Densely Connected Convolutional Networks

Gao Huang[1], *Zhuang Liu*[2], *Laurens van der Maaten*[3], *Kilian Q. Weinberger*[1]

[1]Cornell University [2]Tsinghua University [3]Facebook AI Research

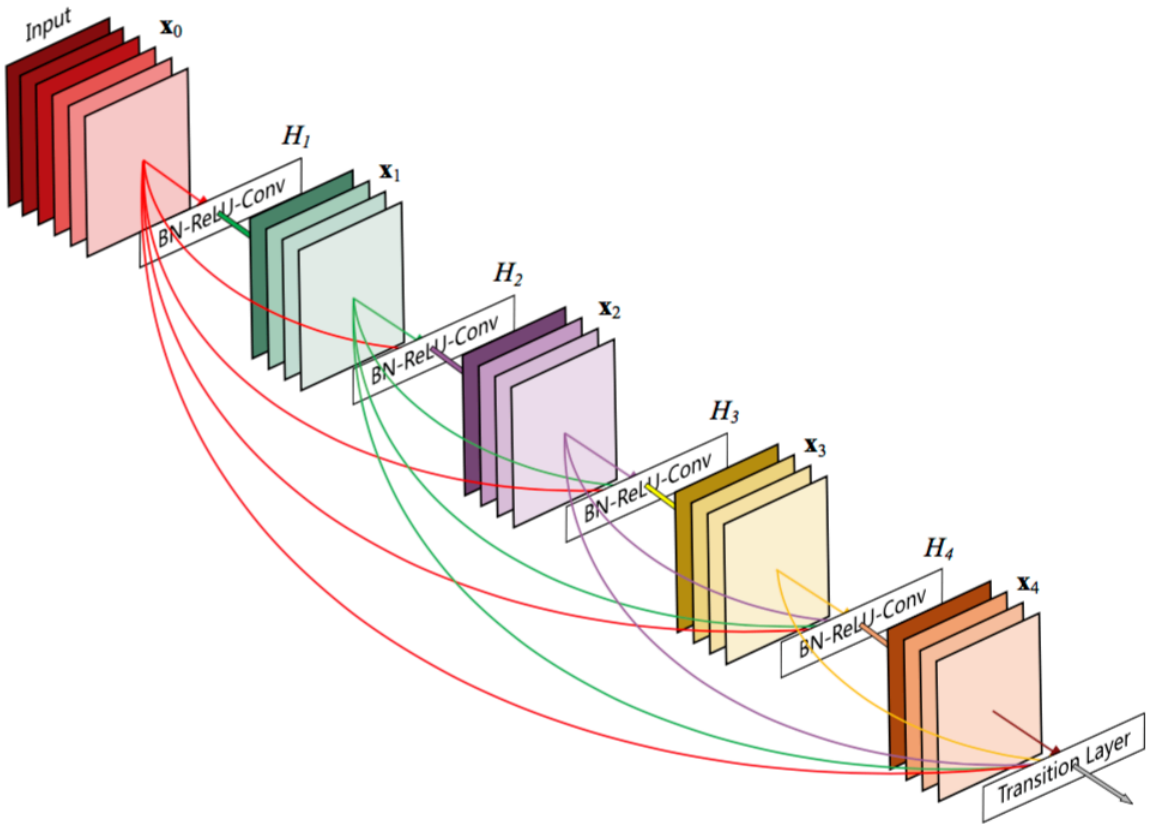
Architecture

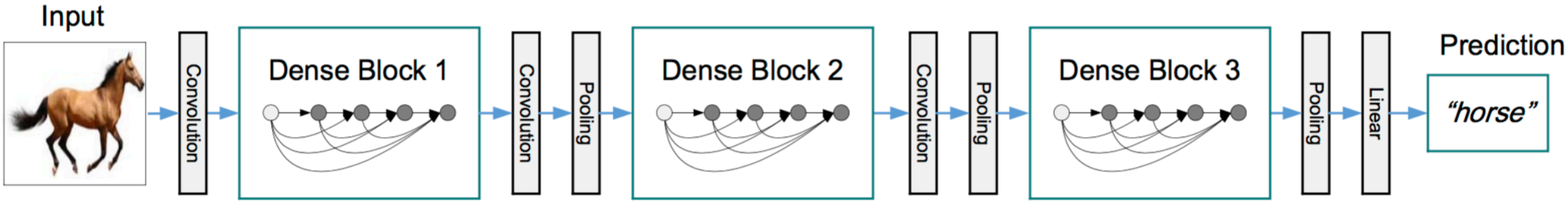
- the ℓ^{th} layer receives the feature-maps of all preceding layers,

