TRAINING ASR MODELS BY GENERATION OF CONTEXTUAL INFORMATION

Kritika Singh, Dmytro Okhonko, Jun Liu, Yongqiang Wang, Frank Zhang, Ross Girshick, Sergey Edunov, Fuchun Peng, Yatharth Saraf, Geoffrey Zweig, Abdelrahman Mohamed

Facebook AI

ABSTRACT

Supervised ASR models have reached unprecedented levels of accuracy, thanks in part to ever-increasing amounts of labelled training data. However, in many applications and locales, only moderate amounts of data are available, which has led to a surge in semi- and weakly-supervised learning research. In this paper, we conduct a large-scale study evaluating the effectiveness of weakly-supervised learning for speech recognition by using loosely related contextual information as a surrogate for ground-truth labels. For weakly supervised training, we use 50k hours of public English social media videos along with their respective titles and post text to train an encoder-decoder transformer model. Our best encoder-decoder models achieve an average of 20.8% WER reduction over a 1000 hours supervised baseline, and an average of 13.4% WER reduction when using only the weakly supervised encoder for CTC fine-tuning. Our results show that our setup for weak supervision improved both the encoder acoustic representations as well as the decoder language generation abilities.

Index Terms— End-to-end ASR, Weak-supervision

1 Introduction

Over the past few years, Automatic Speech Recognition (ASR) has made great strides due to the successful application of supervised Deep Learning (DL) techniques [1, 2, 3, 4]. However, one drawback of such approaches is the heavy reliance on large volume of supervision which can be difficult to acquire for new domains. A good ASR system operating in real environments requires a large volume of training data to marginalize out and deal with different acoustic conditions of background noise, languages, accents, speakers, and their emotional states. This practical need has led to a surge in ASR research in unsupervised acoustic feature learning [5, 6, 7, 8, 9, 10, 11, 12], as well as semi and weakly-supervised learning [13, 14, 15, 16, 17, 18]. In [19], the "island of confidence" technique was used to filter the owner-uploaded video transcripts creating additional weakly-supervised ASR training data. 1 million hours of audio were transcribed using a teacher ASR model to train a production-ready student model [15]. To improve performance in rare words and proper nouns, [16] distilled top hypothesis generated by a contextually-biased ASR system as ground truth for training an encoder-decoder ASR model.

This paper belongs to this last category with a focus on public social media videos, which provide interesting challenges and opportunities for ASR research. On one hand, these videos contain a diverse range of speakers, dialects, topics, and acoustic conditions making automatic recognition difficult. On the other hand, parallel audio, visual and text information (e.g. video title, post text, and comments) is available for social media videos over which joint multi-modal learning is possible. This work focuses solely on utilizing video title and post text as additional contextual information for acoustic model training.

The relationship between contextual text and audio associated with a video ranges from weak semantic relatedness to, sometimes, overlap of exact words, phrases, or quotes taken verbatim from the audio. Training an ASR model to generate such related context information from audio signals exposes it to a large volume of diverse training examples, even if they are far from the exact speech content of audio. The downside is that the audio content may not be related or represented at all in the contextual text, let alone being monotonically aligned to the audio content. In this study, we evaluate the effectiveness of using contextual text from videos as weak labels for large-scale ASR training, and outline a proposal to overcome the aforementioned problems, achieving an average of 20.8% WER reduction over an encoder-decoder baseline system trained only on 1000h of supervised data, and 13.4% when we transfer only the encoder part of the model to be fine-tuned using the Connectionist Temporal Classification (CTC) loss [20].

2 Weakly supervised training

2.1 Datasets

We use two sets of training data: (i) $\{X, Y^s\} \in \mathcal{D}^s$ is the supervised data where X and Y^s are pairs of audio features and label sequences. (ii) $\{X, Y^w\} \in \mathcal{D}^w$ is the weakly-supervised dataset where X and Y^w are pairs of audio features and the corresponding contextual text. The targets Y^s and Y^w are sequences of sub-word units [21].

