Personalizing ASR for Dysarthric and Accented Speech with Limited Data

Joel Shor[1], Dotan Emanuel[1], Oran Lang[1], Omry Tuval[1], Michael Brenner[1], Julie Cattiau[1],
Fernando Vieira[2], Maeve McNally[2], Taylor Charbonneau[2], Melissa Nollstadt[2], Avinatan Hassidim[1],
Yossi Matias[1]

[1]Google, United States
[2]ALS Therapy Development Institute, United States
{joelshor, dotan, oranl}@google.com

Abstract

Automatic speech recognition (ASR) systems have dramatically improved over the last few years. ASR systems are most often trained from typical speech, which means that underrepresented groups dont experience the same level of improvement. In this paper, we present and evaluate finetuning techniques to improve ASR for users with non-standard speech. We focus on two types of non-standard speech: speech from people with amyotrophic lateral sclerosis (ALS) and accented speech. We train personalized models that achieve 62% and 35% relative WER improvement on these two groups, bringing the absolute WER for ALS speakers, on a test set of message bank phrases, down to 10% for mild dysarthria and 20% for more serious dysarthria. We show that 71% of the improvement comes from only 5 minutes of training data. Finetuning a particular subset of layers (with many fewer parameters) often gives better results than finetuning the entire model. This is the first step towards building state of the art ASR models for dysarthric speech.

Index Terms: speech recognition, personalization, accessibility

1. Introduction

State-of-the-art speaker-independent ASR systems are made possible by large datasets and data-driven algorithms. This setup works well when the dataset contains a large amount of data from all types of voices that one would want the system to recognize, but fails on groups not well-represented in the data: ASR models trained on thousands of hours of 'typical users' recognize voices in the typical use-case, but often fail to recognize accented voices or ones impacted by medical conditions.

Getting ASR to work on non-standard speech is difficult for a few reasons. First, it can be difficult to find enough speakers to train a state-of-the-art model. Second, individuals within a group like 'ALS' or a particular accent can have very different ways of speaking. This paper's approach overcomes data scarcity by beginning with a base model trained on thousands of hours of standard speech. It gets around sub-group heterogeneity by training personalized models.

This paper details the technical approach to Project Euphonia, an accessibility project announced at Google I/O 2019 [1, 2]. To demonstrate our approach, we focus on accented speech from non-native English speakers, and speech from individuals living with ALS. ALS is a progressive neurodegenerative disorder that affects speech production, among other things. About 25% of people with ALS experience slurred speech as their first symptom [3]. Most people with ALS eventually lose mobility, so they would especially benefit from being able to interact verbally with smart home devices. The severity of dysarthria for people with ALS is measured with an FRS

score, ranging from 0 to 4, with 0 being incomprehensible and 4 normal [4]. In this paper, we improve state-of-the-art ASR for individuals with ALS FRS 1-3 and heavy accents. The main contributions of this paper are:

1. A finetuning technique for personalizing an ASR model on dysarthric speech that yields word error rates on a test set of message bank phrases [5] of 10.8% for mild dysarthria (FRS of 3) and 20.9% for severe dysarthria (FRS of 1-2). The finetuning technique yields a word error rate of 8.5% on heavily accented data. The finetuned models make recognition mistakes that are distributionally more similar to standard ASR mistakes on standard speech.

2. Demonstrating that the finetuning gives significant improvement over the base model in multiple scenarios, including different non-standard speech (dysarthric and accented), and on different architectures.

3. On the ALS speech, 71% of the relative WER improvement can be achieved with 5 minutes of data, 80% with 10 minutes of data. Furthermore, for some models, just training the encoder produces better results, and just training the layers closest to the input yields 90% of total relative improvement.