  $\mathbf{x}_0, \ldots, \mathbf{x}_{\ell-1}$, as input:

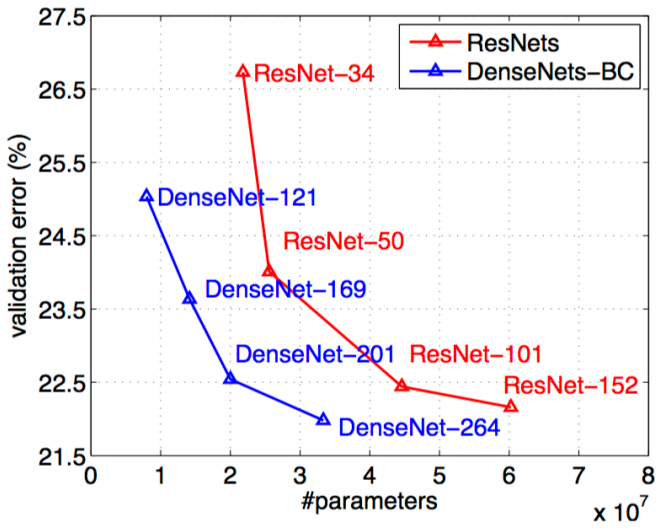
$$\mathbf{x}_\ell = H_\ell([\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{\ell-1}]),$$

