Multi-Domain Neural Machine Translation with Word-Level Adaptive Layer-wise Domain Mixing

Haoming Jiang, Chen Liang, Chong Wang and Tuo Zhao

Phan Anh Tu

September 6, 2022

Motivation

Motivation

- ☹ Limit parallel sentences in certain domains when training

Motivation

- ☹ Limit parallel sentences in certain domains when training
- ➥ Multi-domain Neural Machine Translation (NMT) model

Motivation

- ☹ Limit parallel sentences in certain domains when training
- ➥ Multi-domain Neural Machine Translation (NMT) model
  - ☹ Not well exploit the domain-specific knowledge for each individual domain

Motivation

- ☹ Limit parallel sentences in certain domains when training
- ➥ Multi-domain Neural Machine Translation (NMT) model
  - ☹ Not well exploit the domain-specific knowledge for each individual domain
  - ➥ Assign domain related weights to different samples during training
  - ➥ Design specific encoder-decoder architectures for NMT models

Motivation

- ☹ Limit parallel sentences in certain domains when training
- ➥ Multi-domain Neural Machine Translation (NMT) model
  - ☹ Not well exploit the domain-specific knowledge for each individual domain
  - ➥ Assign domain related weights to different samples during training
  - ➥ Design specific encoder-decoder architectures for NMT models
    - ☹ One single encoder to learn shared embedding $\rightarrow$ not well exploit the domain-specific knowledge

Motivation

- ☹ Limit parallel sentences in certain domains when training
- ➥ Multi-domain Neural Machine Translation (NMT) model
  - ☹ Not well exploit the domain-specific knowledge for each individual domain
  - ➥ Assign domain related weights to different samples during training
    - ☺ Complementary to proposed model
  - ➥ Design specific encoder-decoder architectures for NMT models
    - ☹ One single encoder to learn shared embedding $\rightarrow$ not well exploit the domain-specific knowledge
- ➠ A novel multi-domain NMT model using **individual modules** for each domain
  - ☺ Capturing domain-specific knowledge

