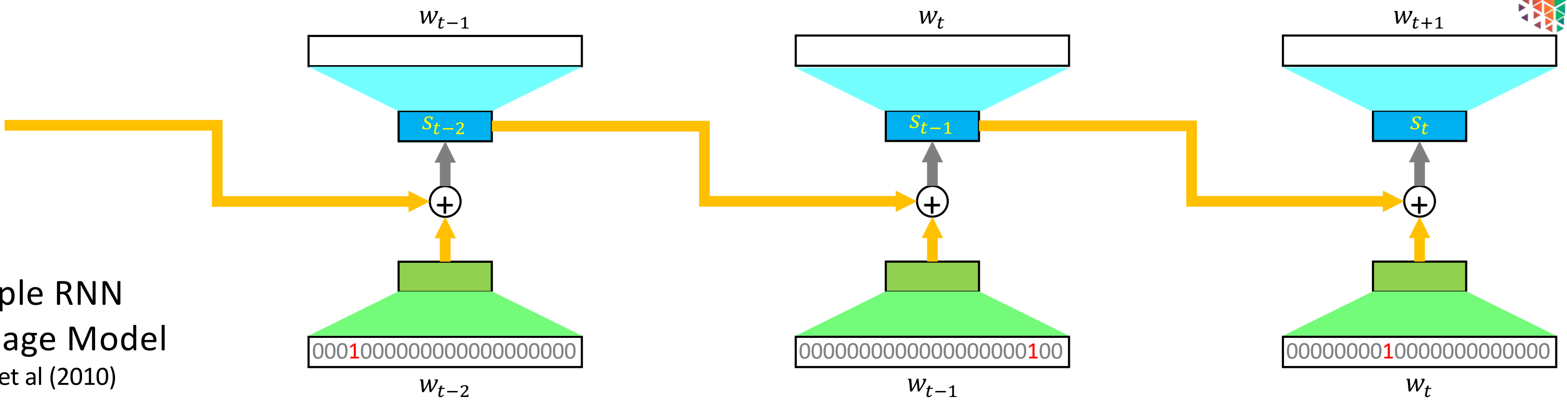
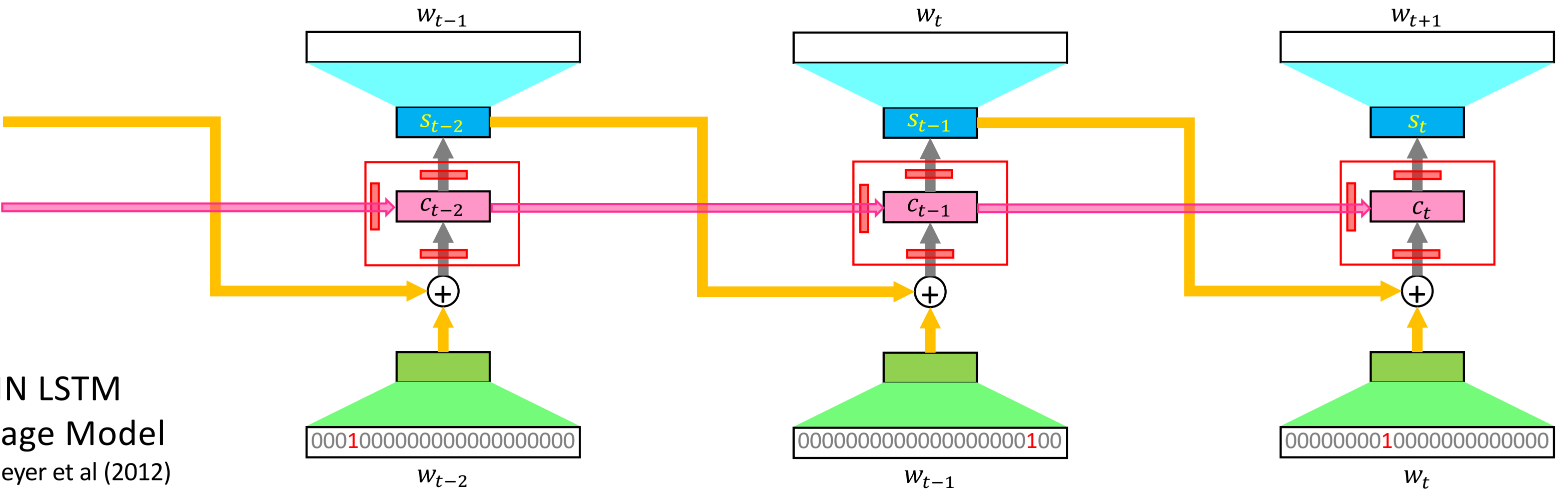
SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