2.2 The proposed approach

Our proposed acoustic model training centers around utilizing distant, weak supervision from contextual text surrounding social media videos. We use an encoder-decoder approach [22, 23] for maximizing the conditional probability of generating the contextual text sequence Y^w given X an input se-

quence of mel-scale log filterbank features where $x_i \in R^d$

$$\mathcal{F}^w = p(Y^w|X;\theta^w) = \prod_{i=1}^{M} p(y_i^w|y_1^w, y_2^w, ..., y_{i-1}^w, x_1, x_2, ..., x_T; \theta^w)$$

The attention-based encoder-decoder approach fits well with the proposed training approach since it offers flexible alignment and unconstrained coverage between input and output sequences. Other ASR training approaches aren't suitable given the abstractive relationship between Y^w and X. The hybrid HMM-NN approach requires a low-level sub-second alignment between input audio and output targets, while the CTC approach assumes a monotonic alignment between inputs and outputs, and it constrains the maximum possible length of the output sequence by the length of the input sequence. The final objective function is $\mathcal{F} = \mathcal{F}^w + \mathcal{F}^s$ where the supervised term, $\mathcal{F}^s = p(Y^s|X;\theta^s)$, is also maximized using the encoder-decoder approach. We share the full model for both types of data, where $\theta^w = \theta^s = \{\theta_{enc}, \theta_{dec}\}$ combines the parameters in the audio encoder and the language generation and attention parameters in the decoder. During training, we alternate, with some mixing ratio, between mini-batches sampled from the two training sets $\mathcal{D}^s$ and $\mathcal{D}^w$.

2.3 The main assumptions

We are making two main assumptions in the proposed training approach:
(1) $|\mathcal{D}^w| >> |\mathcal{D}^s|$, and the diversity of acoustic conditions represented in $\mathcal{D}^w$ is much larger than that in the supervised data $\mathcal{D}^s$. Therefore, training on $\mathcal{D}^w$ has the potential to improve the final model's ability to generalize better to new speakers and recording conditions compared to a baseline model trained only on $\mathcal{D}^s$. To test the importance of this assumption, in our experiments, we present results for pre-training the ASR model using 50x, 10x, and 2x of the supervised data size.
(2) Maximizing $p(Y^w|X;\theta)$ can be used as a proxy for maximizing $p(Y^s|X,\theta)$. This is a rather strong assumption since the best ASR system will not generate a commentary for its speech inputs and vice versa. In other words, θ^{w*}, the optimal model parameters maximizing the conditional likelihood of $\mathcal{D}^w$, may not equal an optimal ASR model's parameters θ^{s*}. To test this assumption, we explore three specific questions:
(i) Given an input sequence X, how close is the user-generated commentary Y^w to Y^s, the true audio content, under some semantic measure of relatedness? We use the set intersection of words in Y^w and hypothesis generated using a baseline ASR as a proxy for relatedness. In our experiments, we test multiple levels of strictness for enforcing this condition.
(ii) Does maximizing $p(Y^w|X;\theta)$ improve $p(Y^s|X,\theta)$ during all phases of model optimization? We distinguish between three learning phases during model optimization: (a) An initial `burn-in` phase (b) A final `fine-tune` phase (c) An intermediate `train-main` phase. More details about these three phases are in Section 2.5. We hypothesize that the maximum transfer between $\mathcal{F}^w$ and $\mathcal{F}^s$ happens during the `train-main` intermediate phase of learning.
(iii) Does training the ASR model on $\mathcal{D}^w$ benefit all model components equally? Does it hurt some of them? To answer this question, we evaluate the impact of weakly supervised training on two ASR supervised fine-tuning setups: (a) One that utilizes both θ_{enc} and θ_{dec} from weak-supervision i.e, the acoustic and the language modeling components, for the final ASR model fine-tuning. (b) A second ASR setup where we only use θ_{enc} for initializing an acoustic-only model that is subsequently fine-tuned with a CTC loss function [20] using an independently trained and fixed language modeling component. Using these two setups, we can distinguish gains due to better encoder acoustic representation from better language generation abilities learned by the decoder component.

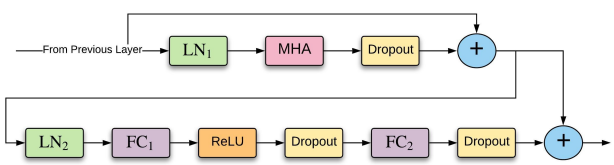

Fig. 1: A block diagram of one transformer block.