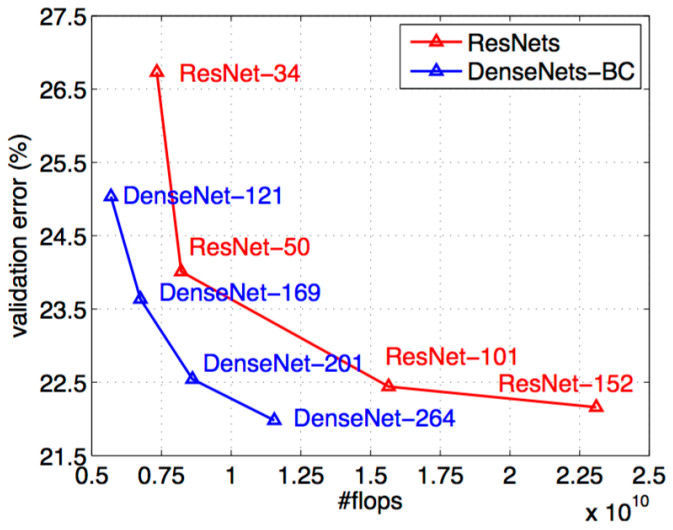
- Bottleneck layers, compression

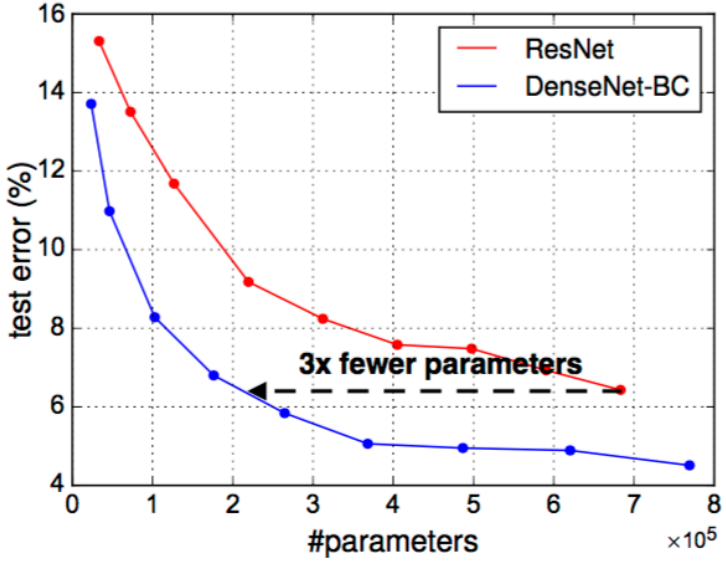


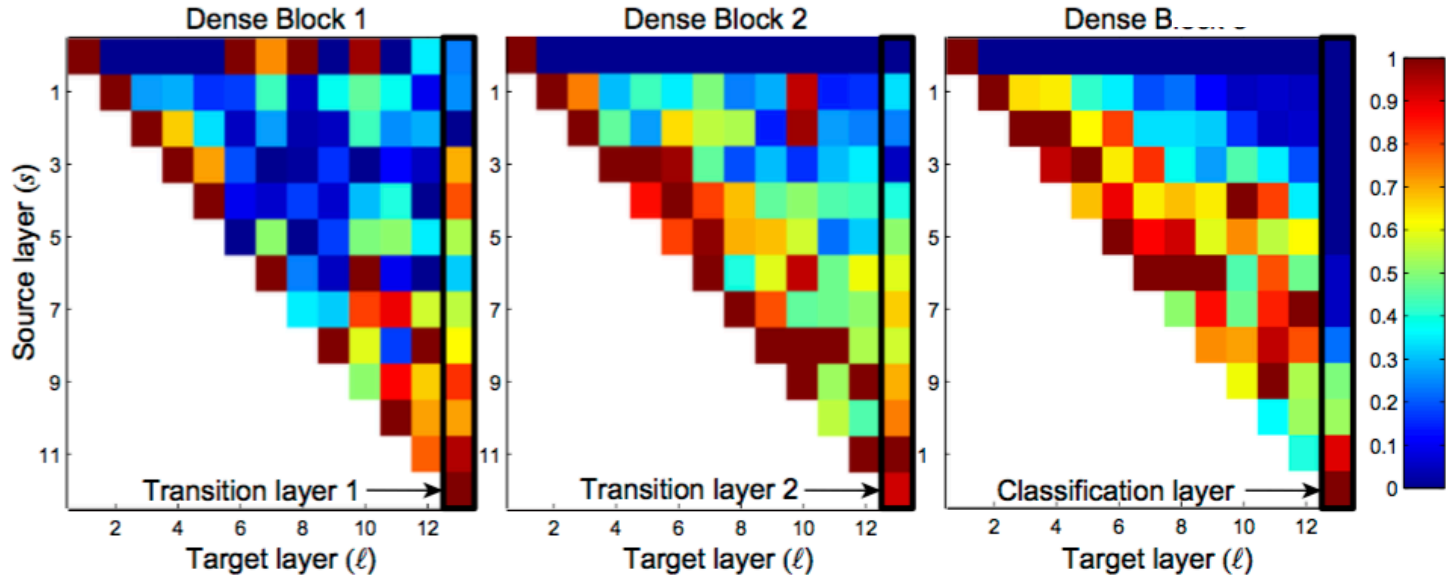


Experiments









Pros

- alleviate the vanishing-gradient problem
- strengthen feature propagation
- encourage feature reuse
- substantially reduce the number of parameters

Cons

- consume lots of memory

Dense Information Flow for Neural Machine Translation

Yanyao Shen[1], *Xu Tan*[2], *Di He*[3], *Tao Qin*[2], *Tie-Yan Liu*[2]

[1]University of Texas at Austin [2]Microsoft Research, Asia [3]Peking University

Dense Architecture

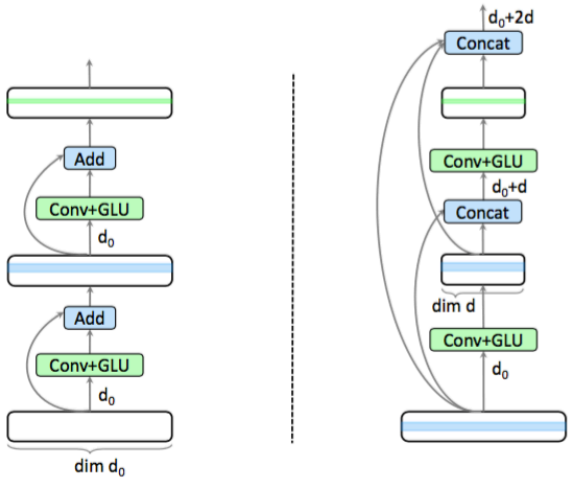


Figure 1: Comparison of dense-connected encoder and residual-connected encoder. Left: regular residual-connected encoder. Right: dense-connected encoder. Information is directly passed from blue blocks to the green block.

