Neural language models

Daniel Hsu

COMS 6998-7 Spring 2025

NLP (Almost) from Scratch

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa; JMLR 12(76):2493−2537, 2011.

Low-level NLP benchmarks circa 2000-2010ish

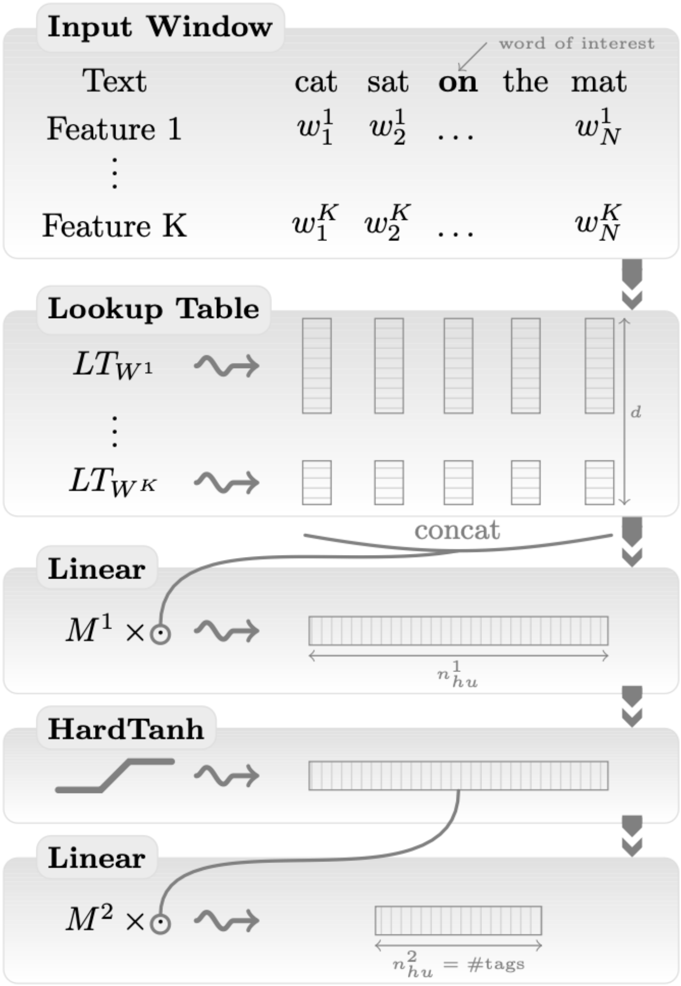
- POS (parts-of-speech tagging)
- Chunking
- NER (named entity recognition)
- SRL (semantic role labeling)