2.4 The model architecture

We use transformer blocks as building blocks in our encoder-decoder ASR model [24, 25, 26], shown in figure 1, and follow the convolutional transformer architecture from [27]. For joint encoding of input content and position, the encoder input applies convolutional blocks each consisting of 2-D convolution, Layer Normalization (LN), ReLU non-linearity, and max-pooling layers:

$$[c_1, c_2, ..., c_\tau] = 2DConvBlocks([x_1, x_2, ..., x_T])$$

The encoder then applies multiple transformer blocks. Multi-headed self-attention (MHA) is the core component of transformer blocks. Self-attention (SA) represents each time step of the sequence $C \in \mathbb{R}^{\tau \times d}$ as a sum of all other time steps weighted by the inner products of their representations where each time step acts as a query q_t, a key k_t, and a value v_t. Scores for each time step are scaled by the inverse square root of the dimension d over which the inner product is computed. A softmax operation is applied over all possible key indices, to encourage soft competition between different time steps, followed by a dropout operation to the combination weights:

$$SA(Q,K,V) = Dropout\left(Softmax(QK^T/\sqrt{d})\right) * V$$

Multi-head attention (MHA) extends self-attention by repeating it number-of-heads times, H, using different linear projections for each head, and concatenating their outputs:

$$MHA(Q,K,V) = Concat_{h=1}^{H}\left(SA(Qw_q^h, Kw_k^h, Vw_v^h)\right) * w_o$$

$\{w_q^h, w_k^h, w_v^h\} \in \mathbb{R}^{d \times d_i}$ project Q, K, and V matrices differently for each head to the desired inner product dimension

d_i. $w_o \in \mathbb{R}^{Hd_i \times d}$ projects the concatenated self-attention vectors to the output dimension d. Following the MHA sub-block, each transformer block applies a fully connected feed-forward sub-block, which is composed of two linear transformers and a ReLU non-linearity in between, to each time step. To avoid vanishing gradients, residual connections are added around the MHA and the fully connected sub-blocks, and LN operations are applied before them. The only difference between the encoder and decoder architectures is the use of a 1-d convolution operation to represent previously generated output tokens $[y_1, y_2, ..., y_{i-1}]$. The decoder component uses two transformer blocks each with multi-head cross attention for summarizing the final encoder representations. The dot-product cross-attention makes neither coverage nor monotonicity assumptions about the relationship between the input and output sequences. Similar to story and dialogue response generation [28, 29], segments of the input sequence can be covered in the output sequence once, multiple times or none at all, and similarly for segments of the output sequence.

2.5 The training process

In our experiments, we study the impact of the three training phases introduced in 2.3: (1) An initial supervised `burn-in` phase in which the decoder cross-attention learns to properly communicate gradient information to adjust encoder acoustic features. (2) A training phase driven by a mixture of the supervised and the weakly supervised loss functions, we refer to it as the `train-main` phase, in which the model expands its inventory of audio features and mappings between acoustic and linguistic cues. (3) A final supervised-only `fine-tune` phase which utilizes either the full encoder-decoder model trained in the `train-main` step, or the encoder component to be refined by the CTC loss. Given the transformer's ability to reconstruct input sequences in any desired arbitrary order, an extra transformer block is optionally added to the encoder layers before fine-tuning to smooth out the transition from the encoder-decoder cross-entropy loss to the CTC loss which enforces monotonicity between input and output sequence.

3 Experiments

Data: Both the supervised and weakly-supervised (WS) datasets used in this study are sampled from our in-house datasets. Our supervised dataset consists of 1000 hours of data sampled from public English videos that are anonymized. We use this data exclusively in the `burn-in` and `fine-tune` stages of training, and on a fraction of minibatches in `train-main` determined by the mixing ratio, as well as for training the baseline ASR model.