$$h^{l+1} = \mathcal{H}([h^l, h^{l-1}, \cdots, h^0]).$$

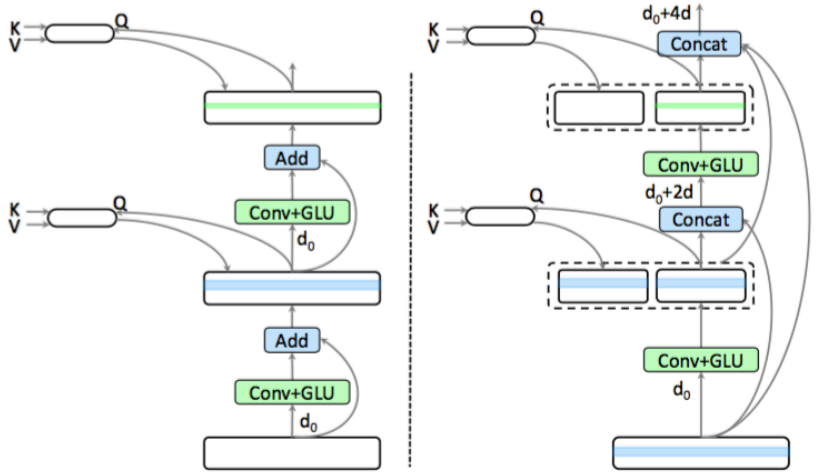
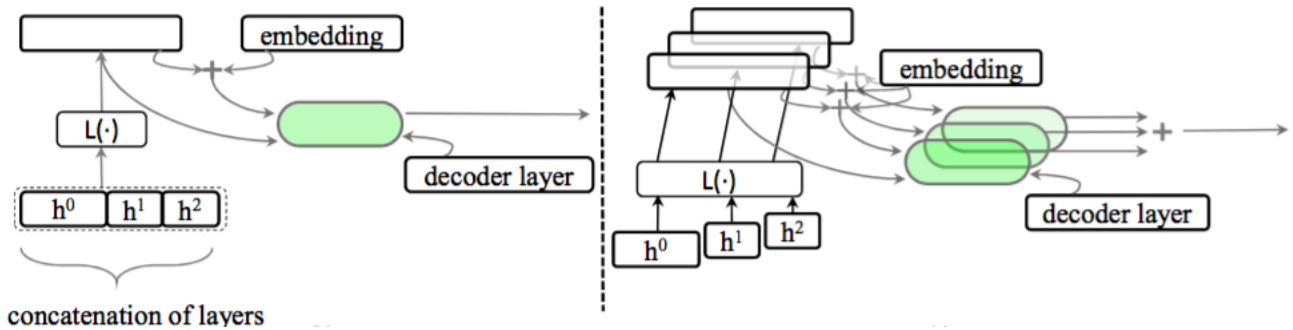


Figure 2: Comparison of dense-connected decoder and residual-connected decoder. Left: regular residual-connected decoder. Right: dense-connected decoder. Ellipsoid stands for attention block. Information is directly passed from blue blocks to the green block.

$$z^{l+1} = \mathcal{H}([z^l, a^l, z^{l-1}, a^{l-1}, \cdots, z^1, a^1, z^0])$$

Dense Attention

$$\mathcal{F}(Q, K, V) = \texttt{Softmax}(Q \times K) \times V,$$



$$a^l = \mathcal{F}\left(\mathcal{L}(z^l), \mathcal{L}\left([\{h^i\}]\right), \mathcal{L}\left([\{h^i\}]\right) + \mathcal{L}(h^0)\right)$$

$$a^l = \sum_{i=1}^{L} \mathcal{F}\left(\mathcal{L}(z^l), \mathcal{L}(h^i), \mathcal{L}([h^i, h^0])\right)$$

Summary Layer

Experiments

		De-En			Tr-En			Tr-En-morph		
	Embed size	64	128	256	64	128	256	64	128	256
	Model size (M)	8 ± 1	11 ± 1	17 ± 1	11 ± 1	17 ± 1	28 ± 1	13 ± 1	21 ± 1	36 ± 1
4L	BASE-4L	28.97	29.99	30.43	**19.80**	20.26	20.99	18.90	18.81	20.08
	DenseNMT-4L-1	**30.11**	**30.80**	31.26	19.21	20.08	21.36	18.83	20.16	21.43
	DenseNMT-4L-2	29.77	30.01	**31.40**	19.59	**20.86**	**21.48**	**19.04**	**20.19**	**21.57**
8L	BASE-8L	30.15	30.91	31.51	20.40	21.60	21.92	20.21	20.76	22.62
	DenseNMT-8L-1	**30.91**	**31.54**	32.08	21.82	**22.20**	23.20	21.20	21.73	22.60
	DenseNMT-8L-2	30.70	31.17	**32.26**	**21.93**	21.98	**23.25**	**21.73**	**22.44**	**23.45**

Table 1: BLEU score on IWSLT German-English and Turkish-English translation tasks. We compare models using different embedding sizes, and keep the model size consistent within each column.

