On Leveraging Visual Modality for Speech Recognition Error Correction

Sang Keun Choe [*1] **Quanyang Lu** [*1] **Vikas Raunak** [*1] **Yi Xu** [*1] **Florian Metze** [1]

Abstract

We present our recent efforts on leveraging visual modality for automated speech recognition (ASR) error correction. A visually grounded attention-based Sequence-to-Sequence (S2S) model is trained to correct contextual and functional word errors in transcripts/outputs of unimodal ASR systems. Specifically, our error correction model address the problem of semantic gap in multimodal fusion, which allows high-level visual features to be combined with *comparably* high-level text features. Visual-semantic joint embedding and language model are used to rescore the n-best list output by the error correction model. Tested on the How2 dataset, visually grounded error correction led to only marginal improvements over the unimodal ASR system. We provide error analysis on the output of visually grounded ASR error correction model, and propose a potential solution based on the analysis.

1. Introduction

Humans are able to accurately recognize speech even in the presence of surrounding noise or accent by utilizing various other *contexts*. For instance, when listening to machine learning lectures, we know that the probability of words like 'backpropagation' and 'convolutional' being spoken by a lecturer is high, and use this context to recognize the speech better. Given humans' multimodal perception systems, visual information would be one of those contexts that can help speech recognition as demonstrated in the well-known McGurk effect (McGurk & MacDonald, 1976).

In recent years, deep neural networks have revolutionized signal processing and machine learning research by achieving state-of-the-art results on many problems in computer vision (He et al., 2015), speech processing (Chan et al., 2015), and natural language processing (Vaswani et al., 2017). In particular, (Chan et al., 2015; Chiu et al., 2017; Bahdanau et al., 2016) have shown that Recurrent Neural Networks (RNN) based sequence-to-sequence (S2S) models with attention mechanism can be successfully exploited for automated speech recognition (ASR). While the above approaches have demonstrated promising performance in terms of Word Error Rate (WER), they are still limited in the sense that they do not utilize additional contexts such as visual information unlike humans.

Accordingly, there have been several recent attempts to improve performance of ASR by incorporating multimodal information as additional contexts into existing ASR systems (Sanabria et al., 2018). However, combining features from different modalities by simply concatenating multimodal features is inefficient due to the semantic gap between the modalities. For example, the features in most ASR systems are characters, produced as the output of decoder, and they are low-level features compared to visual features such as scene contexts, motions and objects in videos.

Therefore, in this paper, we consider augmenting traditional unimodal ASR systems with multimodal information in a more *natural* way. We introduce the interaction between modalities at specific points in the neural network architecture, allowing modules to leverage signals from modalities with similar semantic level. To this end, we propose a visually grounded error correction model and a rescoring scheme, all of which fuse visual information with high-level text features. Tested on the How2 dataset (Sanabria et al., 2018), proposed approaches led to only marginal improvements over the unimodal baseline. We provide extensive analyses on visually grounded error correction on ASR, and propose several future research directions based on the analyses.

2. Related work

2.1. Automated speech recognition

Listen-Attend-Spell, referred to as LAS (Chan et al., 2015) is among the first end-to-end trained neural networks to achieve close to the state-of-the-art results in speech recognition. They proposed a novel pyramidal bidirectional LSTM (pBLSTM) and constructed an encoder-decoder architecture with pBLSTMs and attention mechanism. Based on LAS, (Chiu et al., 2017) introduced a handful of techniques including multi-headed attention and scheduled sampling which significantly improved the performance of S2S-based ASR systems. Though all these works show promising results, their methods do not take multimodal information into account.

Recognizing the importance of multimodal processing, (Palaskar et al., 2018; Caglayan et al., 2018) proposed visually grounded speech recognition systems. They first extracted visual features by using object recognition (He et al., 2016) and action recognition models (Hara et al., 2018), and integrated them with hidden states in encoder and decoder of LAS model. While their method gives improvement over unimodal speech recognition systems, feature-level mismatch between visual features and hidden states of encoder/decoder makes such concatenation possibly sub-optimal.