Contents

Transformer architecture

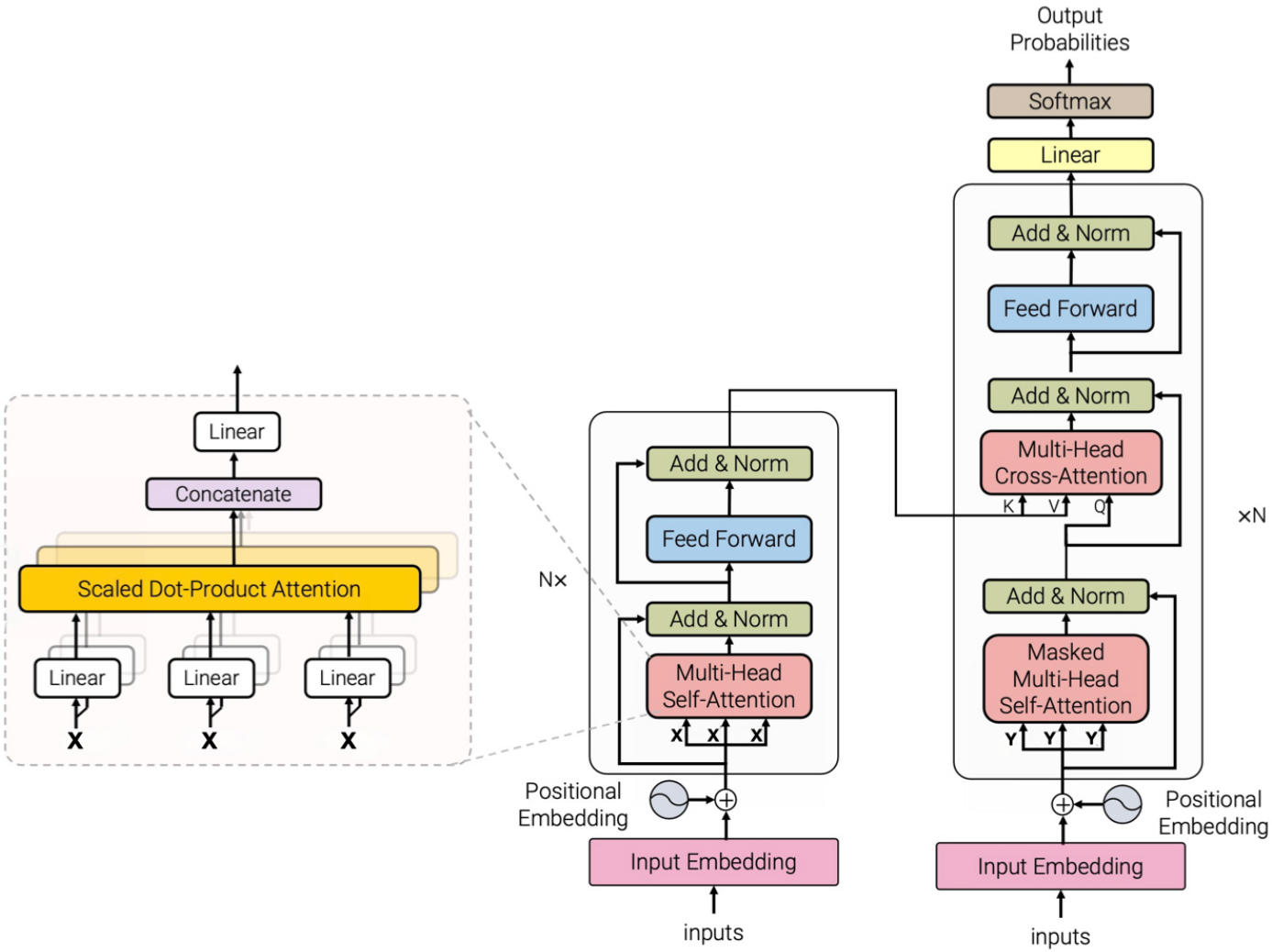


Figure: Transformer Architecture

Self-Attention

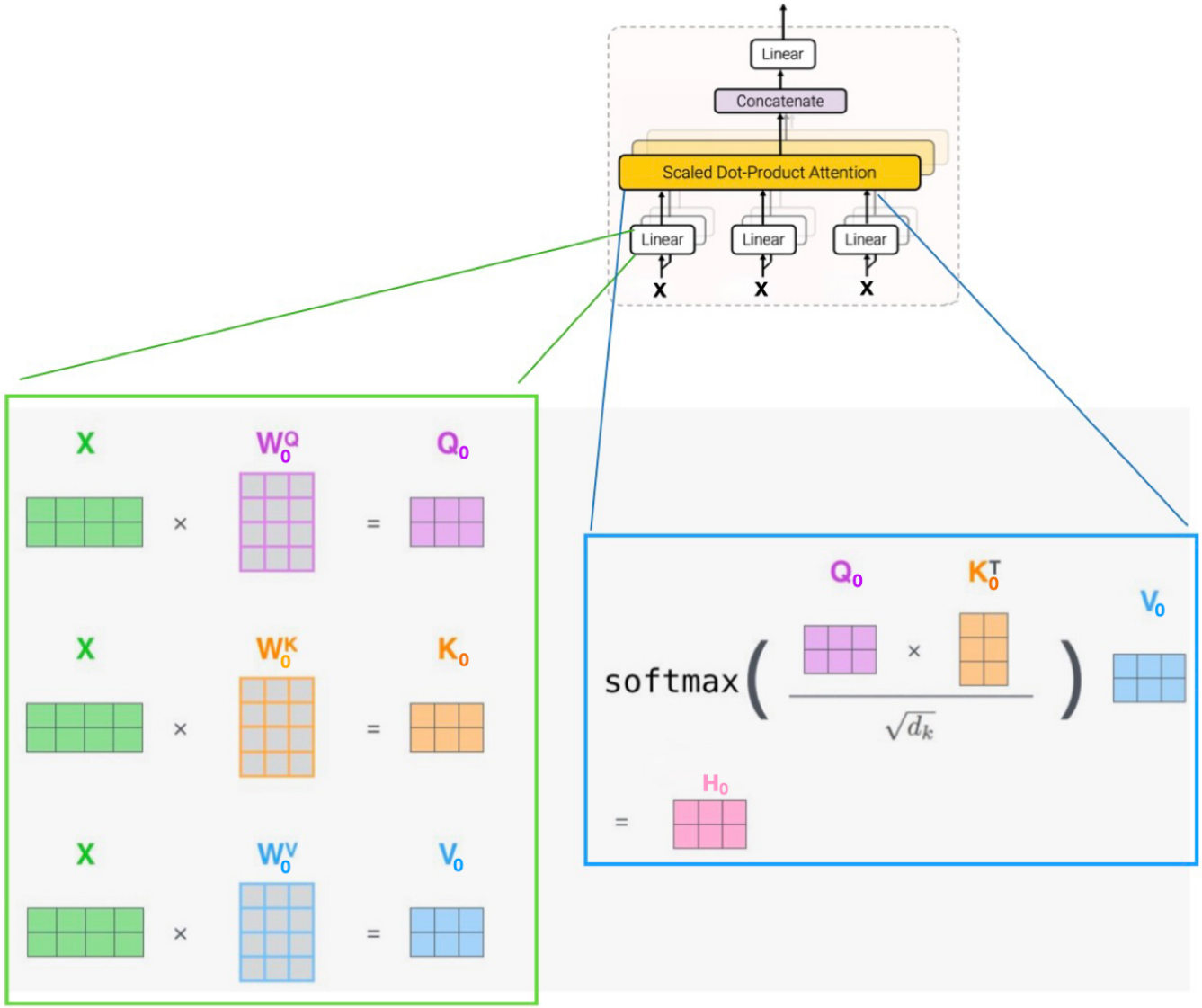


Figure: Matrix view of self-attention

Multi-Head Self-Attention

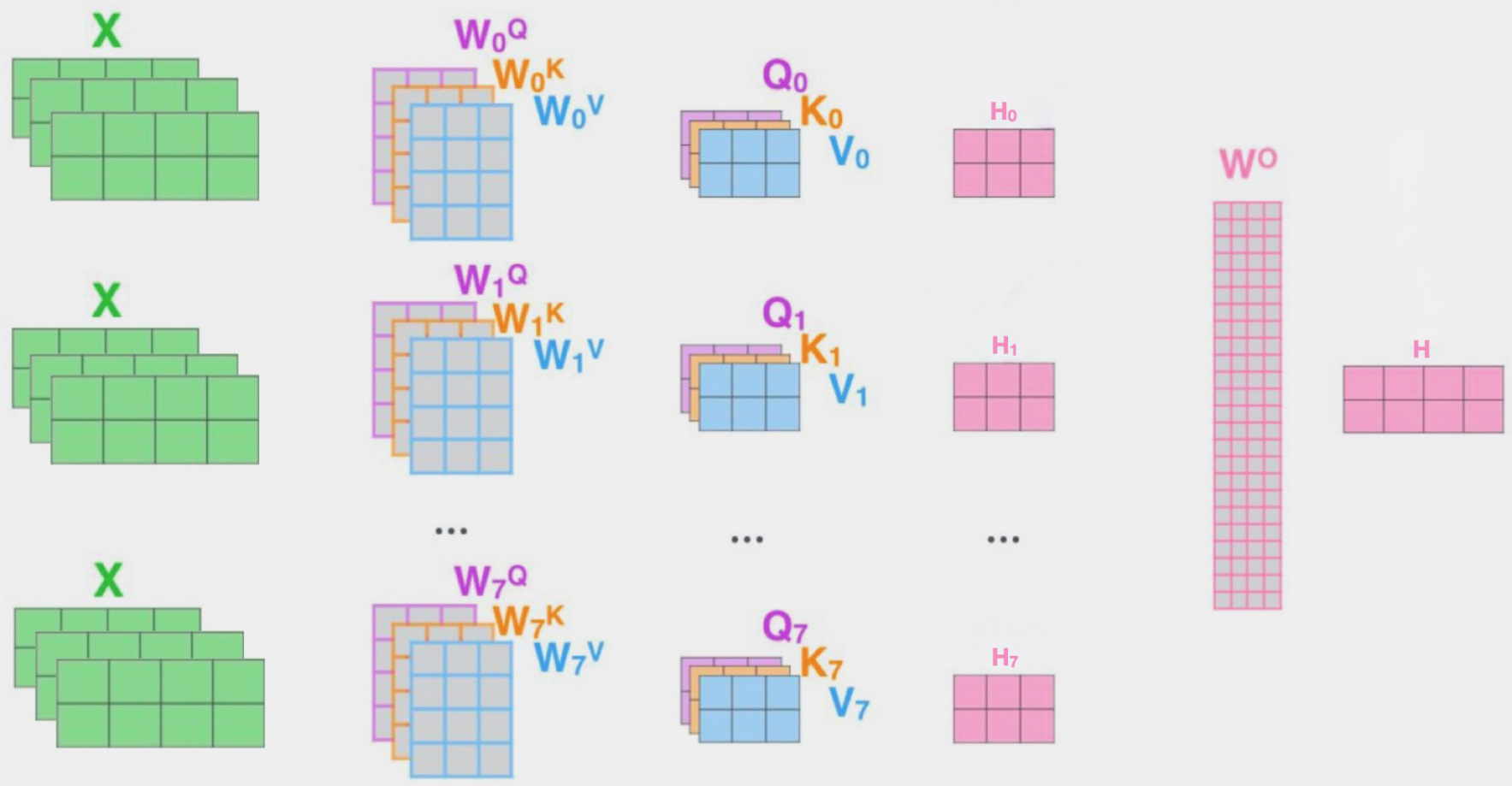


Figure: Matrix view of Multi-Head Self-Attention

Domain Proportion

- Each word in vocabulary has a domain proportion
- Given word embedding $x \in \mathbb{R}^d$, k domains and matrix $\mathcal{R}^{k \times d}$

$$\mathcal{D}(x) = (1 - \epsilon) \cdot \text{softmax}(\mathcal{R}x) + \epsilon/k$$

where $\epsilon \in (0, 1)$