Experiments

	Greedy	Beam
MIXER (Ranzato et al., 2015)	20.73	21.83
AC (Bahdanau et al., 2017)	27.49	28.53
NPMT (Huang et al., 2017)	27.83	28.96
NPMT+LM (Huang et al., 2017)	/	29.16
DenseNMT-8L-2 (word)	29.11	30.33
DenseNMT-8L-1 (BPE)	30.50	32.08
DenseNMT-8L-2 (BPE)	30.80	32.26

Table 4: Accuracy on IWSLT14 German-English translation task in terms of BLEU score.

	BLEU score
GNMT (Wu et al., 2016)	24.61
ConvS2S (Gehring et al., 2017)	25.16
SliceNet-Full (Kaiser et al., 2017)	25.5
SliceNet-Super (Kaiser et al., 2017)	26.1
DenseNMT-En-De-15	25.52

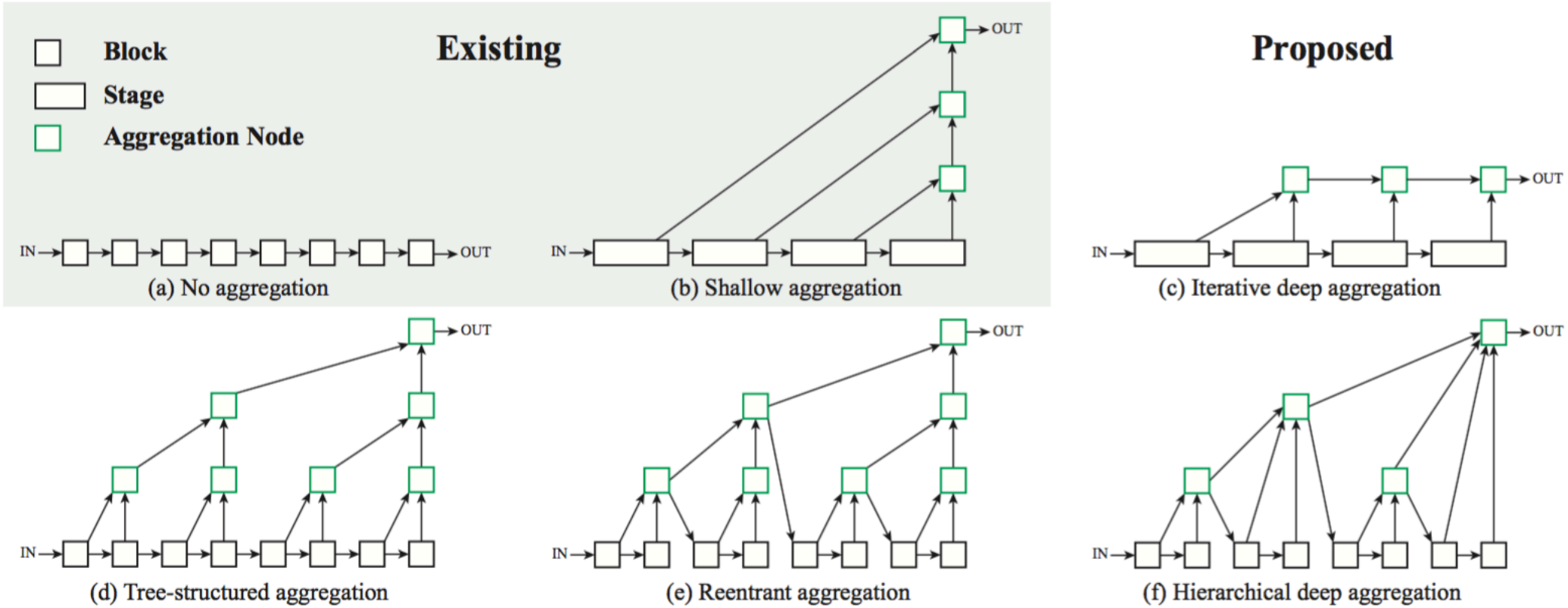
Table 5: Accuracy on WMT14 English-German translation task in terms of BLEU score.