Self-Supervised Representation Learning for Automatic Speech Recognition

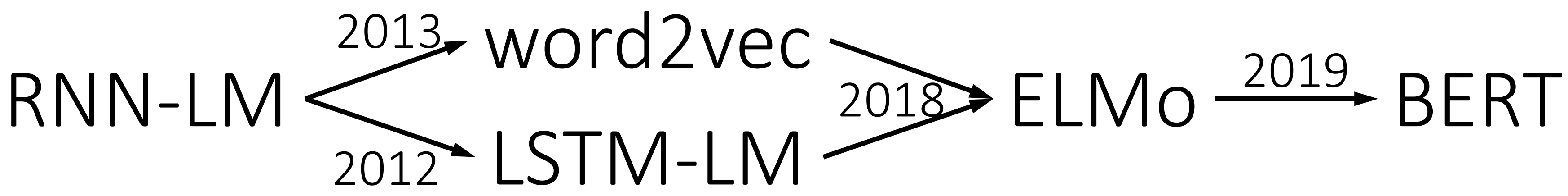
SSL for Speech using Masked Language Model Objective: Hsu et al (2021)
SSL for Speech using Noise Contrastive Objective: Schneider et al (2019), Baevski et al (2020)
Interpreting SSL as Maximum Mutual Information Estimation

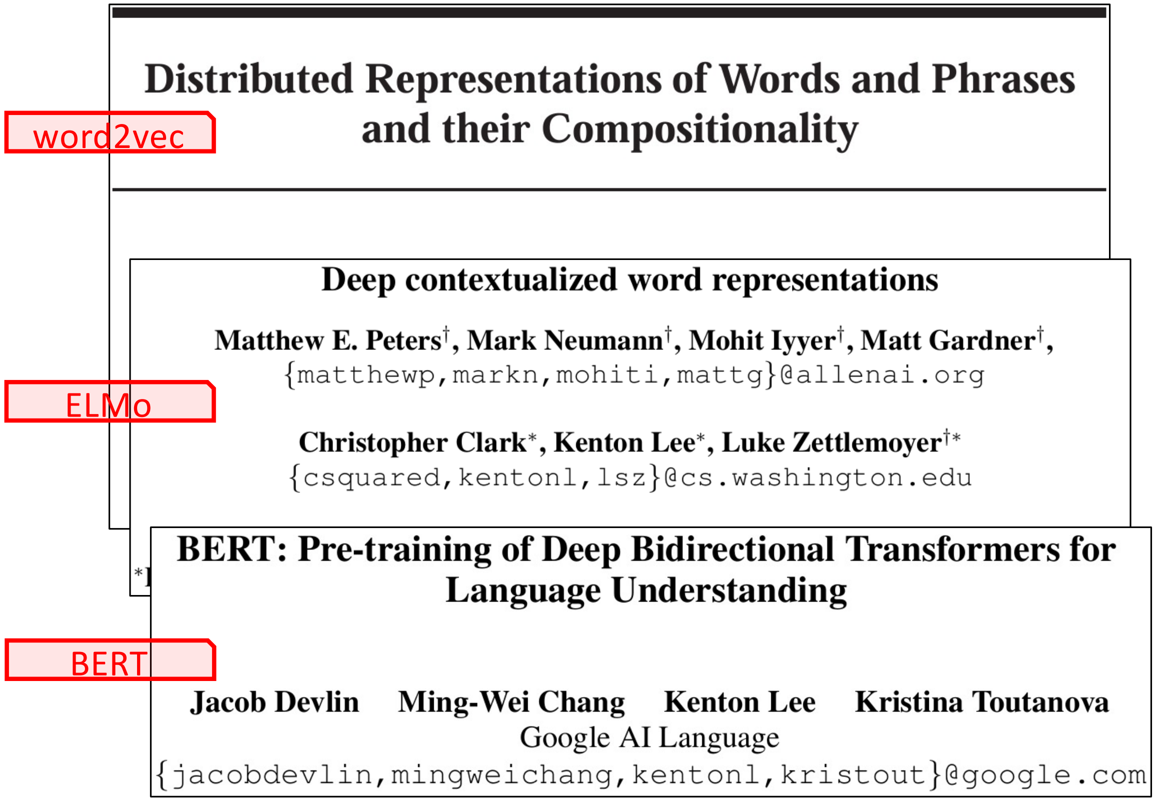


A Simple RNN
Language Model
Mikolov et al (2010)



