Combining Residual Networks with LSTMs for Lipreading

Themos Stafylakis, Georgios Tzimiropoulos

Computer Vision Laboratory
University of Nottingham, UK

{themos.stafylakis, yorgos.tzimiropoulos}@nottingham.ac.uk

Abstract

We propose an end-to-end deep learning architecture for word-level visual speech recognition. The system is a combination of spatiotemporal convolution, residual and bidirectional Long-Short Term Memory networks. We trained and evaluated it on the Lipreading In-The-Wild benchmark, a challenging database of 500-size vocabulary consisting of video excerpts from BBC TV broadcasts. The proposed network attains word accuracy equal to 83.0%, yielding 6.8% absolute improvement over the current state-of-the-art.

Index Terms: visual speech recognition, lipreading, deep learning

1. Introduction

Visual speech recognition (also known as lipreading) is a field of growing attention. It is a natural complement to audio-based speech recognition that can facilitate dictation in noisy environments and enable silent dictation in offices and public spaces. It is also useful in applications related to improved hearing aids and biometric authentication, [1]. Lipreading is the field where the speech recognition and computer vision communities meet each other and combine the advances of each field. The tremendous success of deep learning in both fields has already affected visual speech recognition, by shifting the research direction from handcrafted features and HMM-based models to deep feature extractors and end-to-end deep architectures. Recently introduced deep learning systems beat human lipreading experts by a large margin, at least for the constrained vocabulary defined by each database, [1] [2].

One way to categorize visual and audio-visual speech recognition approaches is (i) to those that model words (e.g. [3] [4]) and (ii) to those that model visemes (e.g. [1] [2]), i.e. visual units that correspond to sets of visually indistinguishable phonemes, [5]. The former approach is considered more pertinent to tasks like isolated word recognition, classification and detection, while the latter to sentence-level classification and large vocabulary continuous speech recognition (LVCSR). Nevertheless, recent advances in speech recognition and natural language processing show that direct modeling of words is feasible even for LVCSR, [6].

The proposed system belongs to the former category, although it can support viseme-level recognition by using viseme instead of word labels at the SoftMax layer. It combined three sub-networks: (i) The front-end, which applies spatiotemporal convolution to the frame sequence, (ii) a Residual Network (ResNet) that is applied to each time step, and (iii) the back-end, which is a two-layer Bidirectional Long-Short Term Memory (Bi-LSTM) network. The SoftMax layer is applied to all time steps and the overall loss is the aggregation of the per time step losses. Finally, the system is trained in an end-to-end fashion.

The rest of the paper is organized as follows. In Section 2, we refer to recent works on visual speech recognition, with emphasis on those that apply deep learning methods. The Lipreading In-The-Wild (LRW) database is discussed in Section 3, while in Section 4 we present analytically the proposed model, together with some useful detail about preprocessing and implementation. Finally, in Section 5 we present our experimental results, together with baseline and state-of-the-art results.

2. Related work

Prior to the advent of deep learning ([7]) most of the work in lipreading was based on hand-engineered features, that were usually modeled by HMM-based pipeline, [8] [9] [10] [11] [12]. Spatiotemporal descriptors such as active appearance models and optical flow, and SVM classifiers have also been proposed, [13]. For an analytic review on traditional lipreading methods we refer to [14] and [15]. More recent works deploy deep learning methods either for extracting "deep" features ([16] [17] [18]) or for building end-to-end architectures. In [19], Deep Belief Networks were deployed for audio-visual recognition and 21% relative improvement was reported over a baseline multi-stream audio-visual GMM/HMM system. In [20], bottleneck features are extracted using Deep Autoencoder. The bottleneck features are concatenated with DCT features and the overall system is trained jointly using an LSTM back-end. In [3], a fully LSTM architecture is proposed, which attains superior results compared to traditional methods on the GRID audiovisual corpus, [21]. In [1], an end-to-end sentence-level lipreading network (LipNet) is introduced, that combines spatiotemporal convolutional layers, LSTMs and Connectionist Temporal Classification (CTC, [22]). It attains 95.2% sentence-level accuracy on a subset of speakers from GRID database, while trained on the remaining GRID speakers. Finally, in [2], the encoder-decoder with attention mechanism is explored, in both audio-visual and visual settings. Using solely visual information, 97.0% word accuracy is reported on GRID and 76.2% word accuracy on LRW. To the best of our knowledge, the latter results define the current state-of-the-art for both databases, insofar as additional training resources may be leveraged, [2].