2.2. Error correction

S2S models (Sutskever et al., 2014) have been used in text correction for both spelling (typing) errors and grammar errors. Spelling error correction is usually considered in the keyboard typing decoding context. (Ghosh & Kristensson, 2017) developed a S2S model for text typo correction by combining of character-level CNN and GRU encoder with word-level GRU decoder. While their method shows promising performance on Twitter typo dataset, it is only applicable for short phrases with the word length of 7.

Neural network models have also been used in Grammar error correction (GEC). (Yuan & Briscoe, 2016) applied neural machine translation (NMT) S2S model to tackle the GEC task. The rare word problem is addressed using an unsupervised aligner. An alternative approach in (Xie et al., 2016) is to apply a character-level S2S model with attention mechanism, but it has limited capability in leveraging high-level information. Combing the char- and word-level S2S models, (Ji et al., 2017) developed a neural hybrid model for GEC task similar to (Luong & Manning, 2016), and achieved higher scores than word-level GEC model.

All the above-mentioned works are related to unimodal text correction. Recently, (Guo et al., 2019) developed a S2S spelling correction for ASR system output. The spelling correction module consists of a 3-layer stacked LSTM S2S model. The encoder and decoder are both in subword-level using wordpiece model. Using external language models for n-best list rescoring further improve the output accuracy. (Zhang et al., 2019) proposed a transformer-based S2S correction model for CTC-based ASR system.

3. Base unimodal ASR system

Due to limited computational resources, we could not afford S2S model that is as big as the baseline in (Sanabria et al., 2018). Accordingly, we introduce several techniques that can improve the performance of S2S ASR systems without significantly increasing computational costs.

3.1. 2D Pre-convolutional neural network

Unlike normal LAS model solely composed of LSTMs, we plug 2D pre-convolutional neural networks (CNNs) before the encoder of LAS model to extract temporal transition invariance in time domain and spectral invariance in frequency domain, as proposed in (Amodei et al., 2016). This time-and-frequency domain 2D CNNs transform 40 dimensional MFCC features into new audio features, each frame of which contains information spanning within the receptive field of CNNs. Consequently, each time step of following encoder LSTMs can have a direct access to multiple time steps of input MFCC features, which finally lead to better performance of ASR systems. In detail, our 2D pre-CNNs consists of 4 convolutional layers and expands the dimension of input MFCC features by 3 times. We did not apply any pooling or stride (>1) convolution layers, which decrease the sequence length of input MFCC features. 2D pre-CNNs is simultaneously trained with the following LAS model.

3.2. Listen, attend, and spell (LAS)

As our main speech recognition system, we exploit a common S2S framework with attention mechanism. The encoder is composed of 3 layers of pyramidal LSTM (pLSTM) layers as in (Chan et al., 2015), and each pLSTM layer reduces the time dimension by 2 times. This pyramidal structure allows the encoder to convert low-level input audio features into compact and high-level representations that can be easily decoded to characters in the following decoder. The decoder consists of 2 layers of LSTM and uses attention to calculate contexts from encoder representations. Contexts obtained with the attention algorithm are concatenated with the input embedding and the final LSTM output. Probabilities for each character is calculated by applying a fully connected layer and softmax function to the output of decoder LSTM. The hidden dimensions of the encoder and the decoder are respectively set to 160 and 320.

3.3. Multi-headed attention

(Chiu et al., 2017) shows that introducing multi-headed attention into a S2S speech recognition model can significantly improve the final WER. They empirically demonstrate that having more than one attention head allows the decoder to attend to more diverse aspects of encoder representations, and thereby reduces the burden on the encoder of learning ideal representations. For example, one attention head can focus on the actual speech while the other head focus on surrounding noise, which consequently helps the decoder to better distinguish actual speech from noise. Due to computational limits, we trained with 2 heads while we observed that adding more heads leads to better WER score in other datasets. The context dimension of multi-headed

attention in our experiments is set to 80.

3.4. Scheduled Sampling

While using ground truth labels as an input to decoder facilitates learning in the decoder at the early stage of training, it incurs different behaviors of the model in training and testing time. To mitigate this problem, (Bengio et al., 2015) introduces scheduled sampling, which samples the input token at $(i+1)$-th time step from the softmax distribution of the i-th time step output with some probability p, instead of ground truth token. This enables the model to output a correct token in a less guided scheme, and thereby leads to better performance in test time. We start training with the sampling rate $p = 0$ and gradually increase it to the maximum value of 0.25 as training progresses.