Deep Layer Aggregation

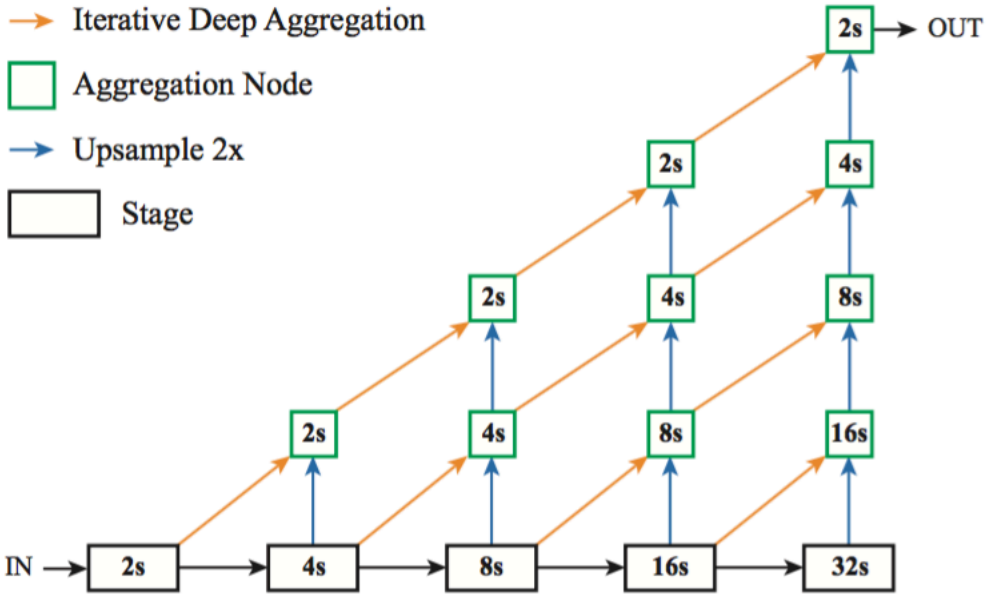
Fisher Yu, Dequan Wang, Evan Shelhamer, Trevor Darrell