2. Related Work

Neural networks are the state-of-the-art systems for ASR in the large data-regime. We explore how well two particular architectures can be finetuned on a small amount of non-standard data. The RNN-Transducer [6, 7] is a neural network architecture that has shown good results on numerous ASR tasks. It consists of an encoder and decoder network and is configured such that unidirectional models can perform streaming ASR. In this paper, we use a bidirectional encoder without attention that was shown to achieve comparable results [8].

We also explore the Listen, Attend, and Spell architecture. It is an attention-based, sequence-to-sequence model that maps sequences of acoustic properties to sequences of language [9]. It has produced state-of-the-art results on a challenging 12,500 hour voice-search ASR task, achieving a 4.1% WER [10]. This model uses an encoder to convert the sequence of acoustic frames to a sequence of internal representations, and a decoder with attention to convert the sequence of internal representations to linguistic output. The best network in [10] produced word pieces, which are a linguistic representation between graphemes and words [11].

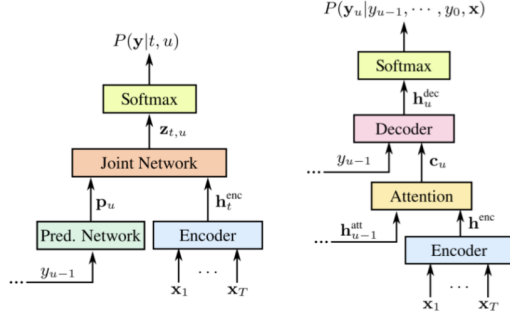There are a number of methods for adapting large ASR models to small amounts of data [12, 13, 14]. This paper's ap-

Figure 1: *Schematic diagrams of the RNN-T architecture (left) and the LAS architecture (right). Diagrams come from [8].*

proach is most similar to [15], which involves selectively finetuning parts of the ASR model.

There are also many published techniques on how to improve ASR specifically for pathological, or dysarthric, speech: [16] appended articulatory features to the usual acoustic ones to achieve a 4-8% relative WER improvement, and [17] adapts ASR models trained on open source datasets to a dataset of dysarthric speech. However, all these approaches are limited either by the quality of the base model or the amount of data available for finetuning.

Numerous studies have explored the acoustic, articulatory, and phonetic differences between standard and modified speech. Some of the conditions explored are Parkinsons [18], age [19], dyslexia [20], and ALS [21]. We add to this body of work by describing the phonetic mistakes that a production ASR system makes on a large collection of ALS audio. We are also able to describe our improved ASR model by characterizing which phonemes the improved models are better able to recognize.

3. Experiments

We create personalized ASR models by starting with a base model trained on standard, unaccented speech. This approach is much more resource efficient than retraining the entire model, from scratch, for each speaker. In order to verify that we are learning something other than just the idiosyncrasies of a particular model, we run most of our experiments starting from two different models: a Bidirectional RNN Transducer (RNN-T) model [6], and a Listen, Attend, and Spell (LAS) model [9]. Both are end-to-end sequence-to-sequence models. We follow the training procedure in [8] for the RNN-T model and [10] for the LAS model (on the 1000 hour, open source Librispeech dataset [22]).

We finetune different layer combinations on different amounts of data, on both ALS and accents datasets. We finetune models per speaker.

3.1. Data

3.1.1. ALS

We collected 36.7 hours of audio from 67 people with ALS, in partnership with the ALS Therapy Development Institute (ALS-TDI). The participants were given sentences to read, and they recorded themselves on their home computers using custom software. The sentences were collected from three sources: The Cornell Movie-Dialogs Corpus [23], a collection of sentences

used by text-to-speech voice actors, and a modified selection of sentences from the Boston Children's Hospital [5]. Note that this corpus is a restricted language domain, but is phonetically very similar to other corpora e.g. Librispeech. The FRS scores of participants were measured by ALS-TDI, and we only evaluate on people with speech FRS 3 and below (17 speakers, 22.1 hours). See the attached multimedia file for audio examples.