4. Error analysis of base unimodal system

With all techniques introduced above, our unimodal LAS shows training WER = 15.4% and test WER = 20.4%. We further analyze the error distribution in terms of insertion (I), deletion (D) and substitution (S) errors. Figure 1 shows error distributions are similar among train, validation, and test set, and the majority of errors ($\sim 60\%$) are substitution errors. Therefore, we focus on substitution errors where the unimodal LAS is making wrong prediction on certain words.

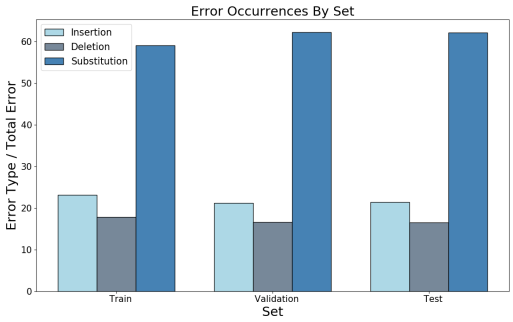

Figure 1. Error Distribution for unimodal LAS model.

Table 1 summarizes the top-10 substitution errors and examples of visual-context errors made by unimodal LAS output on train and test sets. Most errors are 1) functional words or 2) occurred by the lack of high-level contexts. Especially, the second type of errors occurs mainly due to the fact that the unimodal ASR system does not have access to high-level semantic information such as visual context.

Substitution Pairs	Counts
the → a	44
and → in	43
in → and	42
a → the	41
the → to	17
that → the	17
to → the	15
it → that	14
can → could	13
will → would	13
both → bow	5
pipe → pike	4
bedding → betting	4
talk → chalk	4
you → ukulele	4

Table 1. Top-10 substitution errors and examples of visual-context error in unimodal LAS test output

5. Method

Based on the above error analysis on unimodal ASR system, we propose a novel visually grounded ASR error correction model and a rescoring scheme in hopes of 1) correcting functional word errors and 2) incorporating visual information as high-level context features in a *natural* way. The final transcripts is obtained after rescoring n-best list of error correction model output with language model and visual-semantic joint embedding. A schematic illustration of our ASR system is presented in Figure 2.

5.1. Visually grounded error correction model

Our unimodal error correction model consists of 3-layers LSTM S2S model with attention mechanism. To address two problems stated above, we intentionally make our decoder word-level (or subword-level). By doing so, we make features learned in the decoder sufficiently high-level so that it can be naturally fused with visual information such as scene and action features extracted from video. In addition, by having a predefined dictionary for decoder outputs, we can effectively avoid spelling errors, which is a part of substitution errors. We set the embedding and hidden layer sizes for both encoder and decoder to 300 and 512 respectively.

In the visually grounded error correction model, we use the pretrained 2048-dimensional visual action features and reduce the dimension to 50 with a trainable linear layer. Output vector is then concatenated with hidden states of the last LSTM layers in decoder before outputting softmax categorical distribution.