Domain Proportion

- Each word in vocabulary has a domain proportion
- Given word embedding $x \in \mathbb{R}^d$, k domains and matrix $\mathcal{R}^{k \times d}$

$$\mathcal{D}(x) = (1 - \epsilon) \cdot \text{softmax}(\mathcal{R}x) + \epsilon / k$$

where $\epsilon \in (0, 1)$

Larger ϵ encourage the word to be shared across domain

Domain Proportion

- Each word in vocabulary has a domain proportion
- Given word embedding $x \in \mathbb{R}^d$, k domains and matrix $\mathcal{R}^{k \times d}$

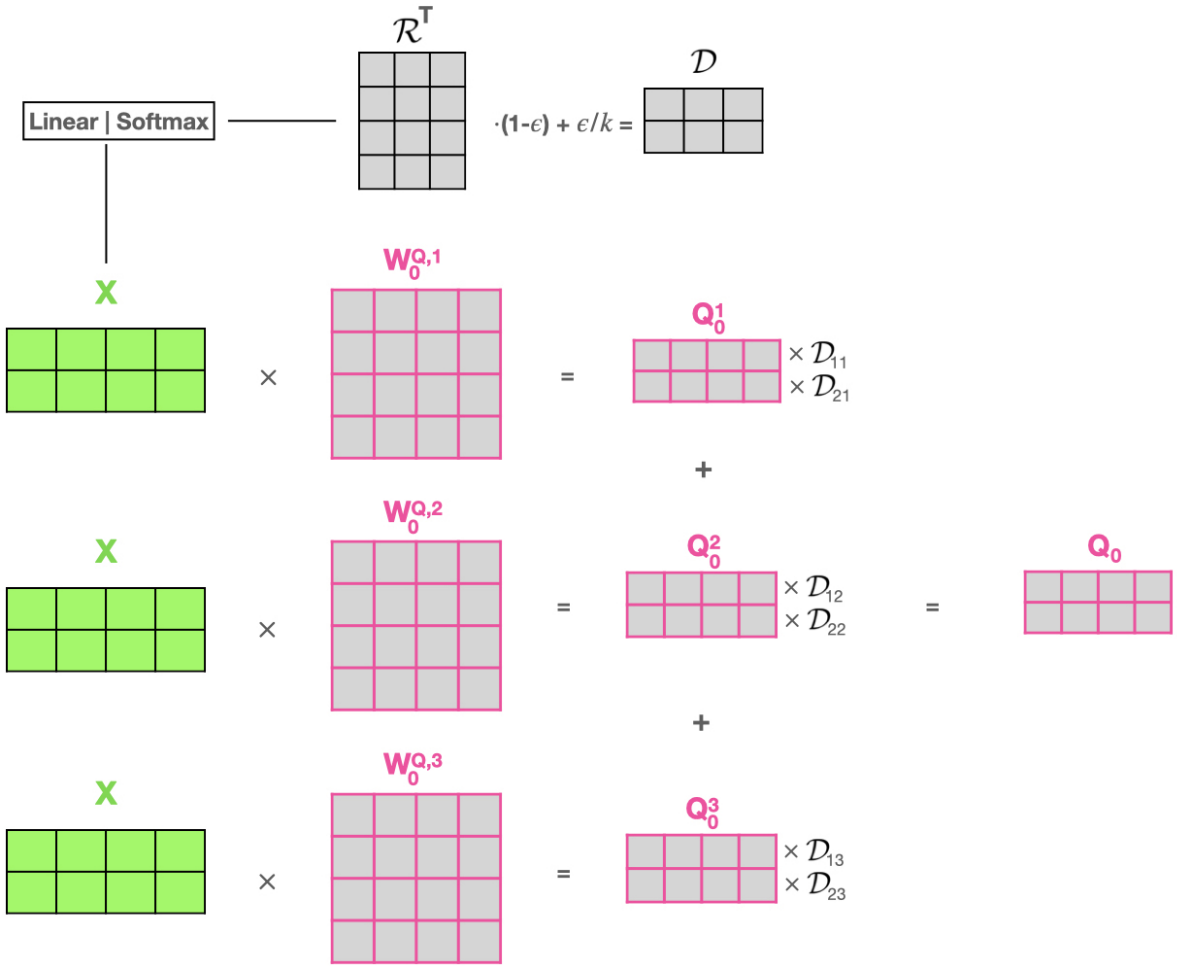
$$\mathcal{D}(x) = (1 - \epsilon) \cdot \text{softmax}(\mathcal{R}x) + \epsilon/k$$