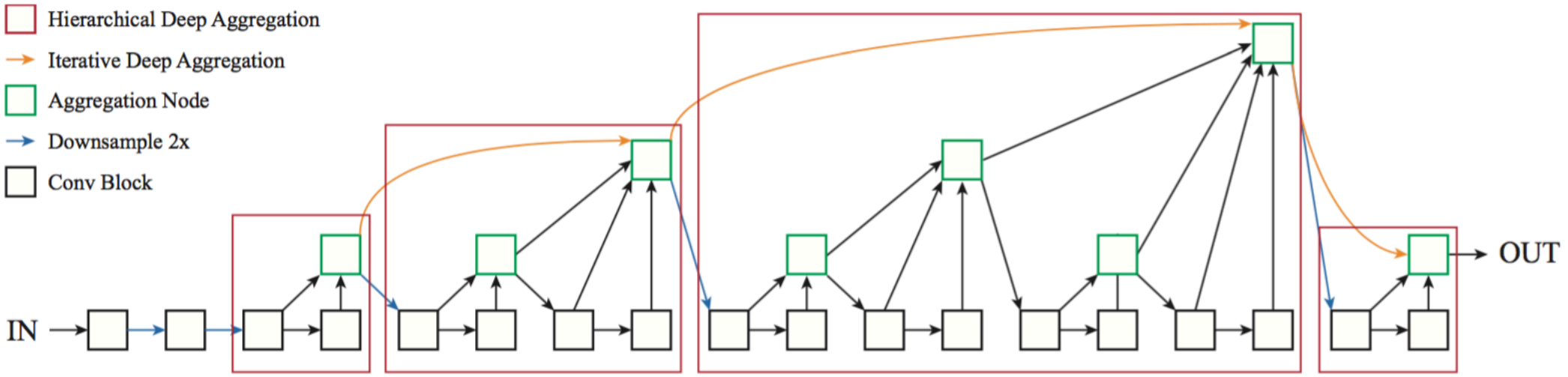
UC Berkeley

Architecture



Architecture





Experiments

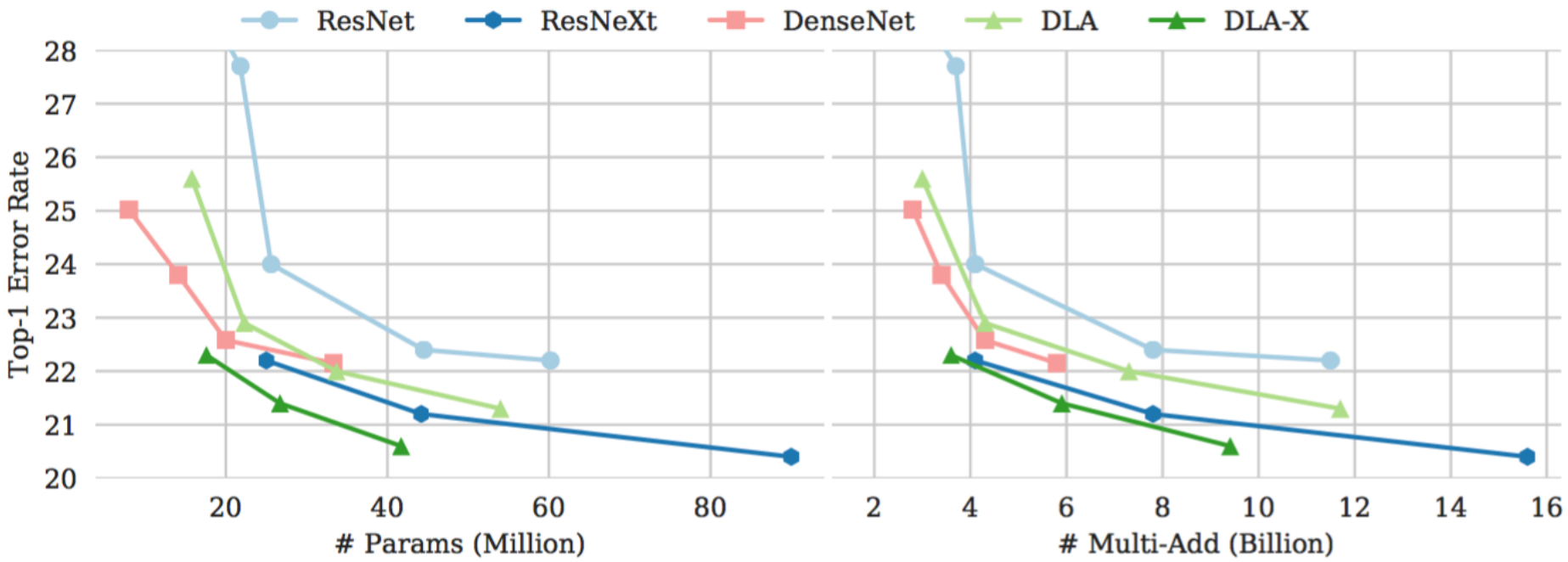


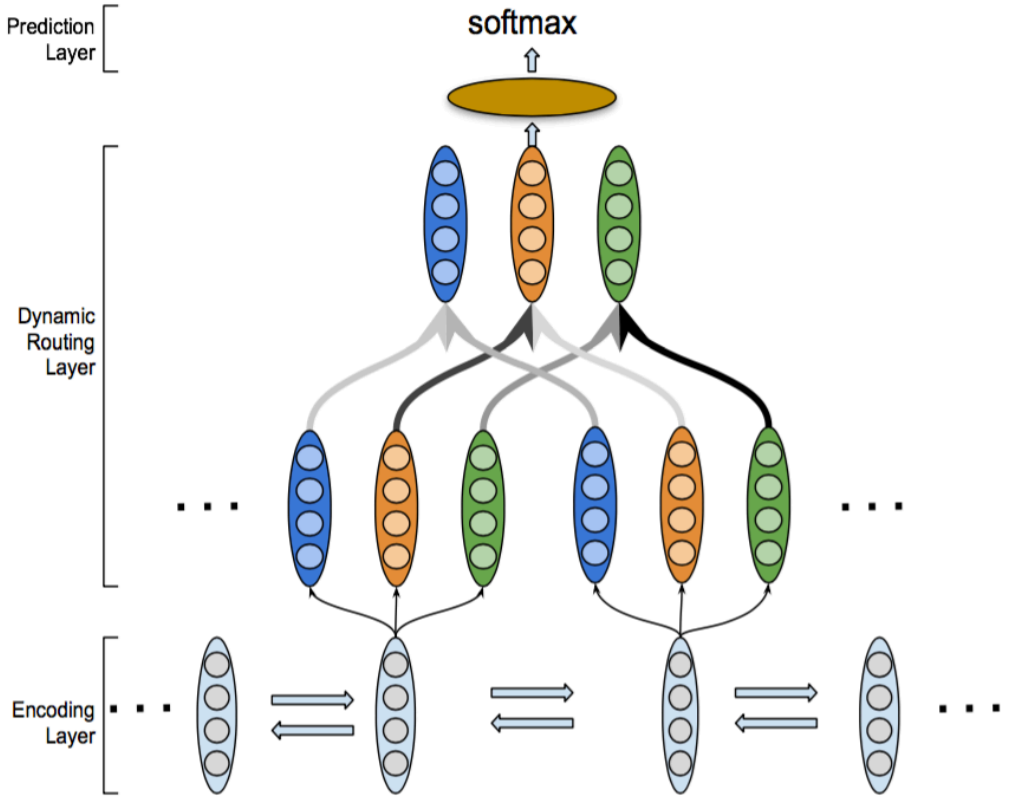
Figure 5: Evaluation of DLA on ILSVRC. DLA/DLA-X have ResNet/ResNeXT backbones respectively. DLA achieves the highest accuracies with fewer parameters and fewer computation.