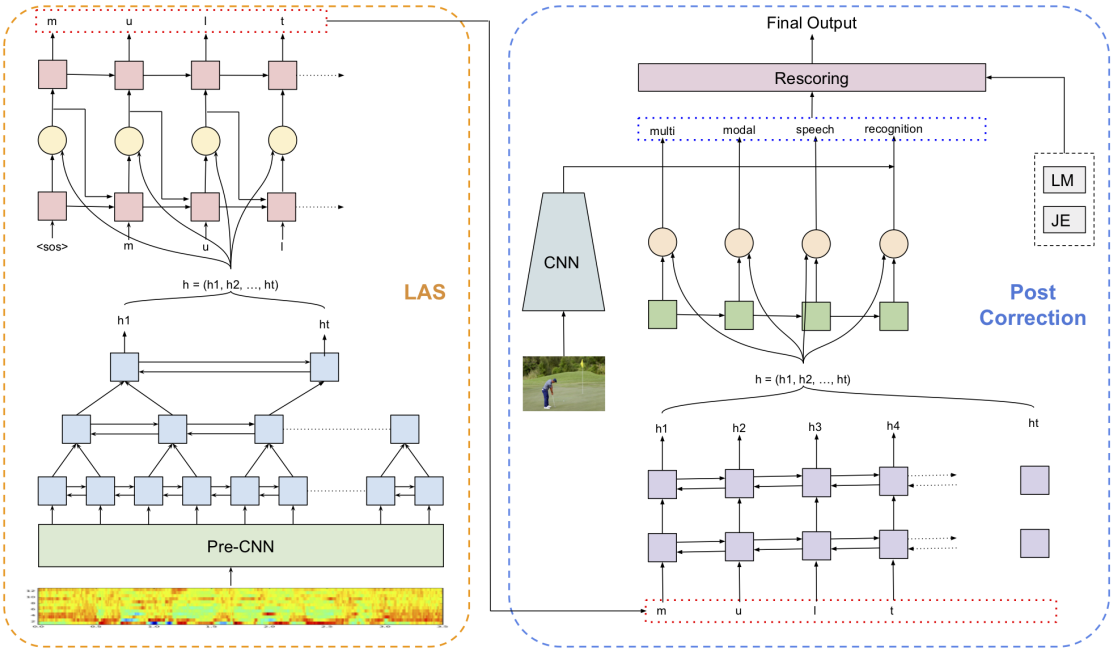

Figure 2. The overall framework of our automated speech recognition system

5.2. Rescoring

By examining the output hypothesis of the error correction model, we found many of the output hypotheses have the ground truth in them. We explored the idea to rescore the n-best list by using offline trained joint embedding and language models to pop up the ground truth to the top of the list. The joint embedding model provides a score ($\mathcal{S}_{JE}$) to evaluate the word probability based on the most related transcripts retrieved from training dataset given visual feature as context. The language model provides the perplexity ($\mathcal{S}_{LM}$) of each sentence in the list. We linearly combine the three scores: the perplexity score ($\mathcal{S}_{ppl}$) generated by the ASR error correction model itself, $\mathcal{S}_{JE}$, and $\mathcal{S}_{LM}$. A grid search is then performed to identify the best hyperparameters α and β.

$$\mathcal{S}_{tot} = (1 - \alpha - \beta)\mathcal{S}_{ppl} + \alpha\mathcal{S}_{JE} + \beta\mathcal{S}_{LM}$$

5.2.1. Joint Embedding

Constructing a joint representation by projecting different modalities onto a latent space can bridge the gap between different modalities (e.g., video, language). In the cross-modal video-text retrieval task (Mithun et al., 2018), a network is learned to minimize the distance between paired video clips and text script while keeping a constant margin between unpaired ones. The retrieval task is performed to find the nearest neighbors in the latent space. Related works (Henning & Ewerth, 2018; Karpathy et al., 2014) have proved that features extracted from video (e.g., scene, actions, objects etc.) are valuable to efficiently retrieve the related texts.

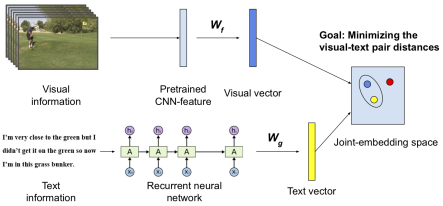

Figure 3. Structure of the joint visual-semantic embedding space

In this work, we train a visual-semantic joint embedding representation of both text and visual features to improve our ASR error correction system by rescoring the n-best list based on its high-level visual context. We use a GRU network to encode the text script as text feature. The trained model achieves recall (R@1) of 17.6%. Conditional word probability is calculated based on top-15 retrieved transcripts in the joint embedding space given visual feature as context. For each sentence in the output n-best list, the joint embedding score ($\mathcal{S}_{JE}$) is defined as the sum of the log-probability of each word in that sentence (Naive Bayes assumption).

5.2.2. Language Model

A common approach to improve the output sequence accuracy in ASR task is to incorporate an external LM (Guo et al.,

2019; Chan et al., 2015). We implement a language model to rescore the output n-best list by error correction model for best candidate sentence. The language model consists of 2-layer unidirectional LSTM for sequence modelling. The embedding and hidden layer sizes are respectively set to 300 and 512. This external language model is then used to generate the perplexity as language model scores ($\mathcal{S}_{LM}$) for rescoring.

