SmallBERT [35] model is employed. We investigate the case when this encoder is frozen, and only the rest of our network weights are updated.

Note that the transformer transducer described in Section 2.1 exploits only the outputs of the audio encoder, $\mathbf{h}_t^{AE}$, and label encoder, $\mathbf{h}_u^{LE}$, to produce the probabilities over tokens, $p(y|t,u) = p(y|\mathbf{h}_t^{AE}, \mathbf{h}_u^{LE})$. In contrast, for our context-aware transformer transducer, the output probability is conditionally dependent on the contextual data as well. Namely, $p(y|t,u)$ becomes

$$p(y|t,u,k) = p(y|\mathbf{h}_t^{AE}, \mathbf{h}_u^{LE}, \mathbf{h}_k^{CE}). \tag{6}$$

Multi-Head Attention based Context Biasing Layer: Given the context embeddings, $\mathbf{h}_k^{CE}$, from Eqn.(5), this module is designed to learn context phrase relevance to an utterance. In this way, the model can pay more attention to the frames corresponding to the entity names or personalized words to help improve their prediction accuracy. Since our base model is transformer-based [28], multi-head attention (MHA) becomes a natural choice to learn the relationships between the context embeddings and the utterance's embeddings. We first investigate the audio embeddings as queries to attend context (See Fig. 1 (c)) because the audio encoder used in this work is bi-directional transformer (i.e. the attention is computed on both previous and future frames), which covers more information than the label encoder (uni-directional transformer) about the input utterance. We also investigate using both audio embeddings and label embeddings as queries to attend context, as shown in Fig. 1 (d).

For clarity, we will omit the MHA here and focus on how we integrate context and audio embeddings with the attention mechanism. The same formulation can be applied to label embeddings as well. Specifically, we create the query, key, and value as follows (Fig. 1 (b)):

$$\begin{aligned} Q^{cb} &= \sigma\left(XW^{cb,q} + \mathbf{1}(\mathbf{b}_i^{cb,q})^\top\right) \\ K^{cb} &= \sigma\left(CW^{cb,k} + \mathbf{1}(\mathbf{b}_i^{cb,k})^\top\right) \\ V^{cb} &= \sigma\left(CW^{cb,v} + \mathbf{1}(\mathbf{b}_i^{cb,v})^\top\right) \end{aligned} \tag{7}$$

Here, $X = [\mathbf{h}_1^{AE}, \ldots, \mathbf{h}_{\mathcal{T}}^{AE}]^\top \in \mathbb{R}^{\mathcal{T} \times d_a}$ is the audio encoder outputs of an utterance and $\mathcal{T}$ is the number of audio frames. $C = [\mathbf{h}_1^{CE}, \ldots, \mathbf{h}_K^{CE}]^\top \in \mathbb{R}^{K \times d_c}$ is the associated context embeddings, where K is the number of context phrases. $\sigma(\cdot)$ is an activation function. $W^{cb,q} \in \mathbb{R}^{d_a \times d}$, $W^{cb,k} \in \mathbb{R}^{d_c \times d}$, $W^{cb,v} \in \mathbb{R}^{d_c \times d}$, and $\mathbf{b}^{cb,*} \in \mathbb{R}^{d \times 1}$ are all learnable weight and bias parameters, and $\mathbf{1} \in \mathbb{R}^{\mathcal{T} \times 1}$ is an all-ones vector.

The cross-attention between audio embeddings and context embeddings is then computed by

$$H^{cb} = \text{Softmax}\left(\frac{Q^{cb}(K^{cb})^\top}{\sqrt{d}}\right)V^{cb}, \tag{8}$$

where the scaling $\frac{1}{\sqrt{d}}$ is for numerical stability [28]. The resulting attention scores (Softmax part of Eqn.(8)) are used to take the weighted sum of context embeddings resulting in the context-aware matrix $H^{cb} \in \mathbb{R}^{\mathcal{T} \times d}$. This context-aware matrix is then fused with the audio embedding matrix through a combiner, which consists of LayerNorm layers, a concatenation and a feed-forward projection layer in order to match the dimension of the label encoder outputs (Fig. 1 (c)). This context integration process can be described as