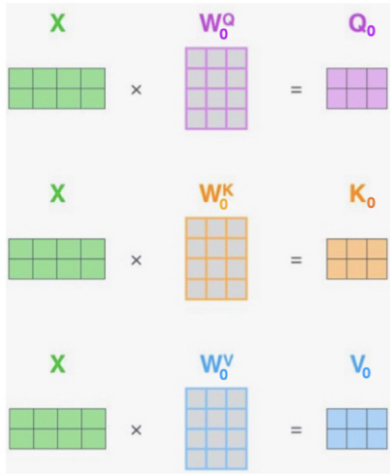
where $\epsilon \in (0, 1)$

Larger ϵ encourage the word to be shared across domain

- $\epsilon = 1 \rightarrow \mathcal{D}(x) = [1/k, 1/k, ..1/k]$

Word-Level Adaptive Domain Mixing





Linear transformation in Transformer

Figure: Matrix view of Word-Level Adaptive Domain Mixing with 3 domains

Layer-wise Domain Mixing

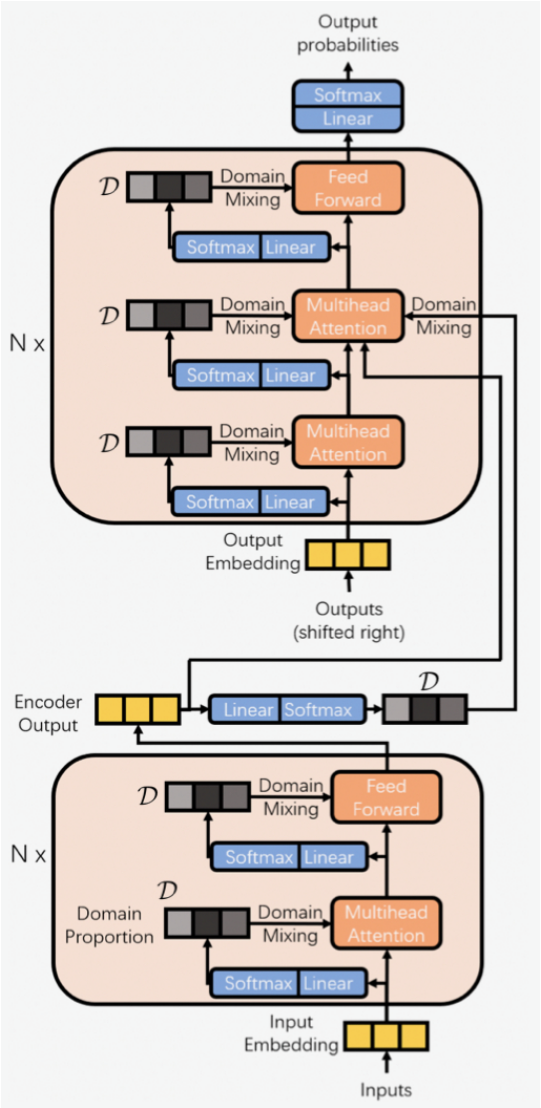


Figure: Multi-domain NMT Model

Training (loss function)

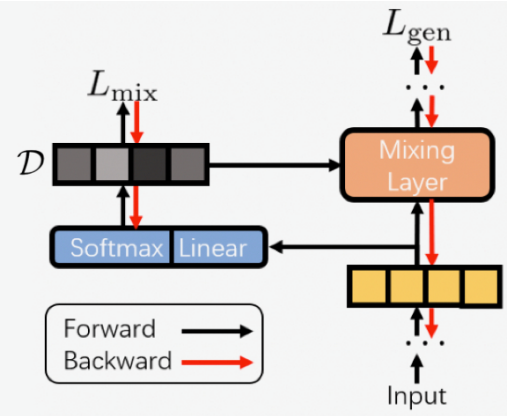
Loss function

$$L^* = L_{gen}(\Theta) + L_{mix}(\Theta)$$

- $L_{gen}(\Theta)$: cross-entropy loss over training data $\{x_i, y_i\}_{i=1}^{n}$
- $L_{mix}(\Theta)$: cross-entropy loss over words and domain labels
  - Each word in sentence x_i has a domain label J (J-th domain)
  - Given embedding x_{ik} of a word k in sentence x_i, cross entropy loss of its is $-\log(\mathcal{D}_J(x_{ik}))$
  - $L_{mix}(\Theta)$ is the sum of the cross entropy loss over all such pair of word and domain label in sentence x_i

Training (parameters update)

- Domain proportion layers $\mathcal{D}$ is trained solely
  $\rightarrow$ Not intervene in the training of the translation model