3.1.2. Accented Speech

To test our finetuning method on another type of non-standard speech, we use the L2 Arctic dataset of non-native speech [24]. This dataset consists of 20 speakers with approximately 1 hour of speech per speaker. Each speaker recorded a set of 1150 phonetically balanced utterances. For each of the 20 speakers, we split the data into 90/10 train and test. All of the sentences which contains proper nouns are used in the training set, in order to remove the possibility of the model to artificially achieve better results on the test set by memorizing them.

3.2. Base Models

All our base networks and finetuning are trained on 80-bin logmel spectrograms computed from a 25ms window and a 10ms hop. We use the same technique presented in [25] and used in [8]: we stack frames in groups of 3 and process them as one 'super-frame.'

We trained RNN-T and LAS architectures (Figure 1). All training was performed with the multicondition training (MTR) techniques described in [8]. During training, we distorted the audio using a room simulator derived from YouTube data. The average SNR of the added noise is 12dB. We use the TensorFlow library Lingvo [26].

3.2.1. Bidirectional RNN-Transducer

In this paper, we primarily work with a bidirectional RNN-Transducer (RNN-T) architecture that achieves near state-of-the-art performance. This architecture was first introduced in [6]. We use the version presented in [8]. The network maps acoustic frames to word pieces, which are a linguistic representation between graphemes and words [11]. It has a 5 layer bidirectional convolutional LSTM encoder, a 2 layer LSTM decoder, and a joint layer. It has 49.6M parameters in total.

3.2.2. Listen, Attend, and Spell model

To verify which results generalize beyond a particular architecture, we also run some of our experiments with a Listen, Attend, and Spell (LAS) model trained on the open source Librispeech dataset [22]. This architecture was first introduced in [9]. The model that we use is described in [10]. It is a sequence-to-sequence with attention model that maps acoustic frames to graphemes. We use an encoder with 4 layers of bidirectional convolutional LSTMs and a 2 layer RNN decoder. The model has a total of 132M parameters.

Our grapheme targets are the 26 English lower-case letters, punctuation symbols, and a space. There are 33 target grapheme symbols. The base LAS model was trained to 1M steps on all 960 hours of the Librispeech dataset. It achieved a 5.5% WER on the clean test split and 15.5% on the non-clean test split (the two standard test splits given in the dataset).

Table 1: *Average WER Improvements*

	Cloud	RNN-T		LAS	
		Base	Finetune	Base	Finetune
Arctic[1]	24.0	13.3	8.5	22.6	11.3
ALS[2]	42.7	59.7	20.9	86.3	31.3
ALS[3]	13.1	33.1	10.8	49.6	17.2

[1] Non-native English speech from the L2-Arctic dataset. [24]
[2] Low FRS (ALS Functional Rating Scale) intelligible with repeating, Speech combined with nonvocal communication.
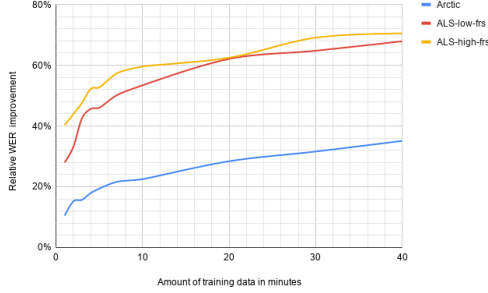[3] FRS-3 detectable speech disturbance. [4]

Figure 2: *Average Relative WER improvement as a function of the amount of training data.*

3.3. Finetuning

All our finetuning uses four Tesla V100 GPUs for no more than four hours.

3.3.1. RNN-T

We started by finetuning 1, 2, and 3 layers in fixed combinations (treating the decoder as a single layer), on both datasets, adjusting hyperparameters as necessary. Let E_i denote the ith layer of the encoder, where lower-numbered layers are closer to the input. We uniformly found that training from E_0 up with or without the joint layer was always better than the other methods, so we focused our search on training from E_0 up, with or without the joint layer.