$$\begin{aligned} H^{concat} &= [\text{LayerNorm}(X), \text{LayerNorm}(H^{cb})] \\ H^{CA} &= \text{FeedForward}(H^{concat}), \end{aligned} \tag{9}$$

where $H^{concat} \in \mathbb{R}^{\mathcal{T} \times (d_a + d)}$ and $H^{CA} = [\mathbf{h}_1^{CA}, \ldots, \mathbf{h}_{\mathcal{T}}^{CA}] \in \mathbb{R}^{\mathcal{T} \times d_{ca}}$. We finally feed the context-aware audio embeddings along with the label embeddings into the joint network. Similarly, when we use the label embeddings to attend context, the inputs used to compute Q^{cb} become the label encoder outputs $Y = [\mathbf{h}_1^{LE}, \ldots, \mathbf{h}_{\mathcal{U}}^{LE}]^\top \in \mathbb{R}^{\mathcal{U} \times d_l}$ of an utterance and $\mathcal{U}$ is the number of tokens that have been predicted.

Fig. 1 (d) illustrates how we use both audio embeddings and label embeddings to attend the context. Specifically, we have two biasing layers, one taking the audio embeddings as queries, and the other taking the label embeddings as queries to

Table 1. The types and examples of contextual words/phrases

Context Type	Example Contextual Phrases
Personalized Device Name	baine's room, grogu's second tv
Named entity	foxtel, xbox
Device Setting	turn on, off, dim
Device Location	living room, basement, hallway

evaluate the importance of context. Each of the context-aware matrices from biasing layers will be fused with the original audio or label embeddings via a combiner (Eqn.(9)). The context-injected audio ($\mathbf{h}_t^{CA}$) and label ($\mathbf{h}_u^{CL}$) embeddings are then fed into the joint network for alignment learning.

3. EXPERIMENTAL SETUP

3.1. Dataset

To evaluate the proposed CATT model, we use 1,300 hours of speech utterances from our in-house de-identified far-field dataset. The amount of training set, dev set, and test set are 910 hours, 195 hours, and 195 hours respectively. We train our model on the train split and report the results on both dev and test splits. The device-directed speech data is captured using a smart speaker, and the contextual words/phrases are provided by the speakers consisting of personalized device names, entity names, device settings, and device locations, as shown in Table 1.

We denote the contextual phrases which are present in the transcription as relevant context, and those that are absent in the transcription as less relevant context. During training in each batch, we use all relevant contextual phrases and cap the maximum number of contextual entities at K=100. If the number of relevant contextual phrases is less than 100, we randomly add less relevant context until we reach a total of 100. In this way, every training utterance would be paired with a variety of context in each training step. This samplings allows flexibility at inference, as the model does not make any assumption about what contextual phrases will be used. Note that we assume the relevant contextual phrases are always present during inference.

During evaluation, we further divide the dev and test sets into two categories, *Personalized* and *Common* sets. An utterance is labeled as *Personalized* if there is any word in the transcription that also appears in the contextual phrase list of personalized device names or entity names. Otherwise, it is categorized as *Common*[1]. Example utterances are illustrated in Table 2 and we present results on these two sets in Section 4.

[1]The utterances in the *Common* set may contain contextual phrases like the device location or/and device setting.

Table 2. Example data in *Personalized* and *Common* sets

Data Split	Example
Personalized	turn off baine's second TV
Common[1]	dim living room light thirty percent

Table 3. WERRs for 14M CATT vs.T-T and SF

	Personalized		*Common*	
WERR (%)	dev	test	dev	test
T-T	0	0	0	0
T-T + SF	2.9	1.6	1.0	2.2
CATT (BLSTM-CE)	1.4	1.6	2.1	2.2
CATT (BLSTM-CE) + SF	2.9	3.1	4.1	4.3
CATT (SmallBERT-CE)	4.3	4.7	4.1	4.3
CATT (SmallBERT-CE) + SF	5.7	6.3	5.2	6.5

3.2. Model Configurations

The proposed CATT, has the following configurations. The audio encoder is a 12-layer transformer (in the 14-million-parameter model case, denoted as 14M below) or a 20-layer transformer (in the 22-million-parameter model case, denoted as 22M below). The label encoder is a 4-layer transformer (in 14M case) or an 18-layer transformer (in 22M case). Both the audio and label encoders have an embedding size of 256 (d_a and d_l in Section 2.1) with 4 attention heads. The joint network is a fully-connected feed-forward component with one hidden layer followed by a $\tanh$ activation function.