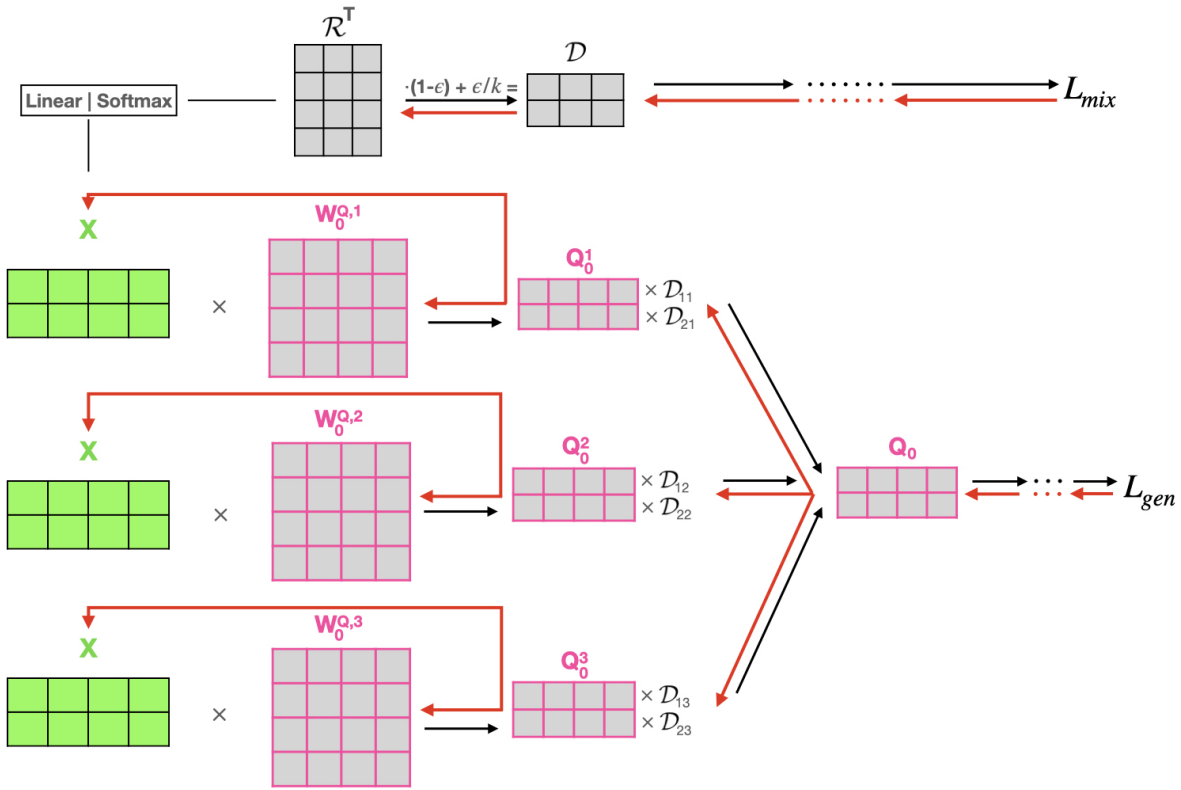


Figure: Computational graph for training the domain proportion layers

Domain Proportions visualization

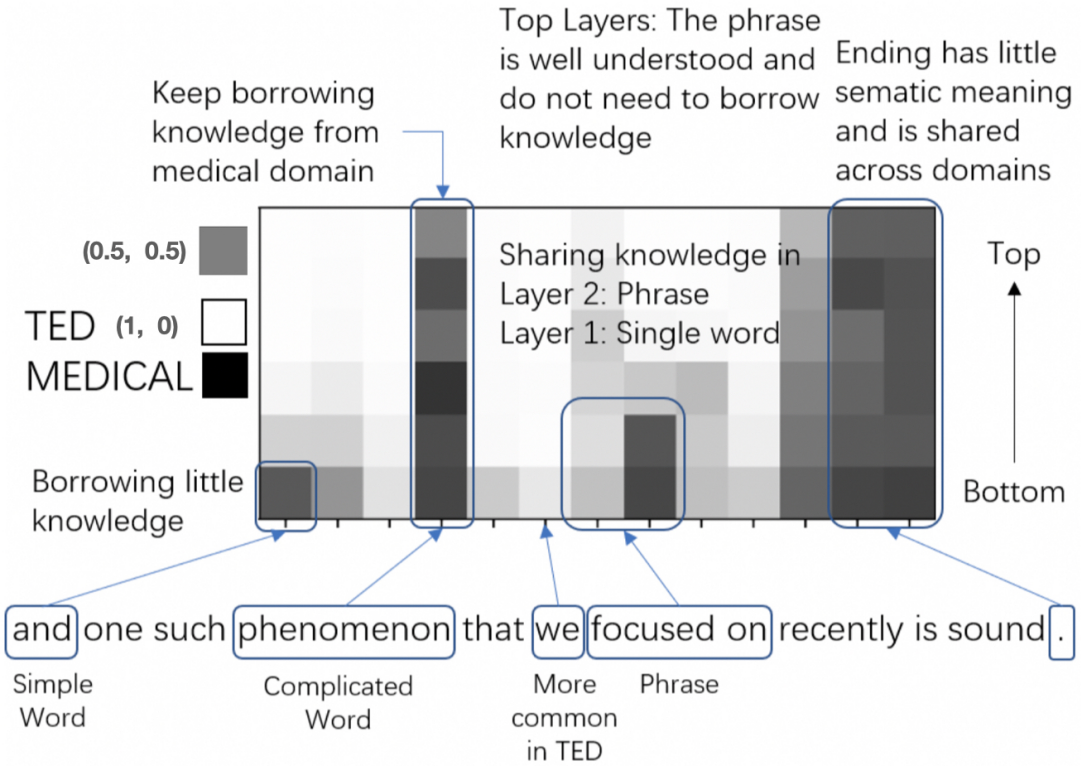


Figure: Domain proportion from TED domain for English-to-French

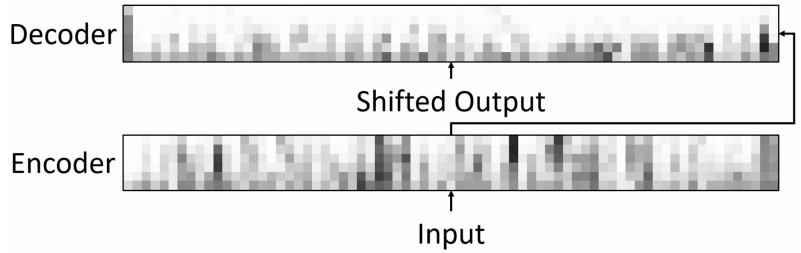


Figure: Domain proportions for English-to-German task of encoder and decoder

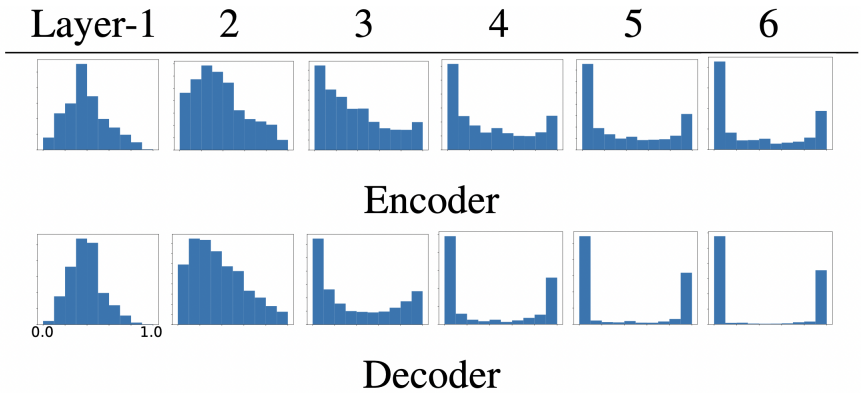


Figure: Histograms of the domain proportions of each layer in English-to-German task

Dataset