3. Database

We train and evaluate the algorithm on the challenging LRW database, [4]. The database consists of audiovisual speech segments extracted from BBC TV broadcasts (News, Talk Shows, a.o.) and it is characterized by its high variability with respect to speakers and pose. Moreover, the number of target works is 500, which is an order of magnitude higher than other publicly available databases (GRID [21], CUAVE [23], a.o.). Another feature that renders the database so challenging is the existence of pairs of words that share most of their visemes. Such examples are nouns in both singular and plural forms (e.g. benefit-benefits, 23 pairs) as well as verbs in both present and

Figure 1: *Random frames from the LRW database*

past tenses (e.g. allow-allowed, 4 pairs). Furthermore, the fact that the clips are excerpts from continuous speech makes correct classification of such words even harder, due to co-articulation with the succeeding word.

The collection of the database was fully automatic, involving OCR on the subtitles, synchronization with the audio (forced alignment), as well as verification that the speaker is visible (see [4] for a detailed description). The training set consists of up to 1000 occurrences per target word, while the validation and evaluation sets both consist of 50 occurrences per word. Each clip is of fixed duration (31 frame with 25fps frame rate) and it usually contains co-articulation from preceding and succeeding words. Random frames from the database are depicted in Fig. 1.

4. Deep Learning modeling and preprocessing

4.1. Facial landmarks and data augmentation

In the first preprocessing step, we discard redundant information in order to focus on the mouth region. To do so, we use the 2D version of the algorithm proposed in [24]. The algorithm tackles regression in two steps. It first applies detection to extract a set of heatmaps (one per landmark) which are used as side information for the subsequent regression network. Based on the 66 facial landmarks, we crop the images and resize them to a fixed 112×112 size. A common cropping is applied to all frames of a given clip, using the median coordinates of each landmark. The frames are transformed to grayscale and are normalized with respect to the overall mean and variance. Finally, data augmentation is performed during training, by applying random cropping (±5 pixels) and horizontal flips, that are common across all frames of a given clip.

4.2. Spatiotemporal front-end

The first set of layers applies spatiotemporal convolution to the preprocessed frame stream. Spatiotemporal convolutional layers are capable of capturing the short-term dynamics of the mouth region and are proven to be advantageous, even when recurrent networks are deployed for back-end, [1]. They consist of a convolutional layer with 64 3-dimensional (3D) kernels of 5×7×7 size (time/width/height), followed by Batch Normalization (BN, [25]) and Rectified Linear Units (ReLU). The extracted feature maps are passed through a spatiotemporal max-pooling layer, which drops the spatial size of the 3D feature maps. The number of parameters of the spatiotemporal front-end is ∼16K.

4.3. Residual Network

The 3D features maps are passed through a residual network (ResNet, [26]), one per time-step. We use the 34-layer identity-mapping version, which was proposed for ImageNet, [27]. Its building blocks are composed of two convolutional layers, and with BN and ReLU, while the skip connections facilitate information propagation, [27]. The ResNet drops progressively the spatial dimensionality with max pooling layers, until its output becomes a single dimensional tensor per time step. We should emphasize that we did not make use of pretrained models, as they are optimized for completely different tasks (static colored images from ImageNet or CIFAR). The number of parameters of the ResNet is ∼21M.

4.4. Bidirectional LSTM back-end and optimization criterion

The back-end of the model is a Bidirectional LSTM network. For each of the two directions, we stack two LSTMs, and the outputs of the final LSTMs are concatenated. The number of parameters of the LSTM back-end is ∼2.4M.

When using word-level recognition without explicit modelling of visemes, several approaches exist in terms of the optimization criterion. One approach is to place the SoftMax layer at the last time step of the LSTM output, i.e. when the overall sequence is encoded by the LSTM. Backpropagation through time is capable of propagating the errors all the way back to the first time step of the sequence, given the resiliency of LSTM to the problem of vanishing gradients, [3]. A second approach is to apply the criterion for each time step. This approach is closer to the typical use of LSTMs in speech recognition, where instead of phoneme/viseme labels, the word label is repeated at every time step. This approach fits well to bidirectional LSTMs, since hidden states have in all time steps access to the overall video, [28]. After experimentation with both approaches, we concluded that the latter leads to much higher word accuracy. Hence, the overall loss is defined as the aggregated loss over all time steps, which coincides to the summation of negative logarithm of word posteriors, i.e.

$$L(\mathbf{X}; w) = -\sum_t \log p(\hat{y}_t = w | \mathbf{X}) \qquad (1)$$

where $\hat{y}_t$ is the estimated label at time step t, w is the word label and $\mathbf{X}$ is the collection of frames of the clip.