We explore two context embedding extractors: (1) 1-layer and 2-layer BLSTM of 256 dimensions (d_c in Section 2.2) for the 14M model and 22M model, respectively; (2) 4-layer, 14M SmallBERT [35] of 256 dimensions. The context biasing encoder consists of a 4 transformer blocks with an embedding size of 256 ($d = d_{ca}$ in Section 2.2) and 4 attention heads; all of the above transformer blocks also contain layer normalizations, feed-forward layers after attention layers and residual links in between layers; there are 512 neurons for the feed-forward layers. The baseline models, i.e. transformer transducers, do not have context embedding extractors or context biasing encoders. We train baselines and CATT models from scratch, except for the SmallBERT context encoder.

The input audio features fed into the network consists of 64-dimensional LFBE features, which are extracted every 10 ms with a window size of 25 ms from audio samples. The features of each frame is then stacked with the left two frames, followed by a downsampling of factor 3 to achieve low frame rate, resulting in 192 feature dimensions. The subword tokenizer [6, 7] is used to create tokens from the transcriptions; we use 4000 subwords/tokens in total. We trained the models by minimizing the alignment loss [3] using the Adam optimizer [36], and varied the learning rate following [4, 28].

4. RESULTS AND DISCUSSION

4.1. Comparisons to the Non-contextual Baselines and Shallow Fusion

The results of all experiments are demonstrated as the relative word error rate reduction (WERR). Given a model A's WER (WER_A) and a baseline B's WER (WER_B), the WERR of A over B is computed as $\text{WERR} = (\text{WER}_B - \text{WER}_A)/\text{WER}_B$. In the following experiments, we use a vanilla transformer transducer (T-T) as our baseline. We denote shallow fusion and context embedding as SF and CE, respectively. We compare the proposed CATT with the following models: (1) T-T, (2) T-T + SF, (3) CATT (BLSTM-CE), (4) CATT (BLSTM-CE) + SF, (5) CATT (SmallBERT-CE), (6) CATT (SmallBERT-CE) + SF. We freeze the SmallBERT model during training, and shallow fusion is only applied during inference time.

Table 3 presents the WERRs of the proposed CATT over the baselines. We can see that CATT outperforms both T-T and SF on both *Personalized* and *Common* sets. The reason CATT also improves the WERs on *Common* sets is that contextual phrases (e.g. the device location and device setting) may appear in the utterances of the *Common* set. Specifically, we can see that CATT with BLSTM context embeddings outperforms the vanilla transformer transducer by up to 2.2%. If we leverage a pre-trained SmallBERT as the context encoder, the CATT model outperforms shallow fusion with 4.7% vs 1.6% WERR on *Personalized* test sets. Combining CATT and SF further achieves 6.3% and 6.5% WERR on both *Personalized* and *Common* test sets.

Note that the SmallBERT encoder has around 14M non-trainable parameters, which is about the same size as the trainable part of the baseline model in Table 3. To fully take advantage of the strong semantic prior knowledge of BERT, we use a larger baseline T-T in Table 5 with 22M trainable parameters. We get 10.6% and 10.3% WERR on *Personalized* and *Common* test sets for CATT with BLSTM context embeddings. CATT with the SmallBERT context embeddings further improves the WERR to 15.3% on *Personalized* test sets. Besides, CATT performs better than shallow fusion with 15.3% vs 3% WERR on *Personalized* test sets. One drawback of shallow fusion is that it is very sensitive to the boosting weights leading to over boosting and degrades the WERs. Instead, CATT consistently improves WERs over the baselines.