Domain	Train	Valid	Test
News	184K	18K	19K
TED	160K	7K	7K

Table: The number of sentences in the dataset

Test Result - Compare with original paper

Method	News	TED
direct-training	19.0 (26.06)	25.0 (28.11)
E/DC	1.0 (27.58)	1.0 (30.33)
E/DC-with-init	1.0	1.0
Encoder	15.0 (27.78)	25.0 (30.30)

Table: BLEU score in English-to-German. The number in bracket bracket is the result in original paper.

Training Experiments

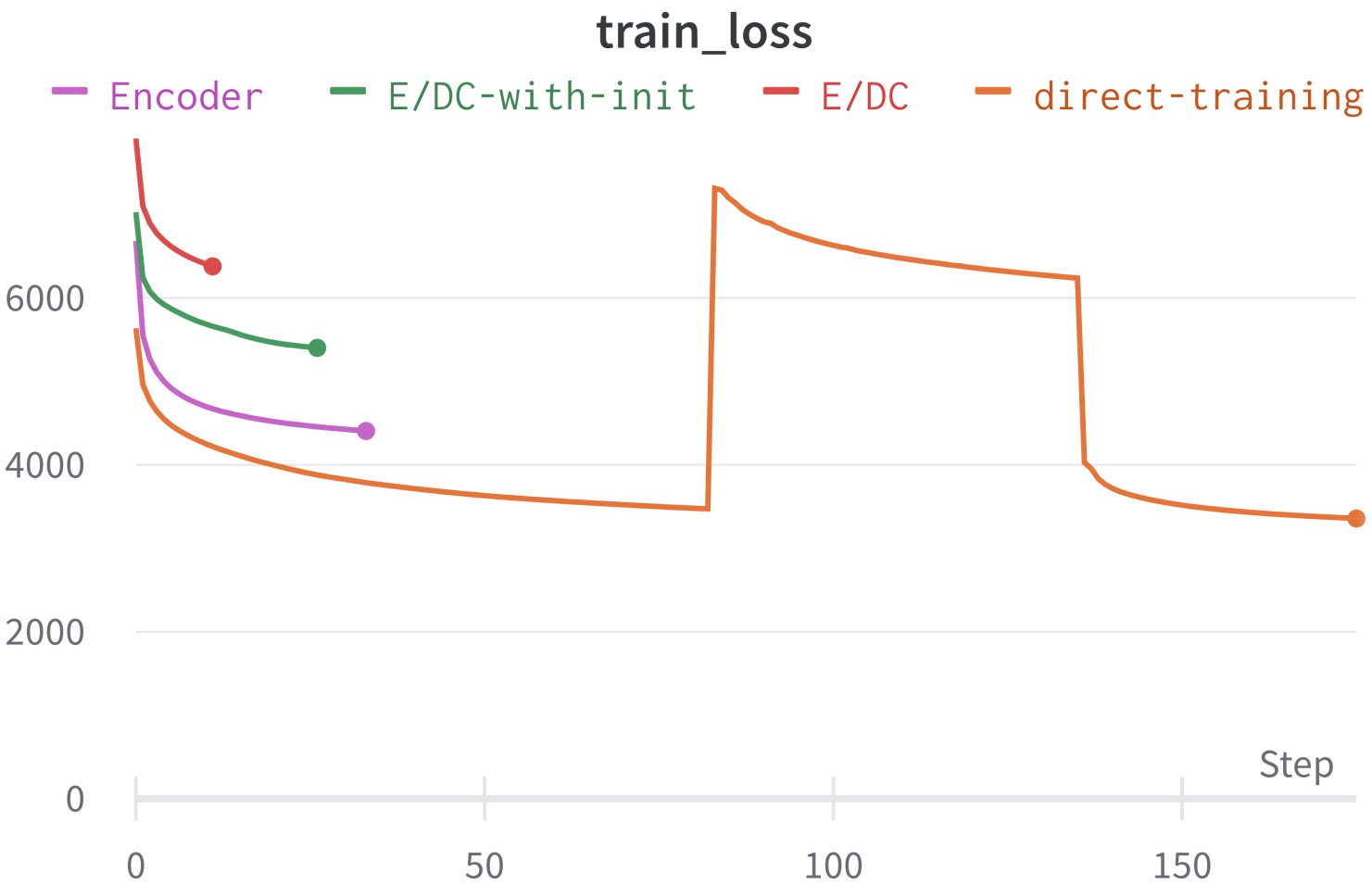


Figure: Training loss

Training Experiments (2)