An RNN LSTM
Language Model
Sundermeyer et al (2012)




word2vec
ELMo
BERT

Distributed Representations of Words and Phrases and their Compositionality

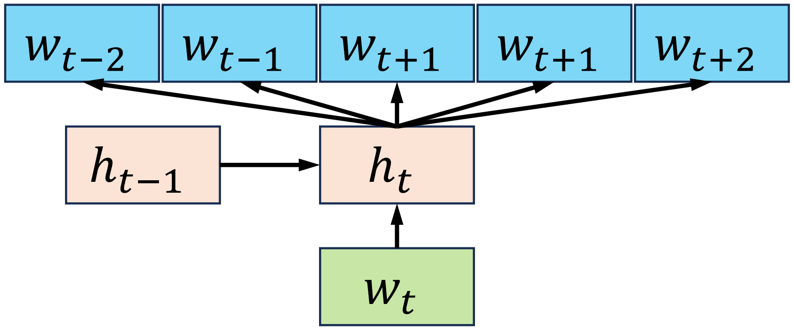
Deep contextualized word representations

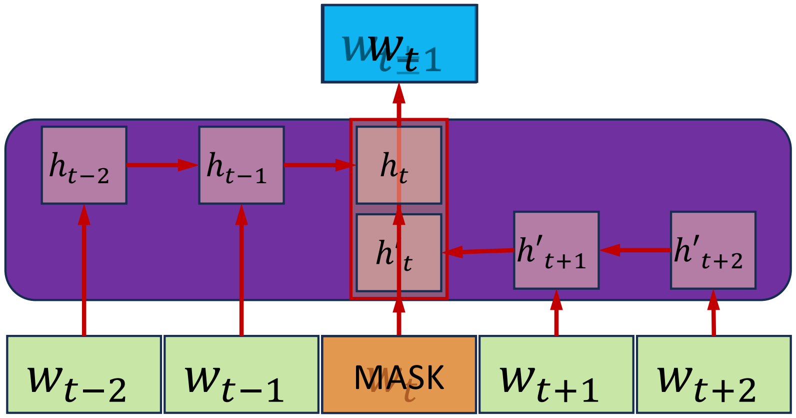
Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
`{matthewp,markn,mohiti,mattg}@allenai.org`

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
`{csquared,kentonl,lsz}@cs.washington.edu`

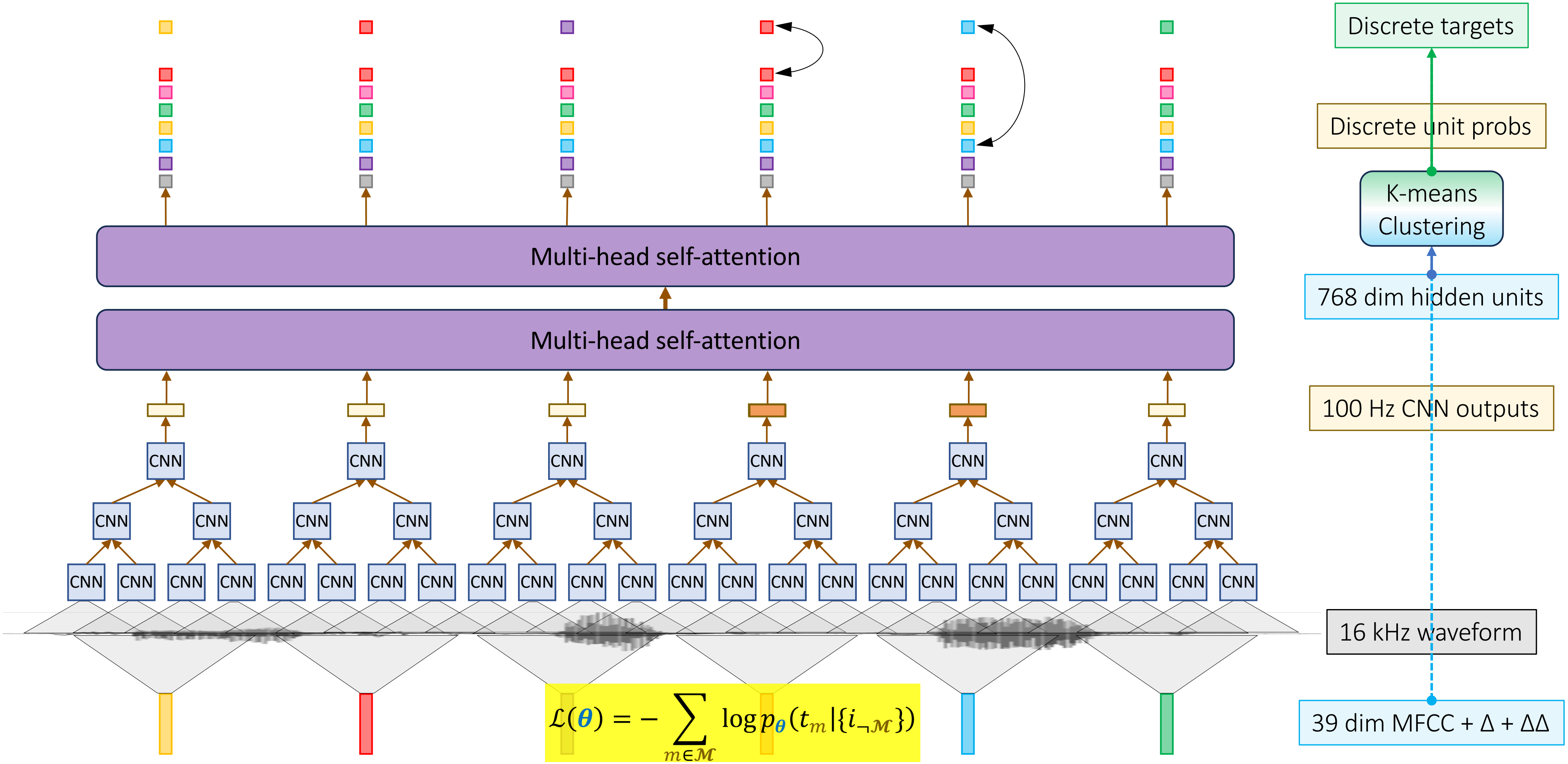
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
`{jacobdevlin,mingweichang,kentonl,kristout}@google.com`