Two example outputs generated by T-T and CATT are shown in Table 4. We also visualize the attention values in Fig. 2, to better understand how the CATT attention mechanism works. As can be seen, the CATT attends to the entities and give them higher attention weights.

4.2. Comparisons of Different Queries to Attend Context

Given the best performed results of CATT with SmallBERT context embeddings in the last section, we take this model and further evaluate it by using the audio embedding as queries (Fig. 1 (c)) and both audio and label embeddings as queries (Fig. 1 (d)) to attend the context. They are denoted as audio-Q and audio+label-Q respectively as follows. In Table 6 and Table 7, we see consistent improvements across different model sizes when using both audio embeddings and label embeddings to attend context. We believe that context embeddings have a correlation with both audio and label embeddings. In addition, we also hypothesize that better alignment can be learned between input audio and output token predictions when both embeddings are calibrated via context.

Table 4. Comparing outputs generated by the T-T baseline and CATT. Entity names and personalized device names are highlighted with bold font in these examples.

Method	**Output**
Reference Utterance	turn on **baine's** room
T-T	turn on basement room
CATT	turn on **baine's** room
Reference Utterance	turn the **foxtel** lights green
T-T	turn the fox lights green
CATT	turn the **foxtel** lights green

Table 5. WERRs for 22M CATT vs.T-T and SF

	Personalized		*Common*	
WERR (%)	dev	test	dev	test
T-T	0	0	0	0
T-T + SF	4.1	3.0	2.0	2.1
CATT (BLSTM-CE)	11.0	10.6	10.8	10.3
CATT (BLSTM-CE) + SF	12.3	10.8	12.8	12.4
CATT (SmallBERT-CE)	15.1	13.6	13.7	12.4
CATT (SmallBERT-CE) + SF	16.6	15.3	14.7	13.4

4.3. CATT with Varying Number of Contextual Entities

As mentioned in Section 3.1, we randomly add less relevant contexts in addition to the relevant ones for the model robustness to the irrelevant contexts. To verify it, we evaluate the best-performing CATT in Table 7 with different number of contextual phrases, K, chosen randomly from the complete context list, and we assume the target context is present in the list. As can be seen in Fig. 3, the WERRs of CATT over T-T remain across different test sets with different values of K.

4.4. Comparisons to Contextual LAS

Finally, we compare the proposed CATT with an existing deep contextual model – Contextual LAS (C-LAS) [9], which was also designed to bias towards the personalized words during training and inference. We follow [9] to set up the configurations of audio encoder, context encoder, and the attention module, except that the number of hidden units are adjusted to match our model sizes (14M and 22M parameters). The details of C-LAS configuration we used are described below (The number before "/" is for 14M model while the number after "/" is for 22M case): The audio encoder's architecture

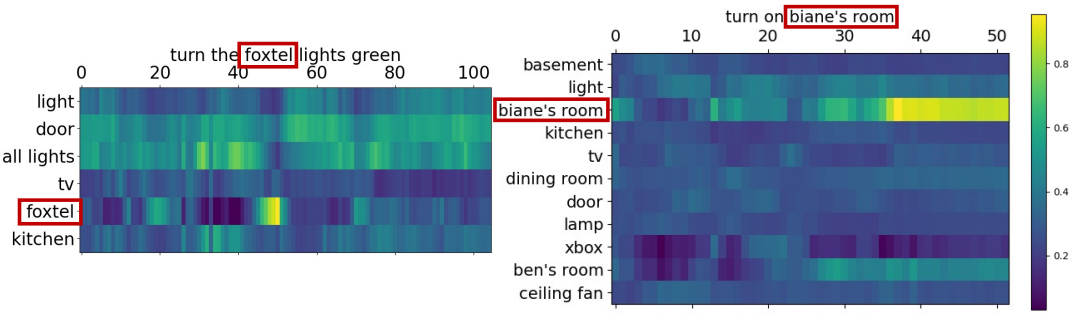

Fig. 2. Visualizing attention weights from Eqn.(8) for two examples with entity names which are marked by red-colored rectangles. The x-axis shows the audio frame indices and the corresponding predictions, and the y-axis shows the context phrases associated with the example. While the darker colors represent values close to zero, the brighter ones represent higher weights.

