Output: Softmax over Vocabulary

Outputs of the RNN are:

1. Projected (scaled up) to the size of the vocabulary V,
2. Normalized with softmax.

$\Rightarrow$ Distribution over all possible target tokens.

$$l(w)_t = W_l h_t + b_l$$

$$p(w)_t = \frac{\exp l(w)_t}{\sum_{w' \in V} \exp l(w')_t}$$

- $l(w)_t$ = logits/energies for word w in time t
- W_l: weight matrix (hidden state × voc. size)
  ... this is **big**.
- Softmax normalization: $\frac{\exp \cdot}{\sum \exp \cdot}$
  ... this is costly.
- Tricks what to do with it
  (negative sampling, hierarchical softmax)
  – not frequently used