Next, for each ALS and accented individual, we exhaustively searched our finetuning space, with various amounts of data. For the RNN-T, this meant finetuning each of E_0, E_0-E_1, E_0-E_2, etc.. the entire encoder, with or without the joint layer.

3.3.2. LAS

For LAS architectures, we finetuned various layer combinations and consistently found that the best results from this network came from finetuning all layers. All results reported in this paper for the LAS network are on finetuning the entire network, unless otherwise specified.

4. Results

4.1. Performance

We report our absolute word error rate in Table 1. The results show dramatic improvement over Google cloud ASR [27] model for very non-standard speech (heavy accents and ALS
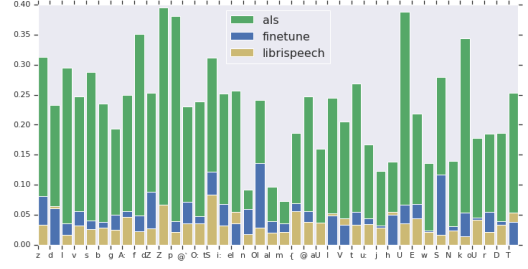

Figure 3: *The distribution of phoneme mistakes before and after finetuning. x-axis is SAMPA phoneme. y-axis is number of times that phoneme was deleted or substituted divided by number of phoneme occurances in the ground truth transcripts.*

speech below 3 on the ALS Functional Rating Scale [4]) and moderate improvements in ALS speech that is similar to healthy speech. The comparison to Google cloud ASR demonstrates that a healthy speech model finetuned on non-standard speech produces strong results, but we acknowledge that the comparison is not perfect: the LAS and RNN-T models use the entire audio to make predictions while Google cloud ASR model only looks backwards in time. This lets Google cloud ASR support streaming, but it is also less accurate.

The relative difference between base model and the finetuned model demonstrates that the majority of the improvement comes from the finetuning process, except in the case of the RNN-T on the Arctic dataset (where the RNN-T baseline is already strong).

4.2. Limited Data

On the ALS dataset, finetuning on five minutes and ten minutes of data yields 75% and 85% of the WER improvement compared to a model trained on 40 min of data, respectively. On the Arctic dataset, fourteen and twenty minutes of training data yields 70% and 81% of the WER improvement, respectively. Figure 2 shows more details.

4.3. Layers

The LAS model consistently performed best when the entire network was finetuned. The RNN-T model achieved 91% of the relative WER improvement by just finetuning the joint layer and first layer of the encoder (compared to finetuning the joint layer and the entire encoder, the average WER across all participants regardless of FRS score was 18.1% vs 15.1%). On Arctic, finetuning the joint layer and the first encoder layer achieved 86% of the relative improvement compared to finetuning the entire network (11.0% vs 10.5%).

4.4. Phoneme mistakes

To better understand how our models improved, we looked at the pattern of phoneme mistakes. We started by comparing the distribution of phoneme mistakes made by Google cloud ASR model on standard speech (Librispeech) to the mistakes made on ALS speech. We map the ground truth transcripts and the ASR model outputs to sequences of SAMPA phonemes, then compute the edit distance between the two phoneme sequences. We aggregate over speakers and utterances.

First, we compute the probability of a mistake for a particular phoneme in the ground truth transcript. Since the standard

speech and ALS speech transcripts are different, we normalize by the number of ground truth transcript phonemes. The phonemes with the five largest differences between the ALS data and standard speech are p, U, f, k, and Z. These five account for 20% of the likelihood of a deletion mistake. Next, we investigate which phonemes are mistakenly added. We compute the probability that a particular phoneme will mistakenly appear in the recognized transcript. By far the biggest differences are in the n and m phonemes, which together account for 17% of the insertion / substitution mistakes.

Finally, we perform the same analysis on some of our finetuned models. The unrecognized phoneme distribution becomes more similar to that of standard speech (see Figure 3). Also, surprisingly, the phoneme distribution that the model produced when it made a mistake was much more similar to mistakes on standard speech after finetuning (KL of 0.26 between standard and ALS, 0.10 after finetuning).