Our weakly-supervised dataset consists of 4M public English videos that are 30 to 60 seconds in duration with contextual text between 60 to 700 characters, totaling 50,000 hours. We restrict our target context to video titles and post text only. Because the relevance of the contextual text to the audio content might greatly impact the expected gain, we filter the weakly supervised data based on the set intersection between contextual text and the baseline ASR hypothesis considering only words of length more than 3 characters, to create two additional sets: (1) A 2,300 hours subset with an intersection of 14 words or more (2) A 12,800 hours subset with an intersection of 6 words or more. To test the impact of quality of contextual text while keeping acoustic richness same, we create two additional subsets: (3) A 2,300 hours subset randomly sampled from the 50,000 hour original set (4) A 12,800 hours subset randomly sampled from the 50,000 hour original set. Additionally, to measure the impact of contextual information for weak supervision in terms of labeled data size, we create: (5) A 2,000 hours set of supervised data that is disjoint to the 1000 hours set used for the baseline. For performance evaluation, we use three test sets, `clean` with 1.3K videos (20 hours), `noisy` with 1.3K videos (20 hours), and `extreme` with 13K videos (77 hours) which is more acoustically and phonetically challenging. For hyper-parameter tuning and model selection, we use a `dev-noisy` subset which consists of 600 videos (9 hours). For inference using the encoder-only model, we use a 5-gram language model which is estimated from 1M utterances containing about 120k distinct tokens.

Experimental setup: The input speech is encoded into 80 dimensions of mel-scale log filterbank features computed over 16ms and shift of 10ms. The encoder two 2-D convolution blocks uses kernel size=3 and output features of 64 and 128 for each block respectively. The max-pooling layers sub-sample the input time steps and frequency channels by a factor of 4. All encoder and decoder transformer blocks, 10 and 2 blocks respectively, use 1k hidden dimension, 16 heads, 4k projection layer before the ReLU nonlinearity, and dropout rate of 0.15. The decoder part uses 4 1-D convolutional layers with kernel size=3 and output features of 256. Supervised labels and contextual text is encoded into 5k sub-word output vocabulary [21]. We use the AdaDelta algorithm [30] with fixed learning rate=1.0 and gradient clipping at 10.0 where total gradients are scaled by the number of utterances in each minibatch. During `train-main` we save checkpoints every 5k model updates and average the last 20 checkpoints to initialize the `fine-tune` phase. We also average the last 20 checkpoints of fine-tuning before decoding. We use a beam size=20 for encoder-decoder model inference without any external language model. We use the 5-gram LM for decoding the CTC fine-tuned models.

Results: Table 1 presents the main results of this study on the three test conditions showing an average WER reduction of 20.8% for the encoder-decoder setup and 13.4% for the CTC setup compared to the baseline supervised model. For all experiments shown in table 1, the `burn-in` phase had 15k model updates, `train-main` had 400k updates with almost one third of the minibatches sampled from the supervised data (mixing ratio=0.3). Supervised `fine-tune` phase used 22k model updates for the encoder-decoder architectures and 150k updates for the encoder-only CTC loss.

While using the full 50,000h weakly supervised data improves upon the baseline system, filtering it for relevant content provided the best performance almost across all setups. Both models show improvement from using more weakly supervised data, however the encoder-decoder model's gains

	Encoder-Decoder			CTC		
	clean	noisy	extreme	clean	noisy	extreme
Supervised baseline						
1000h	22.8	30.2	42.1	21.6	28.5	37.6
Weakly supervised models						
2300h	20.9	27.5	38.2	19.2	25.8	35.2
12,800h	18.6	25.5	34.8	18.7	25.2	34.2
50,000h	18.3	25	34.6	18.6	**24.5**	34.2
Weakly supervised filtered by relevance						
2300h	19.3	26.3	37	18.7	25.2	34.6
12,800h	**17.6**	**24.3**	**33.7**	**18.2**	24.8	**33.7**
Extra true labels instead of weak supervision						
2000h	18.8	25.8	35	18.7	25.1	34

Table 1*:* WERs of the enc-dec and CTC fine-tuned models on the test sets `clean`, `noisy` and `extreme` for different train data sizes

Burn-in phase	N		Y	
Mixing ratio	0	0.3	0	0.3
50,000h	27.6	27.3	24.7	25.4
Filtered 12,800h	28.7	26.1	24.4	25

Table 2*:* WERs of fine-tuned CTC model on `dev-noisy` under different conditions of burn-in and mixing ratio

are slightly higher given that large weak supervision data improves both their acoustic encoder representations and decoder generation abilities, relative to the CTC fine-tuned model which uses a fixed n-gram language model and benefits only from the improved encoder representations.

