How transferable are features in convolutional neural network acoustic models across languages?

Jessica A.F. Thompson[1,2], Yoshua Bengio[2], Marc Schönwiesner[3], Daniel Willett[4]

[1]International Laboratory for Brain, Music and Sound Research
Université de Montréal, Canada

[2]Mila
Montreal, Canada

[3]Institute for Biology
Leipzig University, Germany

[4]Nuance Communications
Aachen, Germany

Introduction

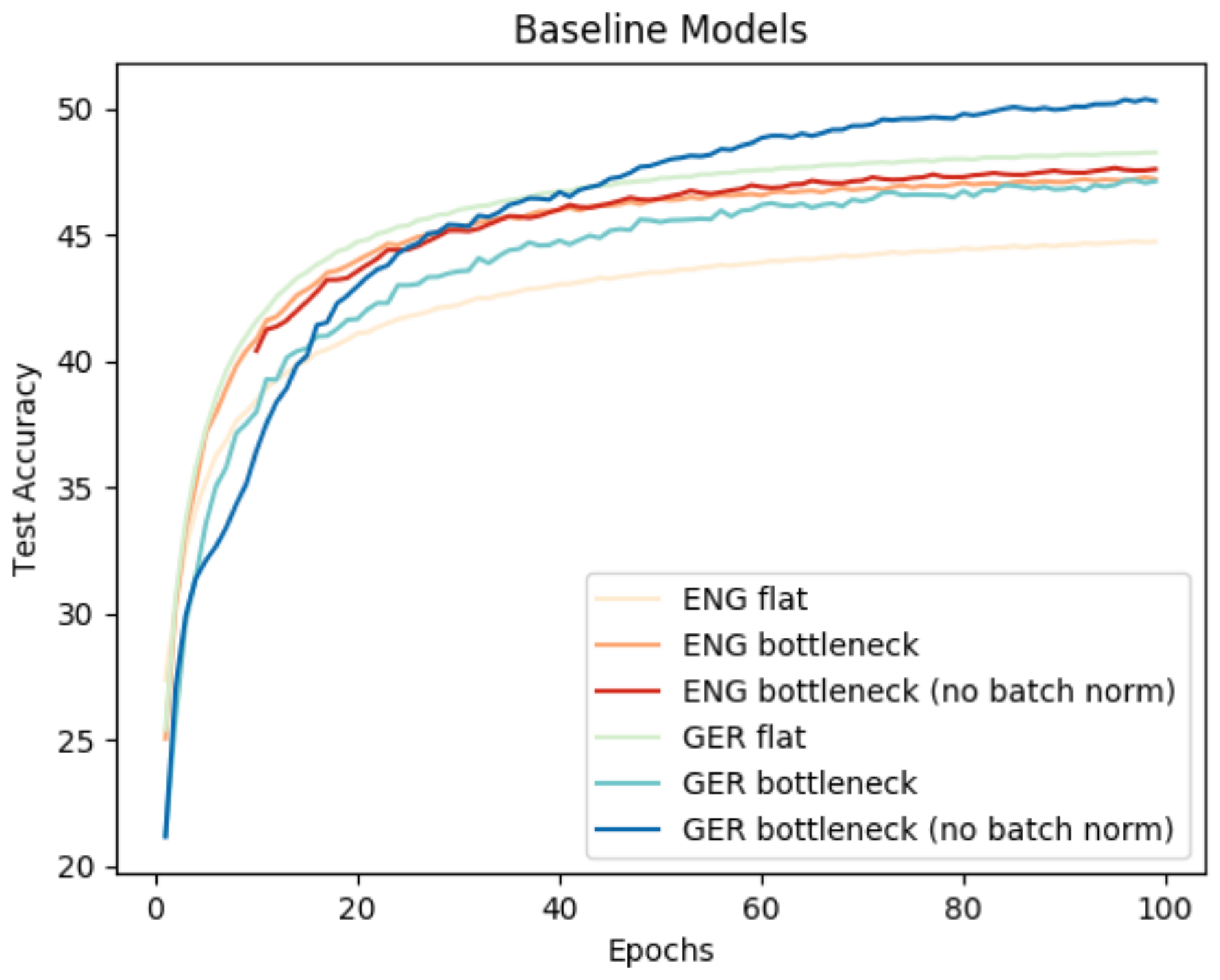
Better characterization of learned representations can help to understand a task and design well-tailored training procedures. We adapt an approach from Yoskinski et al. (2014) to characterize the task specificity at each layer using transferability as a proxy for task specificity.