4.5. Implementation details

Our implementation is based on Torch7 ([29]) and the networks are trained on a NVIDIA Titan X (Pascal Architecture) GPU with 12GB memory. We use the standard SGD training algorithm, with momentum 0.9. BN follows all convolutional and linear layers, apart from the one preceding the SoftMax layer. We do not apply dropout, since it is not part of the ResNet training recipe (BN seems to suffice, [25]). The initial learning rate is 3×10^{-3} and is decreasing on log scale. Training is considered completed when the results on the validation set do no longer improve, with a delay of 3 epochs.

The overall network is trained end-to-end using stochastic gradient descent (SGD). A block-diagram of the network is depicted in Fig. 2. BN layers have been omitted for clarity. The size of the tensors each layer outputs is also presented. For the 3D-convolutional front-end, tensor dimensions denote channels, time, width and height.
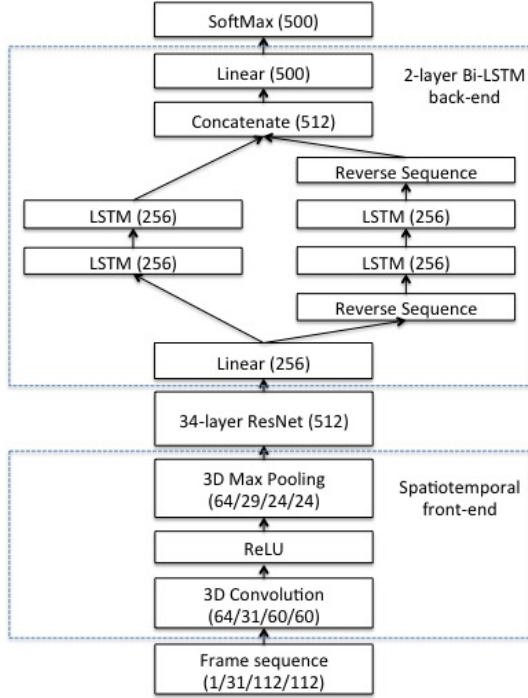
Figure 2: *The block-diagram of the proposed network.*

We should emphasize that although the overall system can be directly trained end-to-end, we use the following three steps approach. Initially, a temporal convolutional back-end is used instead of the Bi-LSTM. After convergence, the temporal convolutional back-end is removed and the Bi-LSTM back-end is attached. The Bi-LSTM is trained for 5 epochs, keeping the weights of the 3D convolution front-end and the ResNet fixed. Finally, the overall system is trained end-to-end. A comparison between the two back-ends is presented in Section 5.

5. Experiments

5.1. Baseline results

The best baseline result published in [4] is the multi-tower VGG-M. It consists of a set of parallel VGG models (towers) with shared weights, which are concatenated channel-wise using pooling, while the rest of the network is the same as the regular VGG-M. The results are presented in Table 1 in terms of word accuracy. Top-1 corresponds to the percentage of times the word was correctly identified, while more generally Top-N corresponds to the percentage of times the correct word was among the N best scores.

Network	Top-1	Top-5	Top-10
Baseline	61.1%	-	90.4%

Table 1: *Word accuracies for the baseline network (VGG-M).*

In [2], an attentional encoder-decoder architecture is proposed, [30]. It is trained on a different set of BBC TV Broadcasts, which contains whole sentences rather than isolated words. A visual-only version of the system (termed "Watch, Attend and Spell", WAS) is evaluated on GRID and on LRW. The network is pretrained on the BBC TV Broadcasts, while the

training sets of GRID and LRW are used for fine-tuning. Word accuracies (Top-1) equal to 97.0% and 74.2% are reported on GRID and LRW respectively, which according to our knowledge represent the current state-of-the-art on both databases.