An interesting observation is that, using weak supervision, the encoder-decoder setups are almost as good or slightly better than their corresponding encoder-only setups, even though encoder-decoder models don't use any external language model. This suggests that weak supervision via context generation was enough to realize a language modeling capability similar to that of the n-gram LM used for the encoder-only models. The best weakly supervised systems are consistently better than using an additional 2000 hours of labeled data for the `train-main` phase. This magnifies the value of weak supervision on reducing the requirement for ASR data labeling. One problem of using weak supervision that is not aligned with the input sequence is that the decoder won't be able to refine encoder representations easily. Hence we included supervised mixing and/or initial burn-in phase during our weakly supervised `train-main` phase. Table 2 shows the effect of both techniques on the learned representations of the encoder-only setup trained by the CTC loss. Comparing cases either with burn-in or mixing shows that mixing helps a bit, however, supervised `burn-in` is much more important than mixing for encoder representations. When burn-in is on, spending almost one third of the mini-batches during `train-main` visiting supervised

Enc-Dec fine-tuning	50,000h	Filtered 12,800h
N	26.4	25.3
Y	25.7	24.8

Table 3*:* The impact of fine-tuning encoder-decoder model on `dev-noisy`

data seems to hurt performance because the model has less chance to observe the more diverse and larger weakly supervised data. Table 3 shows that the encoder-decoder model is ready for recognition with good performance even without the final `fine-tune` phase, but it still benefits from the 22k updates of supervised fine-tuning which, we believe, polish its decoder cross-attention input audio sequences to be more monotonic.

4 Discussion and Related Work

Our work builds on the success of sequence-to-sequence learning for ASR, both the CTC-based [31, 32, 4] and the attention-based [23, 22] variants, by replacing the ground-truth target sequences with semantically related weak supervision. This study is primarily motivated by [33, 34] where hashtag prediction of social media images is successfully used for large CNN pre-training for many image classification and object detection tasks. There is a growing body of research in self-supervised pre-training of models on surrogate tasks for general language representation which have shown great success in several downstream NLP tasks [35, 36]. Grounding language learning and generation [37, 38, 39, 17, 40], spoken keyword spotting, and audio representation [18, 41] on visual cues motivated this work where, similar to this study, inputs and outputs are loosely related with no guarantees of coverage. Weak semantic labels showed significant improvements in phonetic learning in a model of infant language acquisition and vocal commands learning for dysarthric speakers [42, 43]. This work belongs to the growing line of research focusing on reducing the reliance on supervised labels for building ASR systems through unsupervised unit discovery and acoustic representation learning [5, 6, 7, 8, 9, 10, 11, 12], multi- and cross-lingual transfer learning in low-resource conditions [44, 45, 46, 47, 48, 49], and semi-supervised learning [13, 14, 15, 16].

5 Conclusion and Future work

We presented a large-scale weakly supervised training method for speech recognition systems that uses contextual social media information – titles and post text – as a surrogate for transcriptions. An encoder-decoder approach was trained to generate the contextual labels, which are semantically related to the spoken audio content, but have neither monotonicity nor full coverage. Our best models achieved averages of 20.8% and 13.4% WER reduction over supervised baselines. In the future, we would like to combine compare and investigate synergies between semi-supervised student-teacher modeling with our weak-supervision method.

6 References

[1] G. Hinton et. al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, 2012.

[2] W. Xiong et. al., "Toward human parity in conversational speech recognition," *IEEE/ACM TASLP*, vol. 25, no. 12, 2017.

[3] George Saon et. al., "English conversational telephone speech recognition by humans and machines," in *Interspeech 2017*.

[4] Dario Amodei et.al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *ICML 2016*.

[5] A. Park and J. Glass, "Unsupervised pattern discovery in speech," *IEEE TASLP*, vol. 16, no. 1, pp. 186–197, 2008.

[6] Aren Jansen, Kenneth Church, and Hynek Hermansky, "Towards spoken term discovery at scale with zero resources.," in *Interspeech 2010*.

[7] J. Glass, "Towards unsupervised speech processing," in *ISSPA 2012*.