5. Discussion

In this paper, we develop well-performing ASR models for dysarthric and heavily accented individuals by carefully finetuning healthy-speech models. Specifically, we demonstrated:

1. Good absolute performance on average dysarthric ALS speakers, large improvements in very dysarthric speakers

2. Better performance on ALS and accented speech when just training the RNN-T encoder

3. Much of the improvement is from the first 5-10 minutes of training data, and can be achieved by just training the first encoder layer and the joint layer

4. The five most mistaken phonemes in ALS speech account for 20% of the mistakes

Prior to this work, it was unclear how much improvement could be achieved by finetuning on small amounts of non-standard speech. We show that with on the order of 1 hour of data, we can create a personalized ASR model that is significantly better than Cloud-based services. This improvement comes mainly from the finetuning process: we see a relative WER improvement over our base model of 70% for dysarthric speech (for both groups) and 35.1% for accented speech. The finetuning process used four GPUs for fewer than four hours, making this technique very accessible.

We achieve better performance by just training the encoder for the RNN-T architecture but not the LAS architecture. This is likely due to a peculiarity of the RNN-T: the RNN-T architecture is factorized into a component whose activations are solely a function of the current audio (the encoder) and another component whose activations are a function of the current predicted transcript (the decoder). The LAS model has no such factorization, so information about the acoustics and linguistics are likely more evenly distributed throughout the network. This claim isn't precise, since the RNN-T weights used at inference time are determined by the statistical properties of both language and acoustics, but even minor effects from this factorization might mean that finetuning the encoder only helps prevent the network from overfitting to the language seen in the small amount of training data.

A large fraction of the improvement comes from the first 5-10 minutes of audio. One explanation is that the test and train sentences are linguistically similar enough for the model to learn the kinds of things that are said during testing from a small number of examples. Another explanation is that just finetuning part of the network allows it to retain the general acoustic and linguistic information from the general speech model while needing minimal modifications to adapt to a single new speaker. Future work includes testing this hypothesis, possibly by exploring its performance on radically different sentences.

The model might be adapting to general non-standard speech or to the individual. This could be tested, for instance, by training a single model on the entire ALS or Arctic corpus, and comparing it to a single speaker model. We didn't include results from this comparison, and we discuss this in the 'Future Works' section.

We found that finetuning just the joint layer and the first encoder layer achieved 90% of the relative improvement compared to training the joint layer and the entire encoder. This can be explained by a combination of the acoustic vs linguistic properties discussed earlier in this section and by analogy with a popular computer vision finetuning technique. Many computer vision papers publish good results by finetuning the last few layers of a classification network that was pretrained on Imagenet [28]. In that technique, authors usually assume that their data follows roughly the distribution of images in imagenet, and that their image label distribution is different. Our problem has the reverse assumption: we assume that the distribution of language (labels) is the same in our problem as in standard speech, but that the distribution of ALS / accented speech is different from that of standard speech.

Lastly, we find that the five most deleted / substituted phonemes account for roughly 20% of such errors, and that the two most incorrectly inserted phonemes account for just over 20% of such errors. This kind of observation might lead to ALS-detection techniques: one can try to detect ALS degradation by matching the prediction distribution of a production ASR system. We might also be able to improve the finetuning process by collecting more speech that involves the most-often-confused phonemes.

6. Future Work

A major challenge is to build state of the art speech recognition models for strongly dysarthric speech. It is an open question whether there are additional techniques that can be helpful in the low data regime (such as Virtual Adversarial Training, data augmentation, etc). We can also use the phoneme mistakes to weight certain examples during training, or to pick training sentences for people with ALS to record that contain the most egregious phoneme mistakes.

We would like to explore pooling data from multiple speakers with similar conditions, but did not do so in this paper. We believe that such an experiment raises more questions around training speaker-independent non-standard speech ALS models, which we feel are outside the scope of this work.