Application Domain: Acoustic models for automatic speech recognition
What is the most effective way to use speech from multiple sources? For example, how can we best use data from a data-rich language (e.g. American English) to improve performance on other languages? Can we design training procedures that takes advantage of our knowledge about what is being learned at each layer? e.g. Deep Adaptation Networks (Long, Cao, Wang, & Jordan, 2015)

Methods

	English	German
Hours of speech	86	66
Phone set size	54	47
Context-dependent phones	9000	9000

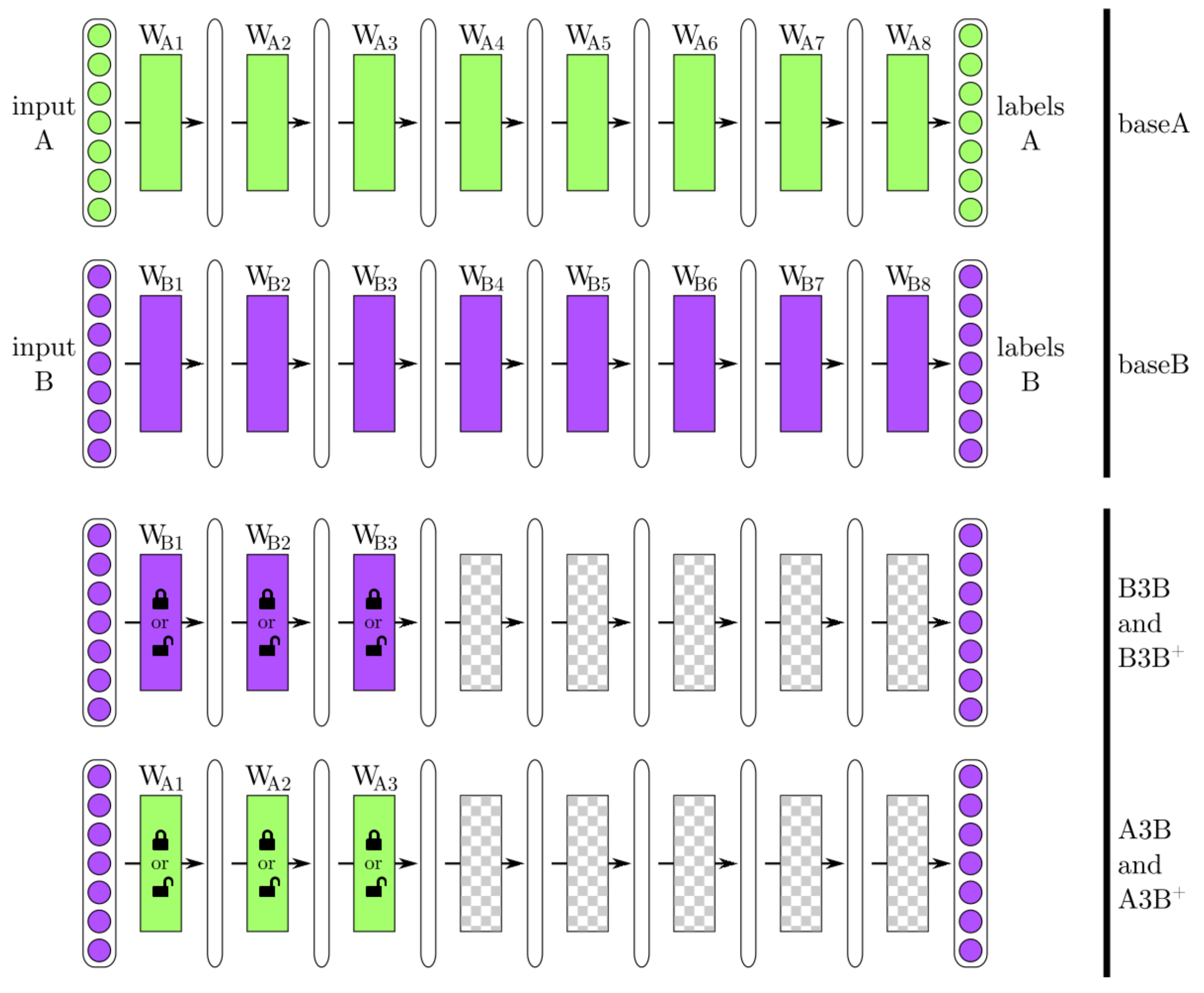
Data:
45D Log mel filter bank features (spectrograms) extracted every 10ms


Baseline Models

Architecture: CNN with 9 conv layers and 3 fully connected layers, resulting in a total of approximately 7.2 million parameters: (7,7, 1024), (3, 3, 256), (3, 3, 256), (3, 3, 128), (3, 3, 128), (3,3, 128), (3, 3, 64), (3, 3, 64), (3, 3, 64), (600), (190), (9000). Trained with ADAM with minibatches of 256 for 100 epochs.