5.2. Results using our network

We begin by using a simpler model than the proposed one, in order to examine the contribution of each individual component of the network. The first network applied 2D convolution instead of 3D. The 2D convolution is followed by the ResNet, while the back-end is based on temporal convolution rather than LSTMs. More specifically, we use two temporal convolutional layers, each of which is followed by BN, ReLU and Max Pooling which reduce the temporal dimensions by a factor of 2. Finally, a Mean Pooling layer is added, followed by a linear and a SoftMax layer. The results are presented in Table 2 (denoted by N1). The results of the same model, but with 3D convolution are also presented in Table 2 (denoted by N2).

In order to verify the effectiveness of the ResNet we replace it with a Deep Neural Network (DNN) of approximately the same number of parameters ($\sim$20M). The DNN is composed of 3 fully connected hidden layers, with BN and ReLU. Its inputs are 3D convolutional maps, treated as vectors (one per time step). The DNN progressively reduces the size of the vectors as $50176 \rightarrow 384 \rightarrow 384 \rightarrow 256$. The results are presented in Table 2 (denoted by N3).

Network	Top-1	Top-5	Top-10
N1	69.6%	90.4%	94.8%
N2	74.6%	93.4%	96.5%
N3	69.7%	90.5%	94.6%

Table 2: *Word accuracies using temporal convolution back-end.*

We now focus on the back-end of the network and use LSTMs instead of temporal convolutions. The first network in Table 3 (denoted by N4) uses a single-layer Bi-LSTM, while the second one (denoted by N5) uses a double-layer Bi-LSTM. These two networks are not trained end-to-end. While training the back-end, the 3D convolutional layer and the ResNet (that are copied from N2) remain fixed. Moreover, the outputs of the two directional LSTMs are added together instead of concatenated together.

Network	Top-1	Top-5	Top-10
N4	78.4%	94.9%	97.4%
N5	79.6%	95.3%	96.3%

Table 3: *Word accuracies using different LSTM in the back-end.*

For the final set of results we use end-to-end training of the overall network. The first network in Table 4 (denoted by N6) is the same as N5, but trained end-to-end, using the weights of N5 as starting point. Finally, N7 is also trained end-to-end and the sole difference with N6 is that the outputs of the two directional LSTMs are concatenated together instead of added together (as depicted in Fig. 2).

5.3. Discussion and error analysis

Several conclusions can be drawn from the results presented above (see also Fig. 3 for clarity). First of all, by comparing the

Network	Top-1	Top-5	Top-10
N6	81.5%	96.0%	98.0%
N7	83.0%	96.3%	98.3%

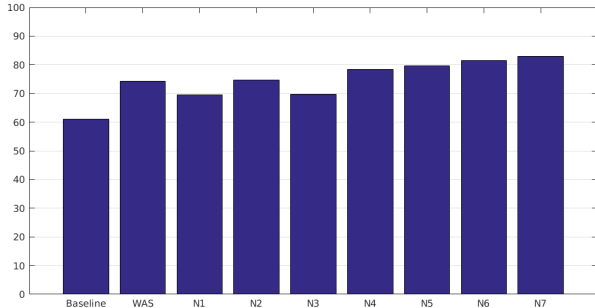Table 4: *Word accuracies using LSTM back-end and end-to-end training.*

Figure 3: *Word accuracy of the networks examined.*

baseline to N1, we observe that the our simplest system yielding 8.5% absolute improvement over the VGG-M baseline. Moreover, the use of 3D (N2) instead of 2D (N1) leads to a further 5.0% absolute improvement, emphasizing the need of modeling the short-term dynamics of the mouth region in the front-end. By comparing N2 and N3 we observe that the ResNet yields 4.9% better work accuracy compared to a 3-layer DNN with the same number of parameters. In addition, by using a single-layer Bi-LSTM (N4) instead of a temporal convolutional back-end, a further 3.8% absolute improvement is attained, highlighting the expressive power of LSTMs in modeling temporal sequences. Furthermore, the use of a two-layer Bi-LSTMs (N5) offers a further 1.2% absolute improvement. The final set of results demonstrates the importance of end-to-end training towards achieving higher word accuracy. By training N5 in an end-to-end fashion (N6) we obtain a 1.9% absolute improvement, while by concatenating (N7) rather than adding together (N6) the Bi-LSTM outputs we obtain our best result, a 83.0% work accuracy.

