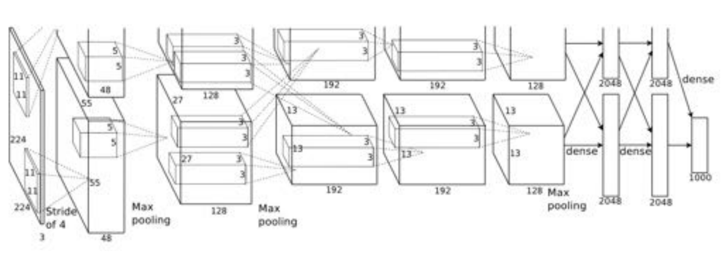
Correspondence between the representations of convolutional neural networks and the activities in inferior temporal cortex measured by electrocorticogaphy
(畳み込みニューラルネットワークの内部表現と下側頭葉における皮質脳波の対応)

伊達 裕人 / Hiroto Date[1,2] 川嵜 圭祐 / Keisuke Kawasaki[2] Mete Ozay[1] 本郷 拓実 / Takumi Hongo[2] 長谷川 功 / Isao Hasegawa[2] 岡谷 貴之 / Takayuki Okatani[1]
1: 東北大院情報科学 / Graduate School of Information Sciences, Tohoku University, Japan
2: 新潟大院医歯学総合研 / Dept of Physiol, Niigata Univ Grad Sch Med Dent Sci, Niigata, Japan

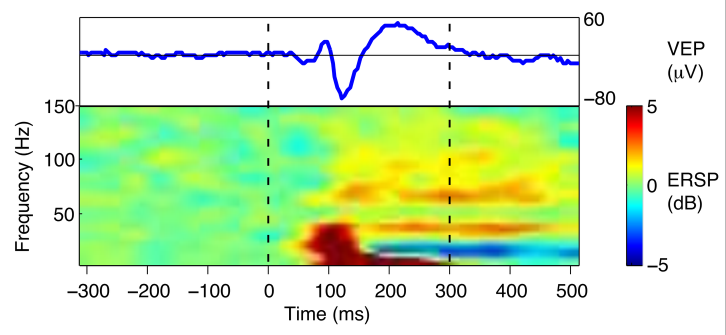
INTRODUCTION

- **Deep convolutional neural networks (CNNs) appear to be the most plausible computational models of visual object recognition in the brain.** CNNs have achieved nearly human-level performance in various computer vision tasks. Moreover, recent studies indicate that internal representations of CNNs are more similar to neural responses than other models of the visual cortex.
- Electrocortocography (ECoG) enables us to record local field potentials (LFPs) with high spatiotemporal resolution. LFPs in various frequency bands may contribute to neural representations at mesoscale, complementary to neuronal firing [2]. In the primate visual cortex, specific frequency bands subserve feedforward or feedback processing. However, it has been unclear **what kind of visual information such frequency-specific activities represent**.

Deep convolutional neural networks (AlexNet [1])



Visual evoked potentials (VEPs) and event-related spectral perturbation (ERSP)



1. **Do predictions of ECoG responses from CNN features have specificity in the frequency domain?**
2. **How are frequency-specific prediction modulated along CNN layers and time?**
3. **What visual properties do the encoding models explain?**

MATERIALS AND METHODS

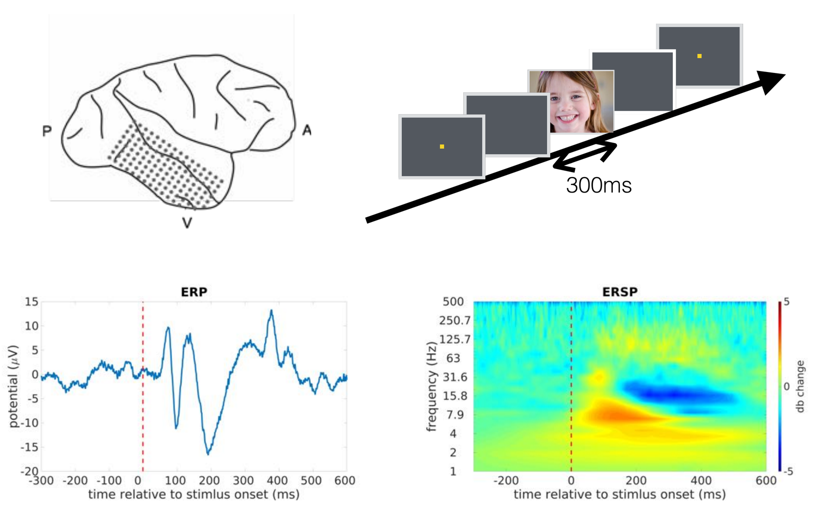
Image set

- Total 12000 natural images (building, body part, face, foliage, fruit, fur, glass, insect, leather, metal, paper, tool)

Recording neural responses in the primate inferior temporal cortex

- We recorded cortical potentials of 128 channel electrocorticography (ECoG) covering from macaque posterior ITC to anterior ITC.
- We computed the amplitude of each frequency (1-500 Hz) by complex Morlet wavelet convolution.
- We downsampled the amplitude for each time window (20 ms), and then conducted trial averaging.