6. Experiments & Discussion

6.1. Data

Throughout our experiments, we use the 300 hours subset of How2 dataset (Sanabria et al., 2018), which contains 300 hours of videos, sentence-level time alignments to the ground-truth English subtitles, and Portuguese translations of English subtitles. Detailed statistics of the dataset is presented in Table 2.

The visual features used in this paper are identical to features used in the previous work (Gupta et al., 2017), which are action features extracted from pre-trained CNNs. All ground-truth transcripts are lowercased and every special character in transcripts is replaced with whitespace.

		Videos	Hours	Clips/Sentences
300h	train	13,168	298.2	184,949
	val	150	3.2	2,022
	test	175	3.7	2,305
	held	169	3.0	2,021

Table 2. Statistics of How2 dataset

6.2. Visually grounded error correction

We first experimented different levels of encoders and decoders in our unimodal S2S error correction models, including word2word, subword2subword, subword2word and char2word. Subword level input is processed using byte pair encoding (Sennrich et al., 2015), and the vocabulary size is set to 10k.

Methods	**WER(%)**
Baseline (LAS)	20.41
+ Error correction (word2word)	26.70
+ Error correction (subword2subword)	23.08
+ Error correction (subword2word)	22.18
+ Error correction (char2word)	20.75
+ EC (c2w)+ Visual	20.37
+ EC (c2w)+ Visual + JE	**20.16**
+ EC (c2w)+ Visual + LM	**20.17**
+ EC (c2w)+ Visual + JE + LM	**20.15**

Table 3. Ablation study on word error rates (WER) of our proposed ASR system: Error correction, joint embedding, language model.

Substitution Pairs	**Counts**	**Change**
the → a	46	↑ 2
and → in	41	↓ 2
a → the	40	↓ 1
in → and	39	↓ 3
that → the	17	-
to → the	16	↑ 1
the → to	15	↓ 2
it → that	14	-
in → on	13	↑ 2
can → could	13	-

Table 4. Top-10 substitution errors in char2word error correction model test output

Methods	**I (%)**	**D (%)**	**S (%)**
Baseline (LAS)	4.51	**3.49**	12.41
+ EC + Visual + JE+ LM	**4.23**	3.80	**12.12**

Table 5. Error distribution (insertion, deletion and substitution) in LAS and char2word error correction model test output

As shown in Table 3, the char2word error correction model show the best performance (WER = 20.75%) among all models. Two subword level models show slightly higher WERs (23.08% and 22.18%) compared to char2word model. Word2word model performs much worse (WER = 26.70%). We believe this is due to the rare words problem (misspelled words) in the outputs of the unimodal ASR system, which are inputs to the encoder of our error correction model. As most rare words appear only 1∼2 times, it is difficult to learn meaningful semantic embeddings for those words, and it consequently makes training of error correction model noisy and unstable.

Even with our char2word model, we were still not able to improve WER (20.75%) over the unimodal ASR system (20.41%). Comparing Table 1 and 4, no significant improvement is found in terms of functional words errors. This is contradictory to our expectation that S2S model can correct such errors and improve the performance of ASR systems. We expect this is mainly due the limited training dataset size (180k). In addition, functional word errors are not strictly defined as grammar errors in (Ng et al., 2014). The error patterns are more vague as they often co-exist with insertion and deletion errors compared to well-defined grammar errors.

With visual information, the WER score of our error cor-

rection model improve from 20.75% to 20.37%. Table 5 show that the substitution error is reduced by ∼0.3%, with visual fusion and rescoring. This demonstrates that visual information fusion with word-level decoder can improve the performance of substitution error correction, and achieve better WER score over the baseline unimodal ASR system.

We believe that the main reason we could not significantly improve ASR performance with visual information is that most errors made by the unimodal ASR system are substitution errors on functional words instead of contextual errors as shown in Table 1. Therefore, stronger language modeling capacity and larger training dataset for the error correction model would be more crucial than having access to high-level contexts such as visual information. One possible solution to this problem would be to train language model on an external large-scale speech transcript corpus and fuse hidden states of language model and of decoder in the error correction model. By doing so, we could have a decoder which is much stronger in terms of language modeling and possibly prevent functional word errors.

6.3. Rescoring

We further implemented the rescoring scheme to re-rank the n-best list in our error correction model output. We combine the perplexity score from visually grounded error correction model with (1) Joint embedding, (2) Language model and (3) Joint embedding + Language model.