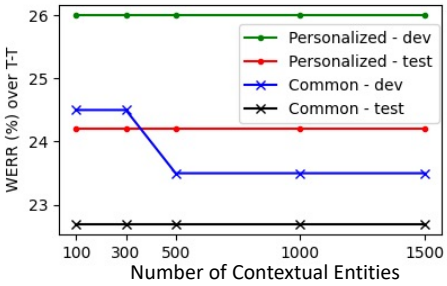

Fig. 3. WERRs (%) for CATT over T-T with different number of contextual entities K

Table 6. WERRs for 14M CATT with audio query and audio+label query

	Personalized		*Common*	
WERR (%)	dev	test	dev	test
T-T	0	0	0	0
CATT(SmallBERT-CE),audio-Q	4.3	4.7	4.1	4.3
CATT(SmallBERT-CE),audio+label-Q	11.4	10.9	9.3	9.7

consists of 10 BLSTM layers, each with 128 / 224 nodes. The audio encoder attention is computed over 256 / 64 dimensions, using 4 attention heads. The bias-encoder consists of a single BLSTM layer with 256 nodes, and the bias-attention is computed over 256 / 64 dimensions. Finally, the decoder consists of 4 LSTM layers with 256 / 128 nodes. The WERR results of CATT over C-LAS are presented in Table 8 and Table 9 for 14M and 22M parameters individually.

With the same type of context encoder, a BLSTM, the proposed CATT has shown $\sim$ 3-6% relative WER improvements. The CATT can be further improved by exploiting the BERT based context embedding and achieve up to 8.1% relative improvement in 22M case with an audio query. Again, CATT with both audio and label embeddings as queries (audio+label-Q) performs the best (by up to 21.7% relative improvement against C-LAS for 22M model).

Table 7. WERRs for 22M CATT with audio query and audio+label query

	Personalized		*Common*	
WERR (%)	dev	test	dev	test
T-T	0	0	0	0
CATT(SmallBERT-CE),audio-Q	15.1	13.6	13.7	12.4
CATT(SmallBERT-CE),audio+label-Q	26.0	24.2	24.5	22.7

Table 8. WERRs for 14M CATT vs.C-LAS

	Personalized		*Common*	
WERR (%)	dev	test	dev	test
C-LAS [9]	0	0	0	0
CATT(BLSTM-CE),audio-Q	4.2	4.6	4.0	3.2
CATT(SmallBERT-CE),audio-Q	6.9	7.6	6.1	5.3
CATT(SmallBERT-CE),audio+label-Q	13.9	13.6	11.1	10.6

Table 9. WERRs for 22M CATT vs.C-LAS

	Personalized		*Common*	
WERR (%)	dev	test	dev	test
C-LAS [9]	0	0	0	0
CATT(BLSTM-CE),audio-Q	5.8	4.8	2.2	1.1
CATT(SmallBERT-CE),audio-Q	10.1	8.1	5.4	3.4
CATT(SmallBERT-CE),audio+label-Q	21.7	19.4	17.2	14.8

5. CONCLUSION

We proposed a novel CATT model to enable a state-of-the-art transformer transducer ASR model to take advantage of the contextual data both in training and inference. We leveraged both BLSTM and BERT-based models to encode context. The relevance of context was measured by the proposed multi-head cross-attention mechanism with audio embeddings alone or together with label embeddings. Our experiments showed that CATT outperformed the non-contextual models, shallow fusion, and C-LAS models on an in-house dataset with a variety of named entities and personalized information.

Acknowledgement We thank Christian Hensel, Zhe Zhang, G. Tiwari, Grant P. Strimel, Thejaswi Muniyappa, Kanthashree Sathyendra, Kai Wei, Markus Mueller and Rupak Swaminathan for feedbacks.

6. REFERENCES

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[2] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[3] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[4] Linhao Dong, S. Xu, and Bo Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," *ICASSP*, 2018.

[5] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *ICASSP*, 2018.

[6] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *ACL*, 2016.

[7] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[8] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. Chang, K. Rao, and A. Gruenstein, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP*, 2019.