HuBERT – Quantized speech “tokens” and BERT-like loss



$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_{m\in\mathcal{M}} \log p_{\boldsymbol{\theta}}(t_m|\{i_{\neg\mathcal{M}}\})$$

Discrete targets

Discrete unit probs

K-means Clustering

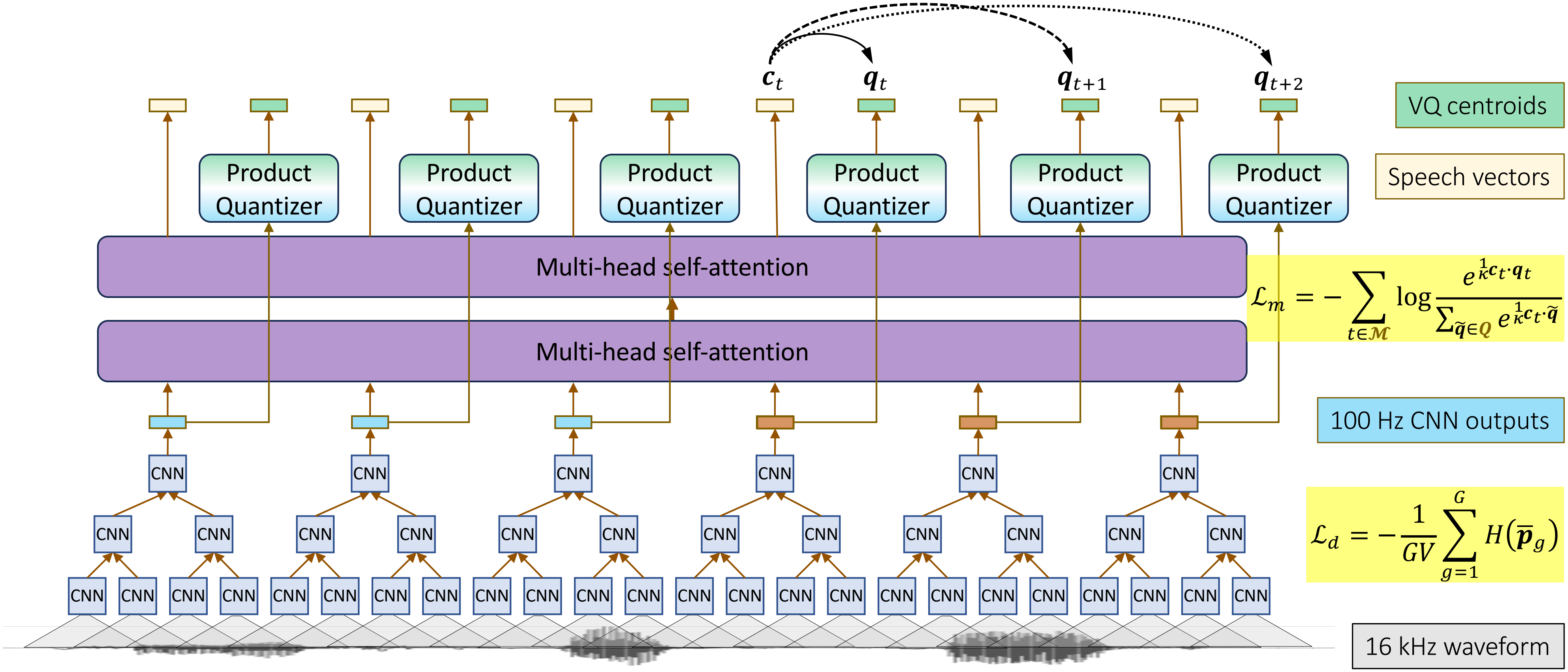
768 dim hidden units

100 Hz CNN outputs

16 kHz waveform

39 dim MFCC + Δ + ΔΔ

wav2vec 2.0 – Noise Contrastive Estimation and Learnt VQ



$$\mathcal{L}_m = -\sum_{t\in\mathcal{M}} \log \frac{e^{\frac{1}{\kappa}\boldsymbol{c}_t\cdot\boldsymbol{q}_t}}{\sum_{\tilde{\boldsymbol{q}}\in Q} e^{\frac{1}{\kappa}\boldsymbol{c}_t\cdot\tilde{\boldsymbol{q}}}}$$

$$\mathcal{L}_d = -\frac{1}{GV}\sum_{g=1}^{G} H(\bar{\boldsymbol{p}}_g)$$

Common misinterpretations of deep representations