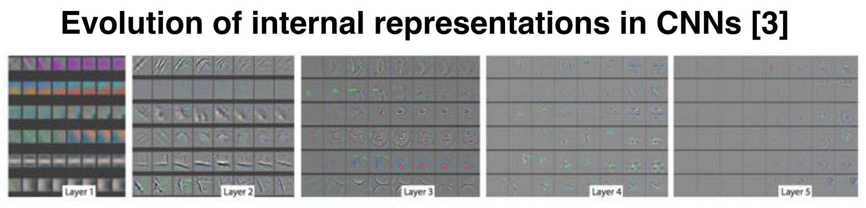
Frequency specific responses recorded by ECoG



Diverse image features from deep convolutional neural networks

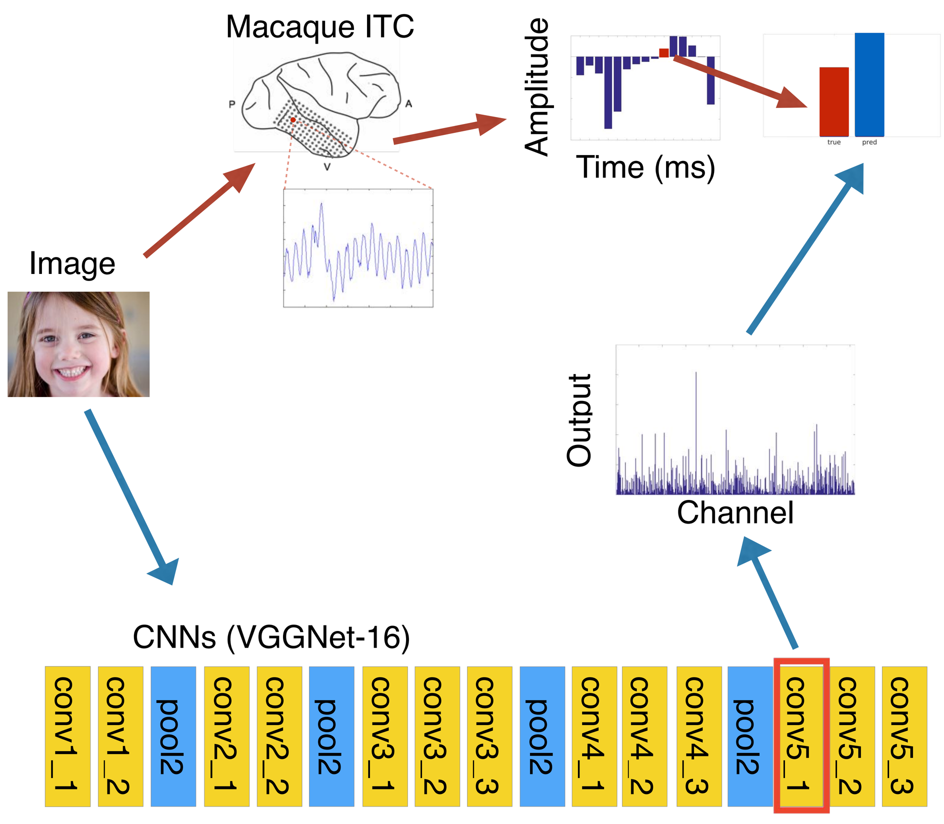
- Deep convolutional neural networks (CNNs) have achieved nearly human-level performance in various computer vision tasks.
- Higher layers in CNNs have higher-level, more abstract and spatially invariant representations [3].
- We used a pretrained model of VGGNet-16 [4], which has 13 convolution layers.
- We extracted outputs at each convolution layer using the same image set.