[9] Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao, "Deep context: end-to-end contextual speech recognition," in *SLT*, 2018.

[10] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen, "Joint grapheme and phoneme embeddings for contextual end-to-end asr.," in *Interspeech*, 2019.

[11] Antoine Bruguier, Rohit Prabhavalkar, Golan Pundak, and Tara N Sainath, "Phoebe: Pronunciation-aware contextualization for end-to-end speech recognition," in *ICASSP*, 2019.

[12] Mehryar Mohri, Fernando C Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput. Speech Lang.*, vol. 16, pp. 69–88, 2002.

[13] Ian Williams, Anjuli Kannan, Petar S Aleksic, David Rybach, and Tara N Sainath, "Contextual speech recognition in end-to-end neural network systems using beam search.," in *Interspeech*, 2018.

[14] Linda Liu, Yile Gu, Aditya Gourav, Ankur Gandhe, Shashank Kalmane, Denis Filimonov, Ariya Rastrow, and Ivan Bulyko, "Domain-aware neural language models for speech recognition," in *ICASSP*, 2021.

[15] Mahaveer Jain, Gil Keren, Jay Mahadeokar, and Yatharth Saraf, "Contextual rnn-t for open domain asr," *arXiv preprint arXiv:2006.03411*, 2020.

[16] Da-Rong Liu, Chunxi Liu, Frank Zhang, Gabriel Synnaeve, Yatharth Saraf, and Geoffrey Zweig, "Contextualizing asr lattice rescoring with hybrid pointer network language model," *arXiv preprint arXiv:2005.07394*, 2020.

[17] Aditya Gourav, Linda Liu, Ankur Gandhe, Yile Gu, Guitang Lan, Xiangyang Huang, Shashank Kalmane, Gautam Tiwari, Denis Filimonov, Ariya Rastrow, et al., "Personalization strategies for end-to-end speech recognition systems," in *ICASSP*, 2021.

[18] Ding Zhao, Tara N Sainath, David Rybach, Pat Rondon, Deepti Bhatia, Bo Li, and Ruoming Pang, "Shallow-fusion end-to-end contextual biasing.," in *Interspeech*, 2019.

[19] Petar Aleksic, Mohammadreza Ghodsi, Assaf Michaely, Cyril Allauzen, Keith Hall, Brian Roark, David Rybach, and Pedro Moreno, "Bringing contextual information to google speech recognition," 2015.

[20] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[21] Rongqing Huang, Ossama Abdel-Hamid, Xinwei Li, and Gunnar Evermann, "Class lm and word mapping for contextual biasing in end-to-end asr," *arXiv preprint arXiv:2007.05609*, 2020.

[22] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L. Seltzer, and Christian Fuegen, "End-to-end contextual speech recognition using class language models and a token passing decoder," *ICASSP*, 2019.

[23] Shubham Toshniwal, A. Kannan, Chung-Cheng Chiu, Y. Wu, T. Sainath, and Karen Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," *SLT*, 2018.

[24] A. Kannan, Y. Wu, P. Nguyen, T. Sainath, Zhijeng Chen, and Rohit Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," *ICASSP*, 2018.

[25] Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates, "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[26] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[27] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L. Seltzer, and Christian Fuegen, "Joint grapheme and phoneme embeddings for contextual end-to-end asr," in *Interspeech*, 2019.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[29] Zhengkun Tian, Jiangyan Yi, Jianhua Tao, Ye Bai, and Zhengqi Wen, "Self-attention transducers for end-to-end speech recognition," *Interspeech*, 2019.

[30] Ching-Feng Yeh, Jay Mahadeokar, Kaustubh Kalgaonkar, Yongqiang Wang, Duc Le, Mahaveer Jain, Kjell Schubert, Christian Fuegen, and Michael L Seltzer, "Transformer-transducer: End-to-end speech recognition with self-attention," *arXiv preprint arXiv:1910.12977*, 2019.

[31] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP*, 2020.

[32] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] M. Schuster and K.K. Paliwal, "Bidirectional recurrent neural networks," *Trans. Sig. Proc.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[35] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint arXiv:1908.08962v2*, 2019.

[36] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.