Illustrated using a correctly written but often misunderstood paper

LAYER-WISE ANALYSIS OF A SELF-SUPERVISED SPEECH REPRESENTATION MODEL

Ankita Pasad, Ju-Chieh Chou, Karen Livescu

Toyota Technological Institute at Chicago
`{ankitap, jcchou, klivescu}@ttic.edu`

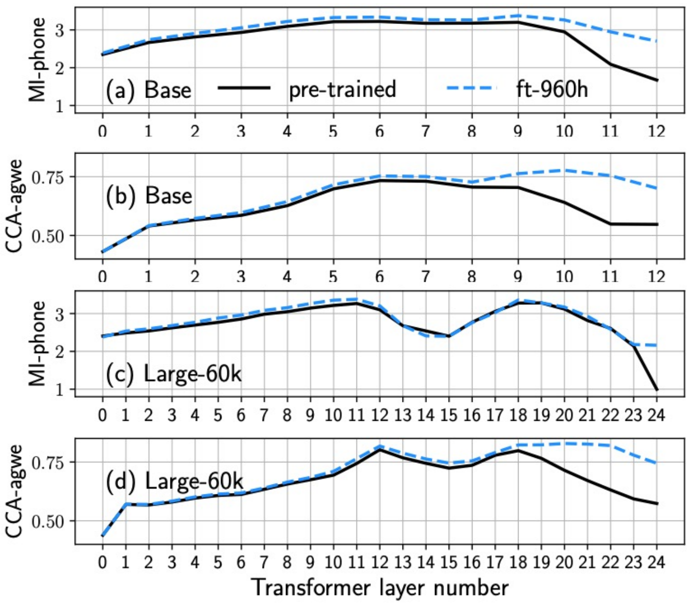


Fig. 5. *MI with phone labels (max: 3.6) and CCA similarity with AGWE.*

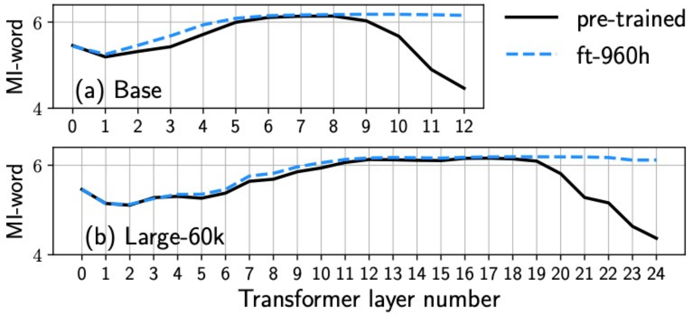


Fig. 6. *MI with word labels (max: 6.2).*

Data processing inequality (Cover & Thomas, pp32)

$$I(A;W) \;\geq\; I(\boldsymbol{f}_k(A);W) \;\geq\; I\big(\boldsymbol{f}_l(\boldsymbol{f}_k(A));W\big)$$

$$I\big(\bar{f}_{l,t_1:t_2}(A);w_i\big) = I\big(\bar{f}_{k,t_1:t_2}(A) + \bar{g}_{k:l,t_1:t_2}(A);w_i\big) \gtrsim I\big(\bar{f}_{k,t_1:t_2}(A);w_i\big)$$