DET	NOUN	VERB	DET	NOUN
the	sailor	dogs	the	hatch

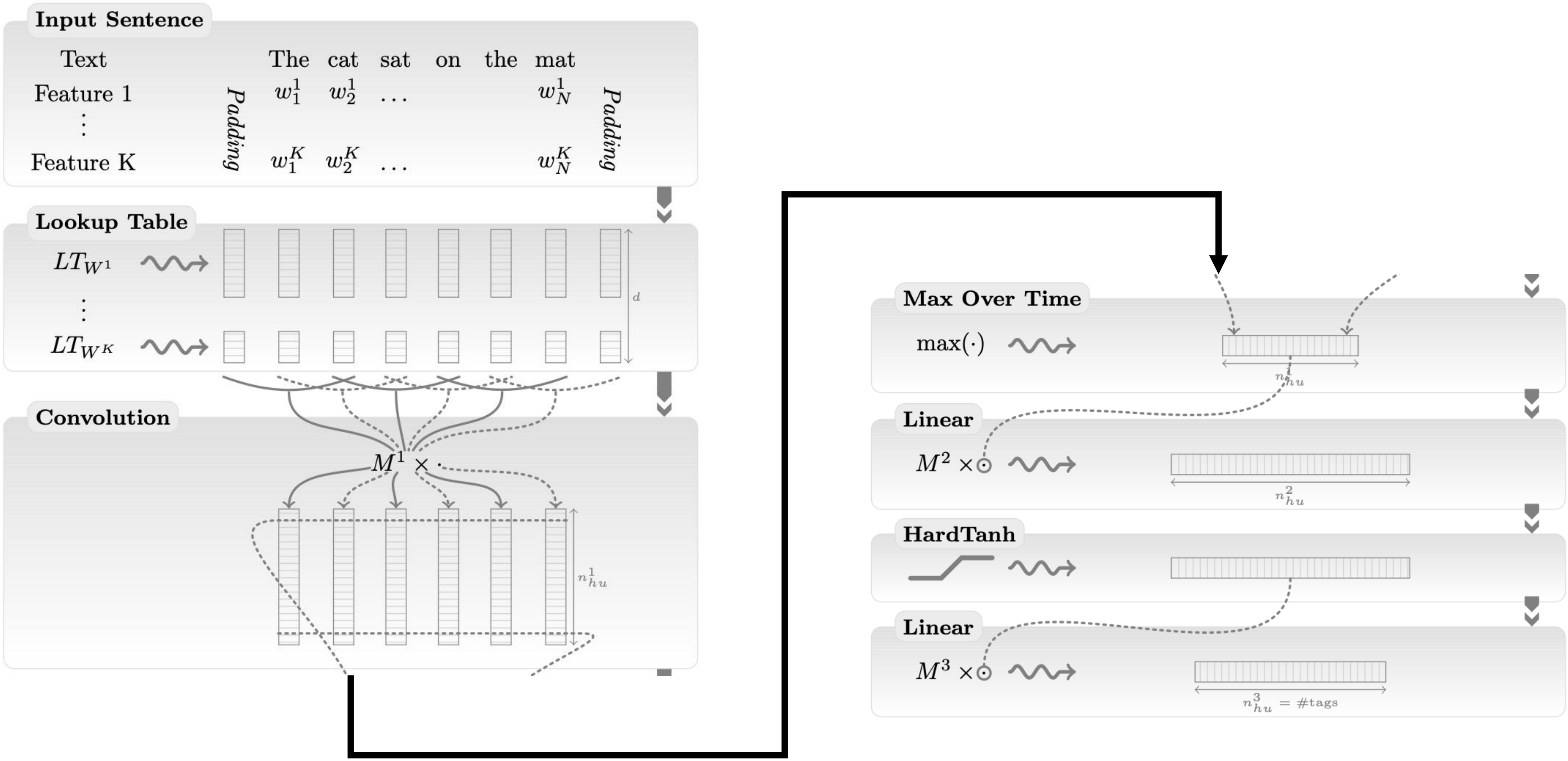
fastens (something) with a dog

- Train tagging models using supervised learning

Architectures: window approach



Architectures: sentence approach



What's hard-coded and what's trainable?

- Hard-coded:
  - Use of discrete word features
  - Extra discrete features
    - Some are generic (e.g., "stemmed" versions of words)
    - Some are task-specific (e.g., in SRL: "distance to target word")
  - Architecture
    - E.g., window size, choice of non-linear activation functions
- Trainable:
  - Embedding vectors ("lookup tables") for discrete features
  - Weight parameters for two linear layers

Supervised learning

- Use standard labeled data sets
- Training objectives:
  - Word-level conditional log-likelihood
  - Sentence-level conditional log-likelihood based on Viterbi-like decoding
- Results:
  - Far from state-of-the-art (circa 2008-2011)

Using unlabeled data

- Unlabeled data: English Wikipedia; Reuters "RCV1" news articles
- Goal: "train language models that compute scores describing the acceptability of a piece of text"
  - Use "window approach" architecture
  - Computational burden in standard training approach:

$$\log \hat{P}(w_t | w_{t-n+1}, \ldots, w_{t-1}) = \log \frac{\exp\left(f_\theta(w_t, w_{t-1}, \ldots, w_{t-n+1})\right)}{\sum_w \exp\left(f_\theta(w, w_{t-1}, \ldots, w_{t-n+1})\right)}$$