For (1) and (2), both joint embedding and language model rescoring show independent 0.2% improvements in WER to 20.16% and 20.17%, respectively. For (3), we combined the two rescoring methods with weights α and β for joint-embedding and language model, and $1 - \alpha - \beta$ for error correction model output perplexity as explained in section 4.3. After parameter search, the optimal parameters are $\alpha = 0.1$, $\beta = 0.2$ and beam-size = 13. However, we only see minimal 0.01% improvement to WER = 20.15%. The results indicate that there is only little room to improve when combining joint-embedding and language model in rescoring.

7. Conclusion

In this work, we explored leveraging visual information on ASR error correction by proposing a visually grounded S2S error correction model and rescoring scheme. As we intend to improve the performance of ASR system by 1) having additional visual contexts and 2) correcting spelling and functional word errors, our model only shows limited improvement over the baseline unimodal ASR system. In following analysis, we demonstrate that the most common errors made by ASR system are functional word errors, which cannot be efficiently corrected with additional high-level contexts such as visual information. We accordingly insist that, while having additional visual contexts can improve error correction in ASR systems, it is more important to have stronger language modeling capacity to reduce substitution errors on functional words.

For future work, we would like to address the differences between functional words errors and contextual errors. Currently, such errors are not very well-defined and insertion/deletion/substitution are not adequate in classifying these two types of errors. We also would like to explore more sophisticated S2S model to correct these functional word errors.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. Deep speech 2 : End-to-end speech recognition in english and mandarin. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 173–182, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/amodei16.html.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4945–4949, March 2016. doi: 10.1109/ICASSP.2016.7472618.

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1171–1179, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969239.2969370.

Caglayan, O., Sanabria, R., Palaskar, S., Barrault, L., and Metze, F. Multimodal grounding for sequence-to-

sequence speech recognition. *CoRR*, abs/1811.03865, 2018.

Chan, W., Jaitly, N., Le, Q. V., and Vinyals, O. Listen, attend and spell. *CoRR*, abs/1508.01211, 2015. URL http://arxiv.org/abs/1508.01211.

Chiu, C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, K., Jaitly, N., Li, B., Chorowski, J., and Bacchiani, M. State-of-the-art speech recognition with sequence-to-sequence models. *CoRR*, abs/1712.01769, 2017. URL http://arxiv.org/abs/1712.01769.

Ghosh, S. and Kristensson, P. O. Neural networks for text correction and completion in keyboard decoding. *arXiv preprint arXiv:1709.06429*, 2017.

Guo, J., Sainath, T. N., and Weiss, R. J. A spelling correction model for end-to-end speech recognition. *arXiv preprint arXiv:1902.07178*, 2019.

Gupta, A., Miao, Y., Neves, L., and Metze, F. Visual features for context-aware speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5020–5024. IEEE, 2017.

Hara, K., Kataoka, H., and Satoh, Y. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6546–6555, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Henning, C. and Ewerth, R. Estimating the information gap between textual and visual representations. *International Journal of Multimedia Information Retrieval*, 7(1):43–56, 2018.

Ji, J., Wang, Q., Toutanova, K., Gong, Y., Truong, S., and Gao, J. A nested attention neural hybrid model for grammatical error correction. *arXiv preprint arXiv:1707.02026*, 2017.

Karpathy, A., Joulin, A., and Fei-Fei, L. F. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pp. 1889–1897, 2014.

Luong, M.-T. and Manning, C. D. Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv preprint arXiv:1604.00788*, 2016.

McGurk, H. and MacDonald, J. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976.

Mithun, N. C., Li, J., Metze, F., and Roy-Chowdhury, A. K. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pp. 19–27. ACM, 2018.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pp. 1–14, 2014.

Palaskar, S., Sanabria, R., and Metze, F. End-to-end multimodal speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5774–5778. IEEE, 2018.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347, 2018. URL http://arxiv.org/abs/1811.00347.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., and Ng, A. Y. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*, 2016.

Yuan, Z. and Briscoe, T. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 380–386, 2016.

Zhang, S., Lei, M., and Yan, Z. Automatic spelling correction with transformer for ctc-based end-to-end speech recognition. *arXiv preprint arXiv:1904.10045*, 2019.