[8] Aren Jansen et. al., "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *ICASSP 2013*.

[9] Odette Scharenborg et. al., "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "speaking rosetta" JSALT 2017 workshop," in *ICASSP 2018*.

[10] Yu-An Chung and James Glass, "Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech," in *INTERSPEECH 2018*.

[11] Aäron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[12] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019*.

[13] K. Vesely, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *ASRU 2013*.

[14] Sheng Li, Xugang Lu, Shinsuke Sakai, Masato Mimura, and Tatsuya Kawahara, "Semi-supervised ensemble DNN acoustic model training," *ICASSP 2017*.

[15] S. Krishnan Parthasarathi and N. Strom, "Lessons from building acoustic models with a million hours of speech," in *ICASSP 2019*.

[16] Bo Li, Ruoming Pang, Tara Sainath, and Zelin Wu, "Semi-supervised training for end-to-end models via weak distillation," in *ICASSP 2019*.

[17] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi, "Representations of language in a model of visually grounded speech signal," in *ACL 2017*.

[18] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," in *Interspeech*, 2017.

[19] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *ASRU 2013*.

[20] Alex Graves, Santiago Fernández, and Faustino Gomez, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML 2006*.

[21] Taku Kudo and John Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *EMNLP 2018*.

[22] 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*.

[23] Jan Chorowski et. al., "Attention-based models for speech recognition," in *NeurIPS 2015*.

[24] Ashish Vaswani et.al., "Attention is all you need," in *NeurIPS 2017*.

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

[26] Shigeki Karita et. al., "A comparative study on transformer vs RNN in speech applications," *CoRR*, vol. abs/1909.06317, 2019.

[27] Abdelrahman Mohamed, Dmytro Okhonko, and Luke Zettlemoyer, "Transformers with convolutional context for ASR," *CoRR*, vol. abs/1904.11660, 2019.

[28] Angela Fan, Mike Lewis, and Yann Dauphin, "Hierarchical neural story generation," in *ACL 2018*.

[29] Jiwei et. al. Li, "A diversity-promoting objective function for neural conversation models," in *NAACL-HLT 2016*.

[30] Matthew D. Zeiler, "Adadelta: An adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012.

[31] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML 2014*.

[32] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *CoRR*, vol. abs/1609.03193, 2016.

[33] Dhruv Kumar Mahajan et. al., "Exploring the limits of weakly supervised pretraining," in *ECCV*, 2018.

[34] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache, "Learning visual features from large weakly supervised data," *CoRR*, vol. abs/1511.02251, 2015.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[36] Yinhan Liu et. al., "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[37] Erik D. Thiessen, "Effects of visual information on adults' and infants' auditory statistical learning," *Cognitive Science*, vol. 34, no. 6, pp. 1093–1106, 2010.

[38] Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux, "Learning words from images and speech," in *In NIPS Workshop on Learning Semantics*, 2014.

[39] David F. Harwath, Antonio Torralba, and James R. Glass, "Unsupervised learning of spoken language with visual context," in *NeurIPS 2016*.

[40] Rahma Chaabouni, Ewan Dunbar, Neil Zeghidour, and Emmanuel Dupoux, "Learning weakly supervised multimodal phoneme embeddings," in *Interspeech 2017*.

[41] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NeurIPS 2016*.

[42] Stella Frank, Naomi H. Feldman, and Sharon Goldwater, "Weak semantic context helps phonetic learning in a model of infant language acquisition," in *ACL 2014*.

[43] Vincent Renkens and Hugo Van hamme, "Mutually exclusive grounding for weakly supervised non-negative matrix factorisation," in *Interspeech 2015*.

[44] Haihua Xu et. al., "Semi-supervised and cross-lingual knowledge transfer learnings for DNN hybrid acoustic models under low-resource conditions," in *Interspeech 2016*.

[45] Jia Cui et. al., "Multilingual representations for low resource speech recognition and keyword search," in *ASRU 2015*.

[46] Georg Heigold et. al., "Multilingual acoustic models using distributed deep neural networks," in *ICASSP 2013*.

[47] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *ICASSP 2013*.

[48] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," *ICASSP 2013*.

[49] Ngoc Thang Vu et. al., "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *ICASSP 2014*.