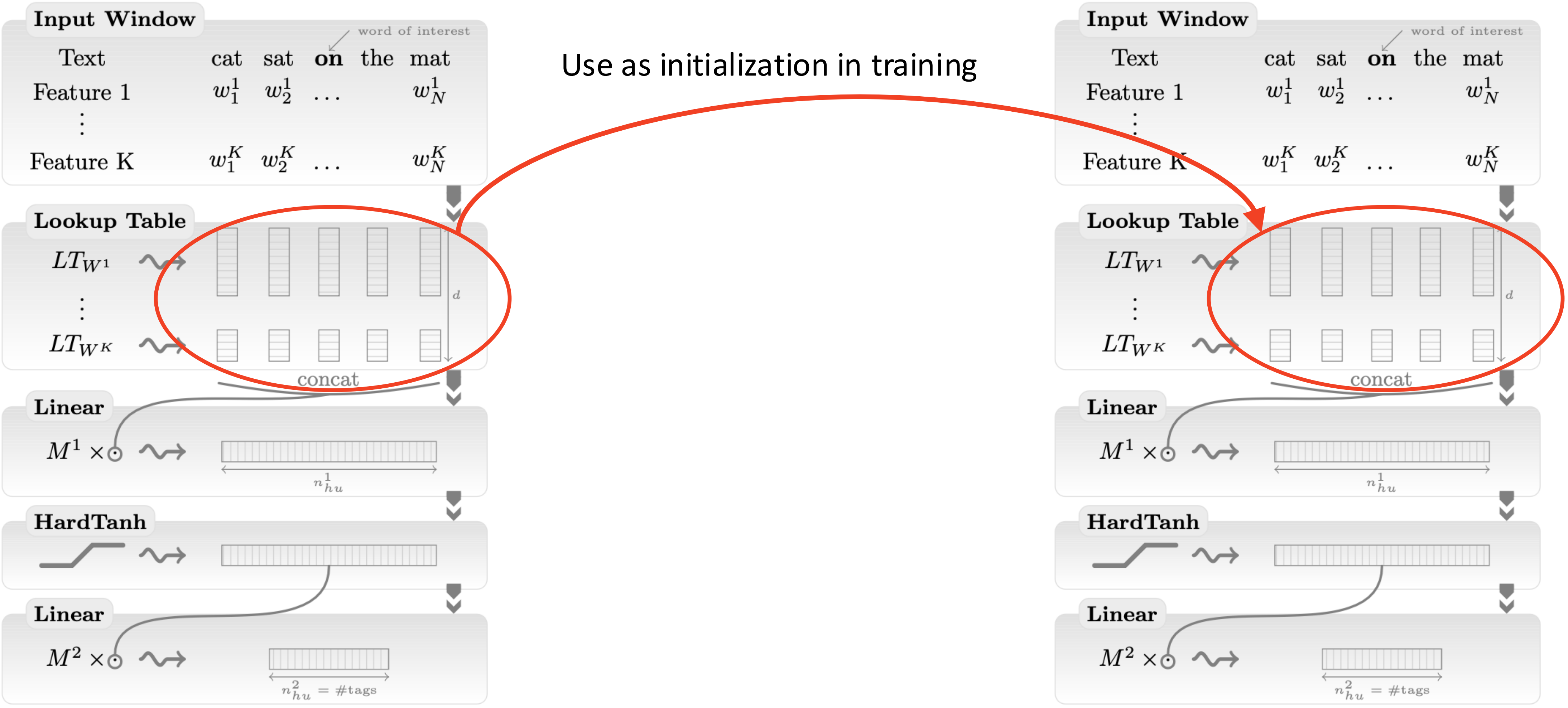
Summation over entire vocabulary!

  - Contrastive ranking alternative: encourage

$$f_\theta(w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}) > f_\theta(w_{t-2}, w_{t-1}, w, w_{t+1}, w_{t+2})$$