Experimental Setup

Similar to Yosinski et al (2014), several '*network surgeries*' were performed. The first *n* layers of a network trained on language A were implanted into a new network of identical architecture where the layers after layer *n* were randomly initialized. This network was trained in four different ways. It was either trained on language A (***selfer network***) or language B (***transfer network***) and the implanted parameters were either **held fixed** or allowed to be **finetuned** during training. This processwas repeated ∀ 1 ≤ *n* ≤ 11 and for both English and German, resulting in 88 networks total. Each network took 13–16 days to train for 100 epochs with the training distributed across four GPUs.



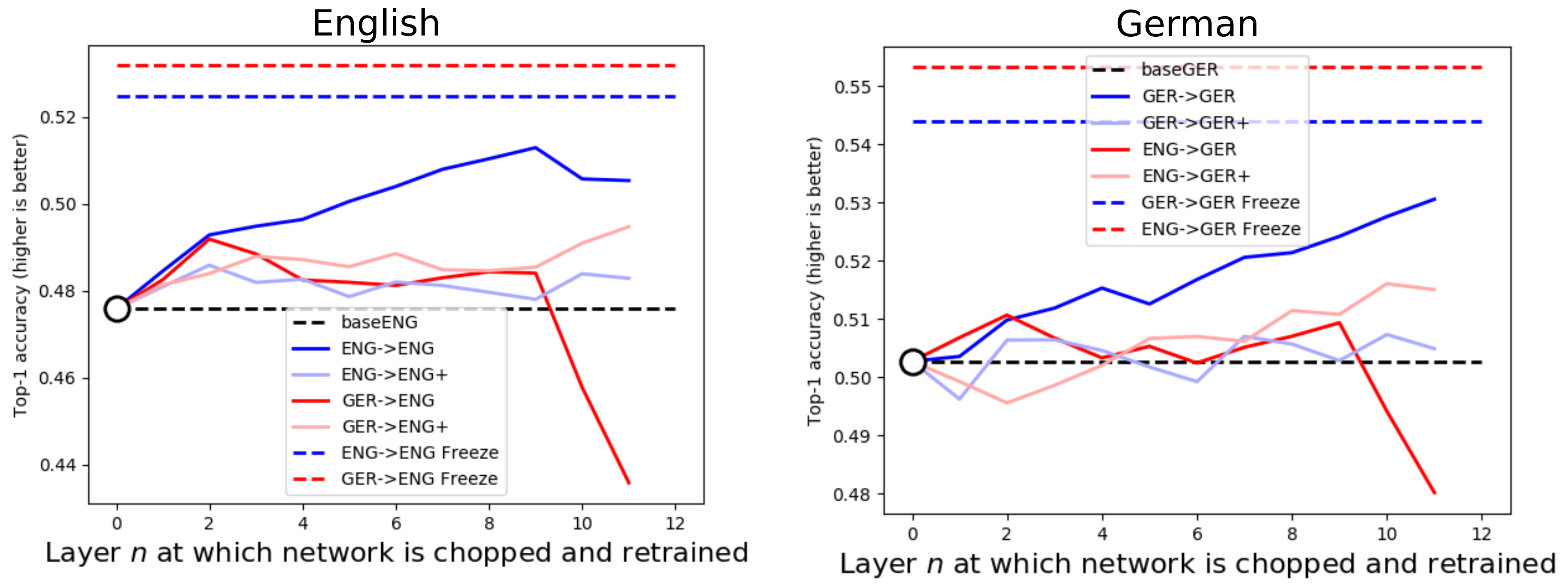
"Selfer" network
(with or without finetuning
on implanted layers)

"Transfer" network
(with or without finetuning
on implanted layers)