Evolution of internal representations in CNNs [3]



Encoding frequency-band specific responses from image features

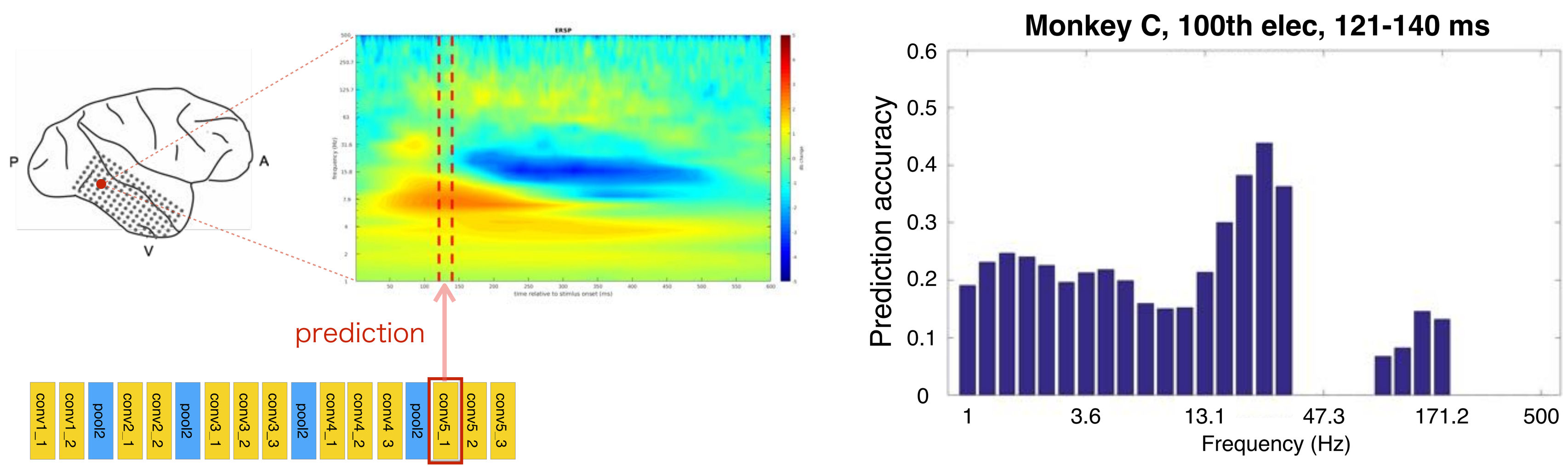
- Encoding ECoG features from CNN features by ridge regression (regularized linear regression)
- An encoding model is specified by one ECoG electrode, time window, frequency, and CNN layer.
- We first optimized each model with training set, and then evaluated each model's prediction accuracy with test set.
- Each model's prediction accuracy was evaluated as Pearson correlation between predicted and true responses.