7. Acknowledgements

8. References

[1] Google, "Project euphonia: Helping everyone be better understood," 2019. [Online]. Available: https://www.youtube.com/watch?v=OAdegPmkK-o&t=1s

[2] J. Cattiau, "How ai can improve products for people with impaired speech," 2019. [Online]. Available: https://blog.google/outreach-initiatives/accessibility/impaired-speech-recognition/

[3] https://www.als.net/what-is-als/.

[4] A. C. T. S. A. P. I.-I. S. Group, "The amyotrophic lateral sclerosis functional rating scale: Assessment of activities of daily living in patients with amyotrophic lateral sclerosis," *Archives of Neurology*, vol. 53, no. 2, pp. 141–147, 02 1996. [Online]. Available: https://doi.org/10.1001/archneur.1996.00550020045014

[5] J. Costello, "Message banking, voice banking and legacy messages," *Boston, MA: Boston Children's Hospital*, 2014.

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

[7] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[8] R. Prabhavalkar, K. Rao, T. N. Sainath, B. Li, L. Johnson, and N. Jaitly, "A comparison of sequence-to-sequence models for speech recognition." in *Interspeech*, 2017, pp. 939–943.

[9] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[10] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[11] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 193–199.

[12] K. C. Sim, A. Narayanan, A. Misra, A. Tripathi, G. Pundak, T. N. Sainath, P. Haghani, B. Li, and M. Bacchiani, "Domain adaptation using factorized hidden layer for robust automatic speech recognition," *Proc. Interspeech 2018*, pp. 892–896, 2018.

[13] C. Wu and M. J. Gales, "Multi-basis adaptive neural network for rapid adaptation in speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4315–4319.

[14] K. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," *Proc. ICASSP 2011*, 2011.

[15] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Pre-training on high-resource speech recognition improves low-resource speech-to-text translation," in *Interspeech*, 2018.

[16] E. Yılmaz, V. Mitra, C. Bartels, and H. Franco, "Articulatory features for asr of pathological speech," *arXiv preprint arXiv:1807.10948*, 2018.

[17] M. B. Mustafa, S. S. Salim, N. Mohamed, B. Al-Qatab, and C. E. Siong, "Severity-based adaptation with limited data for asr to aid dysarthric speakers," *PloS one*, vol. 9, no. 1, p. e86285, 2014.

[18] V. Delvaux, K. Huet, M. Piccaluga, S. Van Malderen, and B. Harmegnies, "Towards a better characterization of parkinsonian speech: a multidimensional acoustic study," in *Interspeech*, 2018.

[19] A. Hermes, J. Mertens, and D. Mücke, "Age-related effects on sensorimotor control of speech production," *Proc. Interspeech 2018*, pp. 1526–1530, 2018.

[20] N. D. C. Blanco, M. Hoen, F. Meunier, and J. Meyer, "Phoneme resistance and phoneme confusion in noise: Impact of dyslexia," in *InterSpeech 2018*, vol. 2018, 2018.

[21] N. Nonavinakere Prabhakera, P. Alku *et al.*, "Dysarthric speech classification using glottal features computed from non-words, words and sentences," in *Interspeech*, 2018.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[23] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, 2011, pp. 76–87.

[24] G. Zhao, S. Sonsaat, A. O. Silpachai, I. Lucic, E. Chukharev-Khudilaynen, J. Levis, and R. Gutierrez-Osuna, "L2-arctic: A non-native english speech corpus," *Perception Sensing Instrumentation Lab*, 2018.

[25] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.

[26] J. Shen, P. Nguyen, Y. Wu, Z. Chen *et al.*, "Lingvo: a modular and scalable framework for sequence-to-sequence modeling," 2019.

[27] "Google cloud speech api," https://cloud.google.com/speech-to-text/.

[28] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *arXiv preprint arXiv:1411.1792*, 2014.