Information Aggregation via Dynamic Routing for Sequence Encoding

Jingjing Gong, Xipeng Qiu, Shaojing Wang, Xuanjing Huang

Fudan University

Architecture



(a) Aggregation via Dynamic Routing

$$\mathbf{m}_{i\to j} = c_{ij} f(\mathbf{h}_i, \theta_j),$$

$$\mathbf{s}_j = \sum_{i=1}^{L} \mathbf{m}_{i\to j},$$

$$\mathbf{v}_j = \frac{\|\mathbf{s}_j\|^2}{1+\|\mathbf{s}_j\|^2} \frac{\mathbf{s}_j}{\|\mathbf{s}_j\|}$$

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})},$$

$$b_{ij} \leftarrow b_{ij} + \mathbf{v}_j^{\mathrm{T}} f(\mathbf{h}_i, \theta_j),$$

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})},$$

$$b_{ij} \leftarrow b_{ij} + \mathbf{v}_j^{\mathrm{T}} f(\mathbf{h}_i, \theta_j),$$

Algorithm 1: Dynamic Routing Algorithm

Data: Input Capsules: $\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_L$, Maximum number of Iterations: T
Result: Output Capsules: $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M$
Initialize $b_{ij} \leftarrow 0$;
for $t = 1$ **to** T **do**
  Compute the routing coefficients c_{ij} for all $i \in [1, L], j \in [1, M]$; /* Eq.15 */
  Update all the output capsule $\mathbf{v}_j, j \in [1, M]$; /* Eq. 13 and 14 */
  Update b_{ij} for all $i \in [1, L], j \in [1, M]$; /* Eq. 16 */
end for
return $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_M$

$$\mathbf{m}_{i \to j} = c_{ij} f(\mathbf{h}_i, \theta_j),$$

$$\mathbf{s}_j = \sum_{i=1}^{L} \mathbf{m}_{i \to j},$$

$$\mathbf{v}_j = \frac{\|\mathbf{s}_j\|^2}{1 + \|\mathbf{s}_j\|^2} \frac{\mathbf{s}_j}{\|\mathbf{s}_j\|}$$

Hierarchical Dynamic Routing for Long Text

- Split a document into sentences, and apply the proposed dynamic routing mechanism on word and sentence levels separately.
  - first encode each sentence into a fixed-length vector
  - convert the sentence encodings into document encoding.

Experiments

	Yelp-2013	Yelp-2014	IMDB	SST-1	SST-2
RNTN+Recurrent (Socher et al., 2013)	57.4	58.2	40.0	-	-
CNN-non-static (Kim, 2014)	-	-	-	48.0	87.2
Paragraph-Vec (Le and Mikolov, 2014)	-	-	-	48.7	**87.8**
MT-LSTM (F2S) (Liu et al., 2015)	-	-	-	49.1	87.2
UPNN(np UP) (Tang et al., 2015)	57.7	58.5	40.5	-	-
UPNN(full) (Tang et al., 2015)	59.6	60.8	43.5	-	-
Cached LSTM (Xu et al., 2016)	59.4	59.2	42.1	-	-
Max pooling	61.1	61.2	41.1	48.0	87.0
Average pooling	60.7	60.6	39.1	46.2	85.2
Self-attention	61.0	61.5	43.3	48.2	86.4
Standard DR-AGG	**62.1**	**63.0**	**45.1**	**50.5**	87.6
Reverse DR-AGG	61.6	62.5	44.5	49.3	87.2

(1) so relentlessly wholesome it made me want to swipe something .
(2) so relentlessly wholesome it made me want to swipe something .
(3) so relentlessly wholesome it made me want to swipe something .

Table 4: A visualization to show the perspective of a sentence from 3 different upper level capsule. A deeper color indicates more information of the associated word is routed to the corresponding capsule.