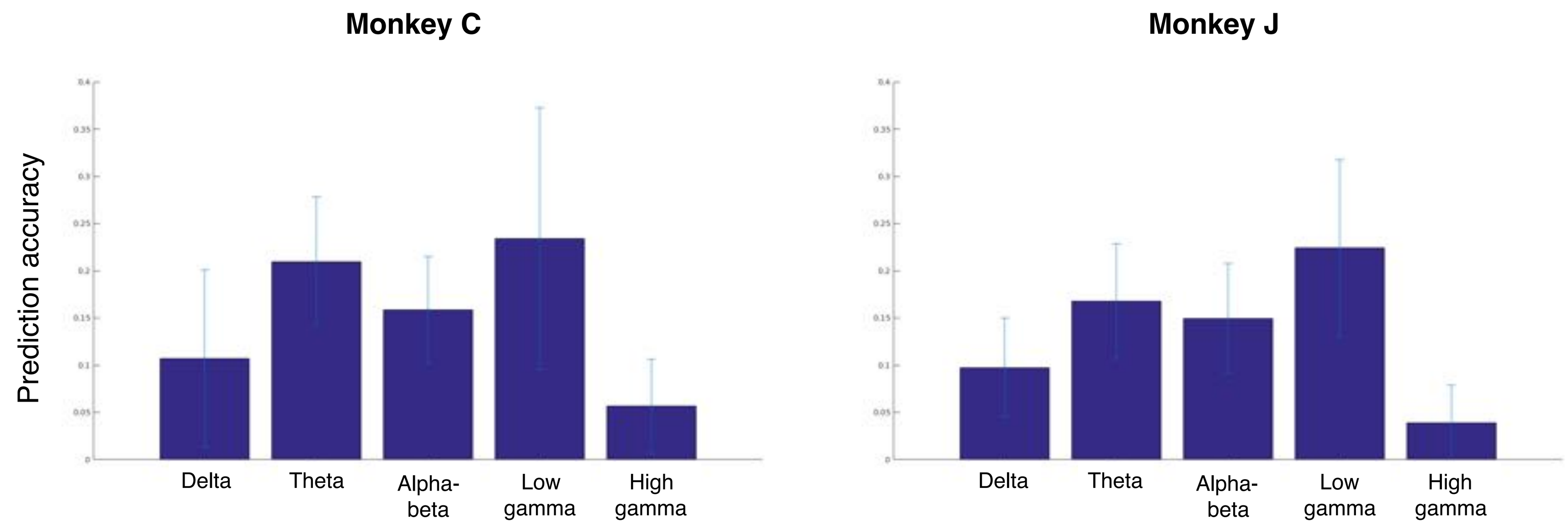
RESULTS

Specificity of prediction accuracy in the frequency domain

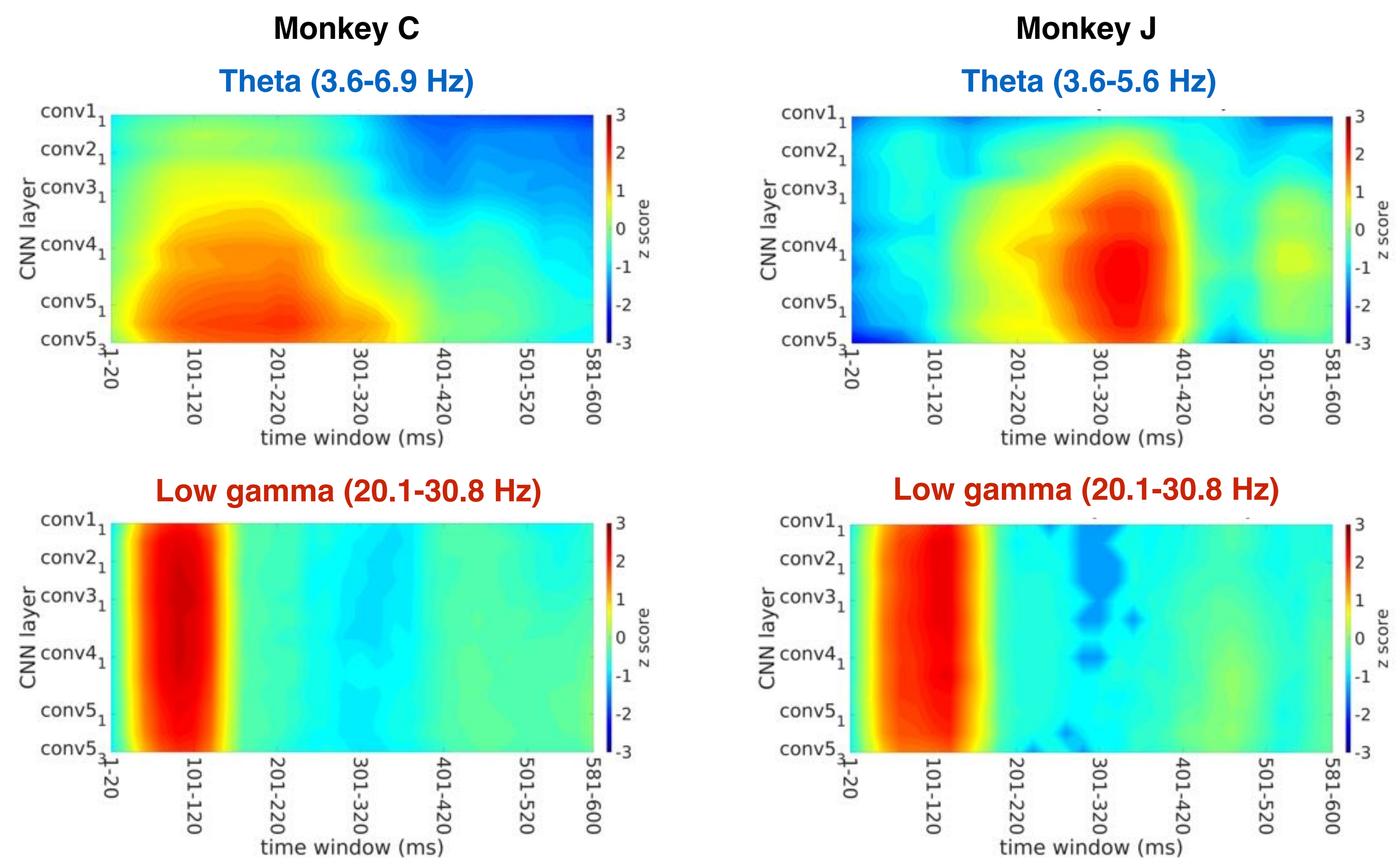
An example: prediction from conv5_1 layer to one electrode and time window