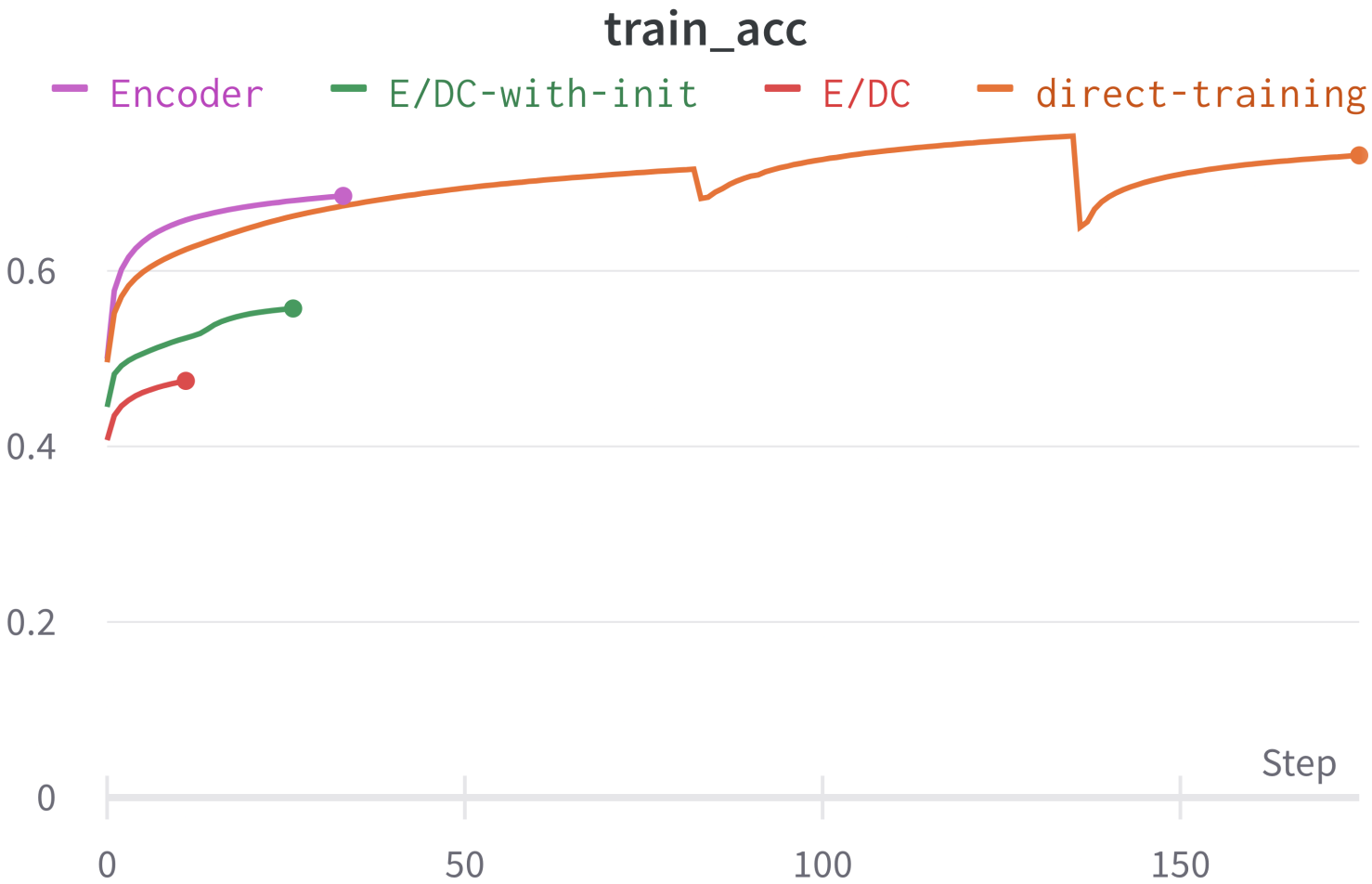


Figure: Training accuracy

Training Experiments (3)