Figure copied from Yosinksi et al 2014

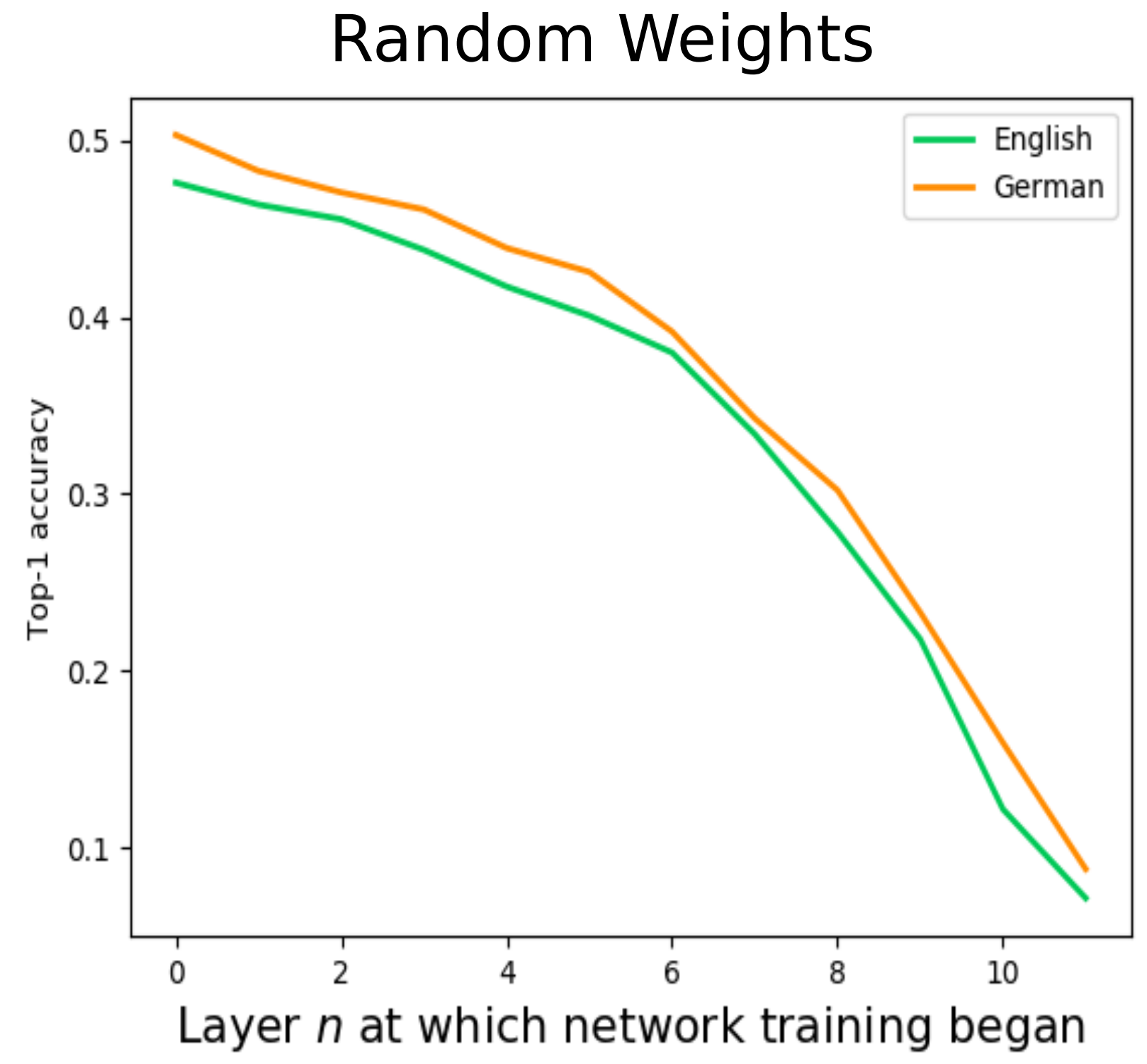
Results

All convolutional layers were easily transfered between languages at no significant loss in performance. Transfer networks that were chopped at fully connected layers (ten and eleven) showed a marked drop in accuracy (up to 4%) compared to the monolingual baseline network. However, those same networks showed a slight gain over the monolingual baseline when the implanted parameters were finetuned on the target language. Surprisingly, the selfer networks without finetuning performed better than all other 'chimera' networks. We hypothesized that this was partly due to the benefit of weight freezing, as in freeze training.


English
German

Freeze Training (Raghu et al. 2017): Parameters were gradually "frozen" over the course of training starting with early layers until only the last layer was being updated for the final epochs. This procedure resulted in the best performance overall, and greatly improved the performance of the transfer networks with fewer weight updates.

Random, Untrained Weights: Performance gradually degrades as fewer parameters are trained and more layers fixed at their random initialization. Yosinski et al. found instead a sharp cutoff at layer layer 3.


Random Weights

Discussion

- Despite a large degree of transferability of intermediate acoustic features between languages, naive approaches to transfer are not the most efficient.
- When studying transfer networks in an underfittimg regime, it is all the more important to compare to selfer control conditions.
- While early layers can be transferred across languages with no loss in performance compared to baseline, there is still some task specificity in later layers such that weight freezing only benefits the selfer networks.
- Consistent with the observation that CNNs converge bottom-up, freeze training, as proposed by Raghu et al (2017), greatly improved performance of both selfer and transfer networks. However, the improvment was greatest for the transfer networks which performed best overall when trained with freeze training.
- Further work needed to explain performance of random, untrained weights.

References

[1] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In Advances in neural information processing systems 27.
[2] Long, M., Cao, Y., Wang, J., & Jordan, M. I. (2015). Learning Transferable Features with Deep Adaptation Networks. In International conference on machine learning (Vol. 37).
[3] Veselý, K., Karafiát, M., Grézl, F., Janda, M., & Egorova, E. (2012). The language-independent bottleneck features. In IEEE workshop on spoken language technology
[4] Raghu, M., Gilmer, J., Yosinski, J., & Sohl-Dickstein, J. (2017). SVCCA: Singular Vector Canonical Correlation Analysis for Deep Understanding and Improvement. In Advances in neural information processing systems 30