Comparison between each frequency

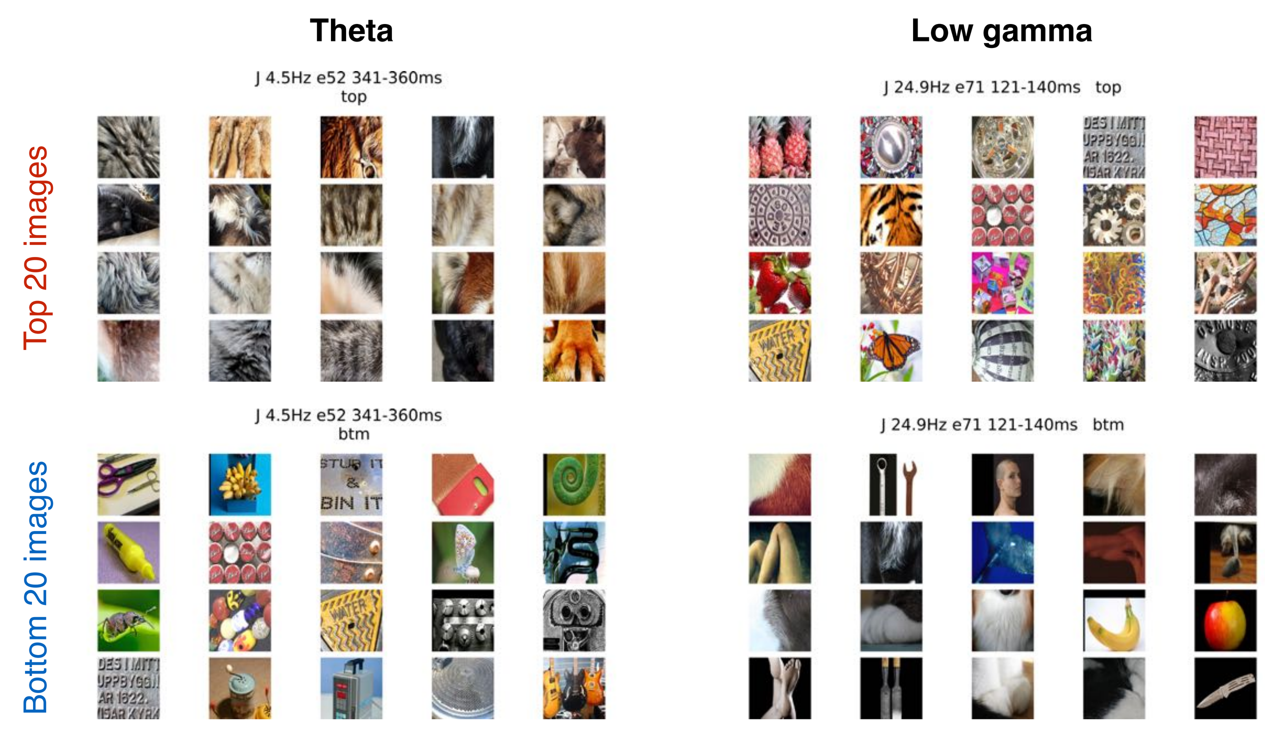


Layer dependence and temporal modulation of prediction accuracy



Visual representations of each frequency-band estimated by encoding models

Predictions for all test images



SUMMARY

- Neural responses in the primate ITC measured by ECoG were predicted by CNN features in a frequency-specific manner.
- Lower-frequency (theta) activities were better predicted by CNN features from middle or higher layers, whereas higher-frequency (low gamma) activities were predicted equally well from almost all the layers.
- Lower-frequency activities were most well predicted at 300-400ms after stimulus onset, whereas higher-frequency activities were at 50-150ms after stimulus onset.
- Visual representations estimated by the best encoding model of each frequency band indicated frequency-specific representations of visual attributes.

References
[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems, 1–9.
[2] Watrous, A. J., Fell, J., Ekstrom, A. D., & Axmacher, N. (2015). More than spikes: common oscillatory mechanisms for content specific neural representations during perception and memory. Current Opinion in Neurobiology, 31, 33–39.
[3] Zeiler, M., & Fergus, R. (2014). Visualizing and understanding convolutional networks. Computer Vision–ECCV 2014, 8689, 818–833.
[4] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. Iclr 2015, 1–14.

Conflict of Interest: we declare no competing financial interest and no conflict of interest.