for uniformly random w

Semi-supervised learning



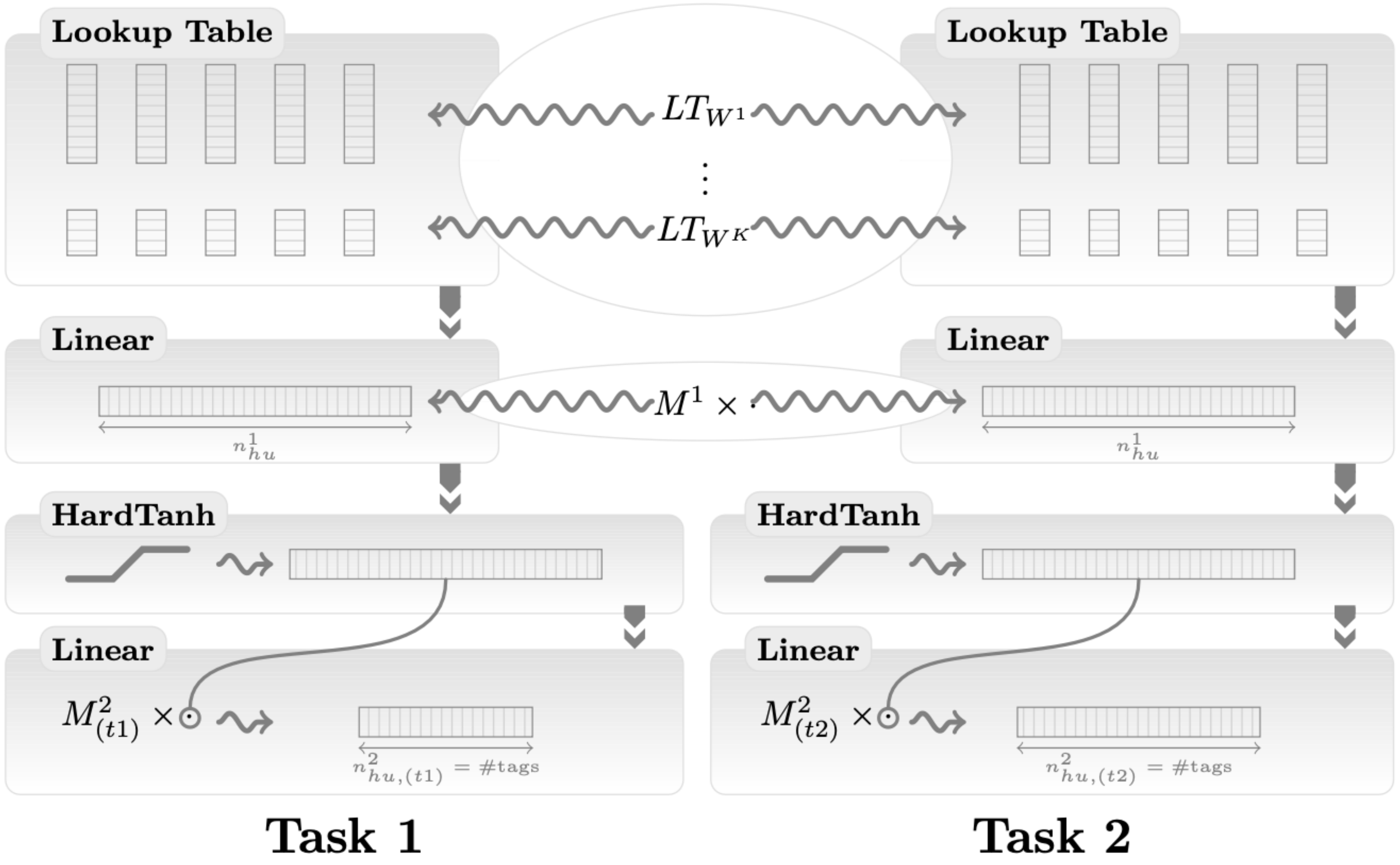
<u>Language model</u> (trained using unlabeled data) <u>POS tagger</u> (trained using labeled data)

Results

Approach	POS (PWA)	CHUNK (F1)	NER (F1)	SRL (F1)
Benchmark Systems	97.24	94.29	89.31	77.92
NN+WLL	96.31	89.13	79.53	55.40
NN+SLL	96.37	90.33	81.47	70.99
NN+WLL+LM1	97.05	91.91	85.68	58.18
NN+SLL+LM1	97.10	93.65	87.58	73.84
NN+WLL+LM2	97.14	92.04	86.96	58.34
NN+SLL+LM2	97.20	93.63	88.67	74.15

LM1 = only using Wikipedia, small vocab
LM2 = Wikipedia + Reuters, big vocab, longer training

Multi-task learning



Theory

Ando and Zhang (JMLR 2005), "A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data"

- <u>Goal of semi-supervised learning</u>:
  - Use unlabeled data to identify "good functional structures" --- i.e., ways to represent data so that target function becomes "smooth" or "simple"
- <u>Why multi-task learning can help</u>:
  - "When we observe multiple predictors for different problems, we have a good sample of the underlying predictor space, which can be analyzed to find the common structures shared by these predictors"
- <u>Method</u>: Train representations using "self-supervised" auxiliary tasks based on unlabeled data; use representations for supervised tasks