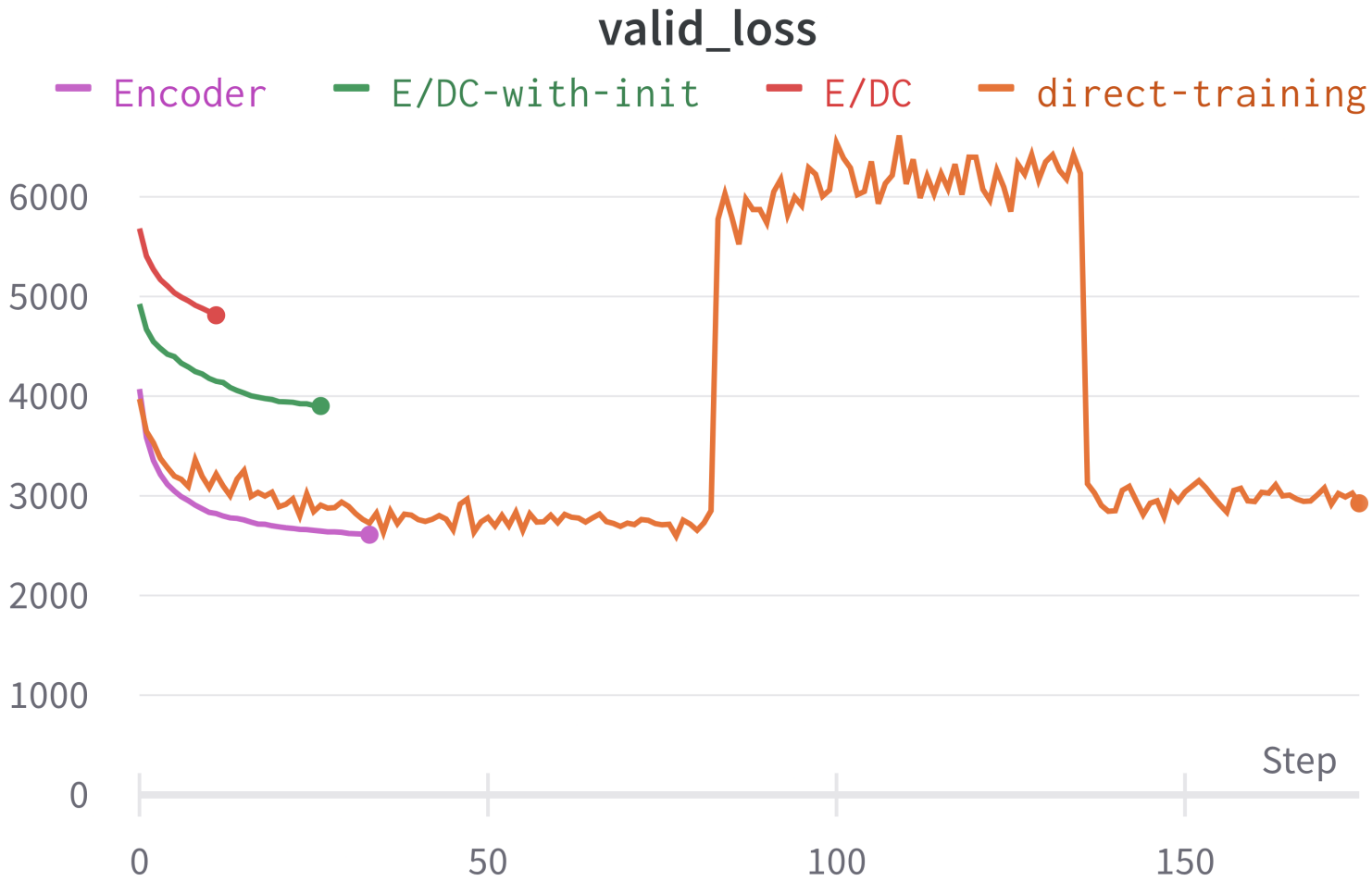


Figure: Validation loss

Training Experiments (4)

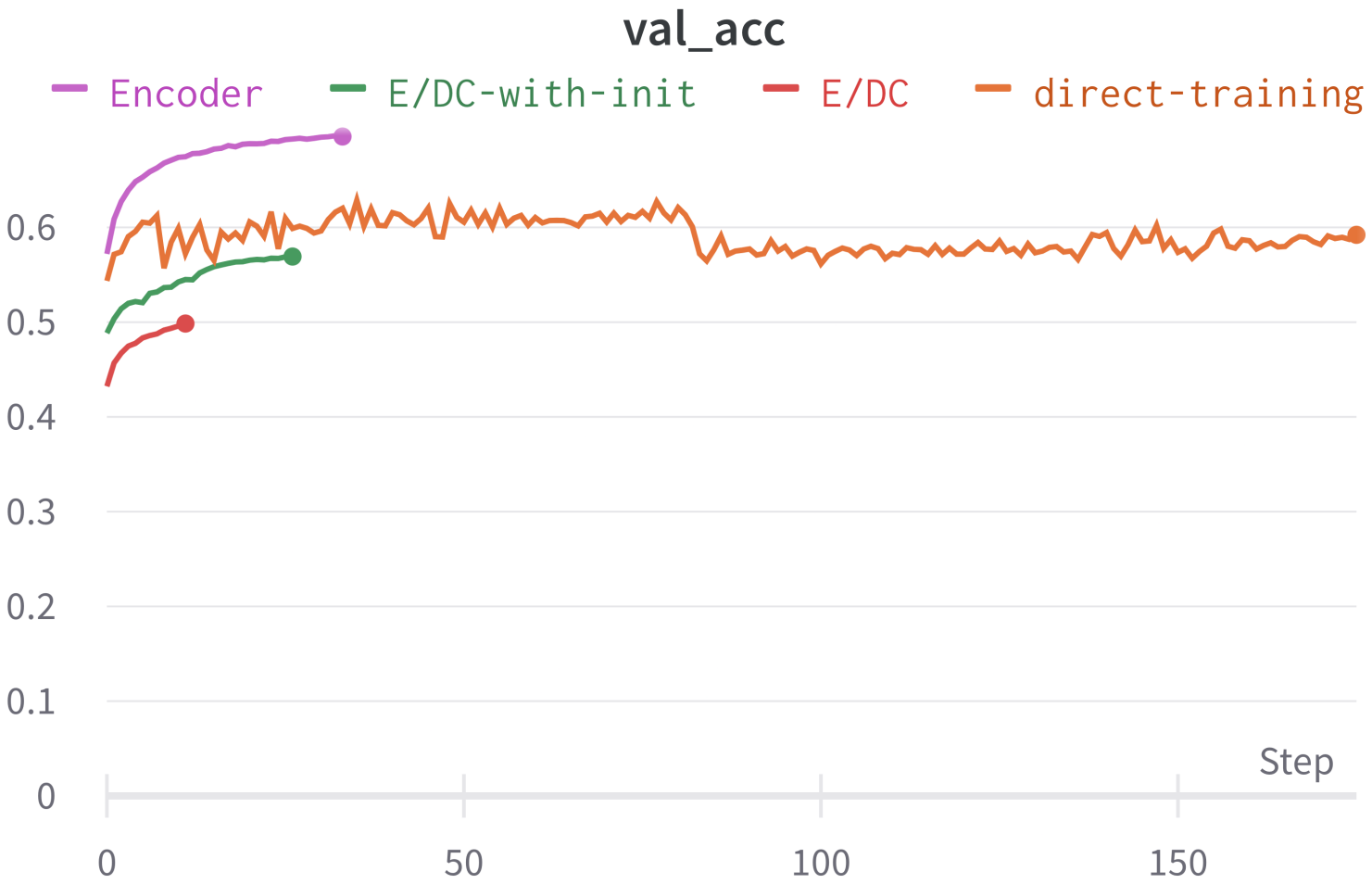


Figure: Validation accuracy