Target Word	Decision	Error Rate (%)
SPEND	SPENT	26
WANTS	WANTED	18
RUSSIAN	RUSSIA	18
BENEFIT	BENEFITS	18
BENEFITS	BENEFIT	16
RUSSIA	RUSSIAN	16
CANCER	AGAINST	16
GIVING	LIVING	16
DIFFERENCE	DIFFERENT	14
MAKES	MEANS	14

Table 5: *Most frequent errors made by the proposed system.*

Table 5 contains the most frequent errors made by our best system (N7). We observe that most of the word pairs are mutually close with respect to their phonetic and "visemic" content. We should emphasize again that the clips contain coarticulation with preceding and succeeding words, as they are excerpted from continuous speech. Hence, correct identification of the first and last visemes of a word is occasionally hard.

The list of words for which the system yields the best and worst performance is presented in Table 6. As expected, the system does very well on words with rich phonetic/visemic content and vice versa. There are 8 words for which the system made no errors, and only 3 words for which the word accuracy dropped below 50%. Recall that the number of evaluation clips is 50 per target word (i.e. 25000 clips overall).

Target Word	Acc (%)	Target Word	Acc (%)
SUNSHINE	100	SPEND	58
ECONOMIC	100	AROUND	58
TEMPERATURES	100	THING	56
WESTMINSTER	100	THEIR	56
POLITICIANS	100	UNTIL	54
SITUATION	100	GETTING	52
OBAMA	100	SAYING	50
INQUIRY	100	THERE	48
MINISTER	98	GOING	48
FAMILIES	98	UNDER	42

Table 6: *Words with the highest accuracy (left) vs. words with the lowest accuracy (right).*

6. Conclusions

We proposed a spatiotemporal deep learning network for word-level visual speech recognition. The network is a stack of a 3D convolutional front-end, a ResNet and an LSTM-based back-end, and trained using an aggregated per time step loss. We chose to experiment with the LRW database, since it combines many attractive characteristics, such as large size ($\sim$500K clips), high variability in speakers, pose and illumination, and non-laboratory in-the-wild conditions. We explored several network configurations, and we demonstrated the importance of each building block of the network as well as the gain in performance attained by training the network end-to-end. The proposed network yielded 83.0% work accuracy, which corresponds to less that half the error rate of the baseline VGG-M network and 6.8% absolute improvement over the state-of-the-art 76.2% accuracy, attained by an attentional encoder-decoder network, [2] [4].

7. Acknowledgements

8. References

[1] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," *arXiv preprint arXiv:1611.05358*, 2016.

[3] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6115–6119.

[4] J. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[5] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, no. 4, pp. 796–804, 1968.

[6] S. Bengio and G. Heigold, "Word embeddings for speech recognition." in *INTERSPEECH*, 2014, pp. 1053–1057.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *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, pp. 82–97, 2012.

[8] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Continuous automatic speech recognition by lipreading," in *Motion-Based recognition*. Springer, 1997, pp. 321–343.

[9] G. I. Chiou and J.-N. Hwang, "Lipreading from color video," *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1192–1195, 1997.

[10] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[11] C. Chandrasekaran, A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar, "The natural statistics of audiovisual speech," *PLoS Comput Biol*, vol. 5, no. 7, 2009.

[12] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 423–435, 2009.

[13] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. C. Azemin, and J. Gubbi, "Lip reading using optical flow and support vector machines," in *3rd International Congress on Image and Signal Processing (CISP)*, vol. 1. IEEE, 2010, pp. 327–330.

[14] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and vision computing*, vol. 32, no. 9, pp. 590–605, 2014.

[15] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, p. 23, 2004.

[16] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[17] K. Thangthai, R. W. Harvey, S. J. Cox, and B.-J. Theobald, "Improving lip-reading performance for robust audiovisual speech recognition using dnns." in *AVSP*, 2015, pp. 127–131.

[18] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2722–2726.

[19] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7596–7599.

[20] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2304–2308.

[21] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[23] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Cuave: A new audio-visual database for multimodal human-computer interface research," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2. IEEE, 2002, pp. II–2017.

[24] A. Bulat and G. Tzimiropoulos, "Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3DFAW) challenge," in *European Conference on Computer Vision*. Springer, 2016, pp. 616–624.

[25] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 448–456.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] ——, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 630–645.

[28] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *International Conference on Artificial Neural Networks*. Springer, 2005, pp. 799–804.

[29] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, no. EPFL-CONF-192376, 2011.

[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention." in *ICML*, vol. 14, 2015, pp. 77–81.