Deep learning

Deep dual learning[1]

Hamid Beigy

Sharif University of Technology

May 8, 2021

[1]Some slides are adopted from Tao Qin, Sreeja R Thoom et al. slides.

Table of contents

Introduction

1. Three Pillars of Deep Learning
   - Big data: web pages, search logs, social networks, and new mechanisms for data collection: conversation and crowd–sourcing.
   - Big models: 1000+ layers, tens of billions of parameters
   - Big computing: CPU clusters, GPU clusters, TPU clusters, FPGA farms, provided by Amazon, Azure, Ali etc.

1. Big-Data Challenge
   - Today's deep learning highly relies on huge amount of human-labeled training data

Task	Typical training data
Image classification	Millions of labeled images
Speech recognition	Thousands of hours of annotated voice data
Machine translation	Tens of millions of bilingual sentence pairs

   - Human labeling is in general very expensive, and it is hard, if not impossible, to obtain large-scale labeled data for rare domains

1. How translate from a source language to a destination language?
2. Main problems
   - How translate words from the source language to the destination language?
   - How order words in the destination language?
   - How measure goodness of translation?
   - What type of corpus is needed? (monolingual or bilingual)
   - How build a sequence of translators? (Persian $\rightarrow$ English $\rightarrow$ French)

1. In NMT, recurrent neural networks such as LSTM or GRU units are used (Bahdanau, Cho, and Bengio 2015).

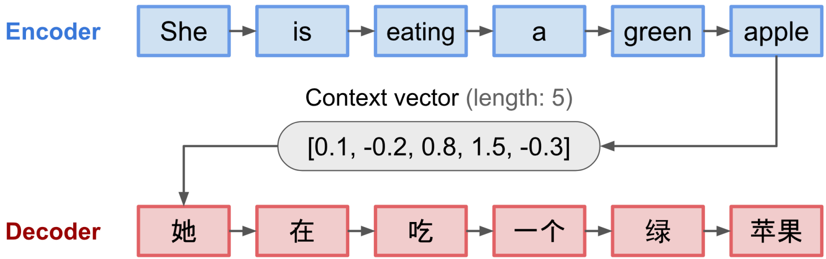
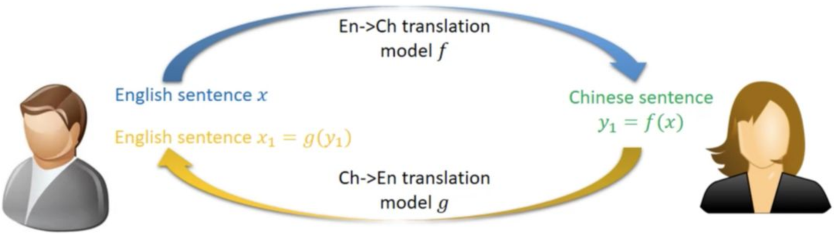


Figure: Lilian Weng

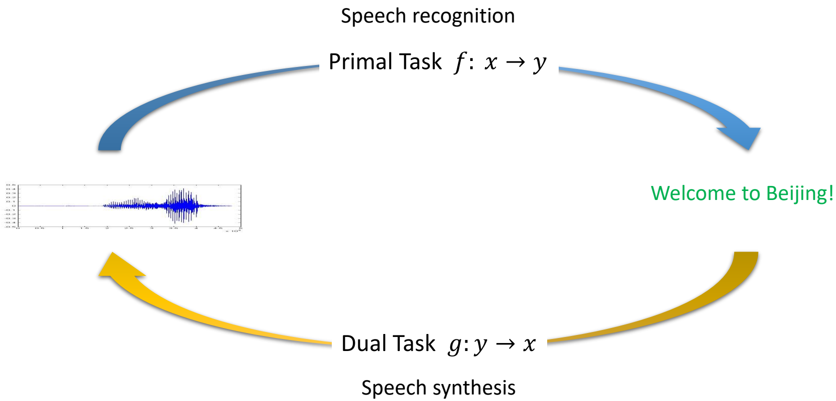
2. A critical disadvantage of this fixed-length context vector design is incapability of remembering long sentences.
3. The attention mechanism was proposed to help memorize long source sentences in NMT
4. Another critical disadvantage of this model is training set. We need a large bilingual corpus.
5. Dual learning was introduced to overcome the need for a large bilingual corpus.

Dual learning

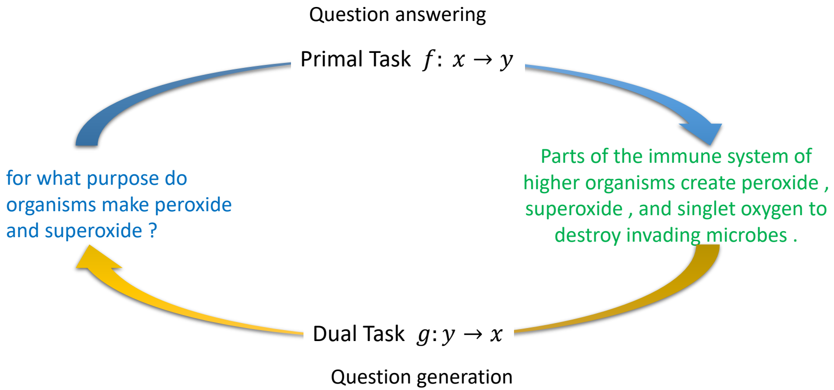
1. Dual learning is a auto-encoder like mechanism to utilize the monolingual datasets (He et al. 2016).



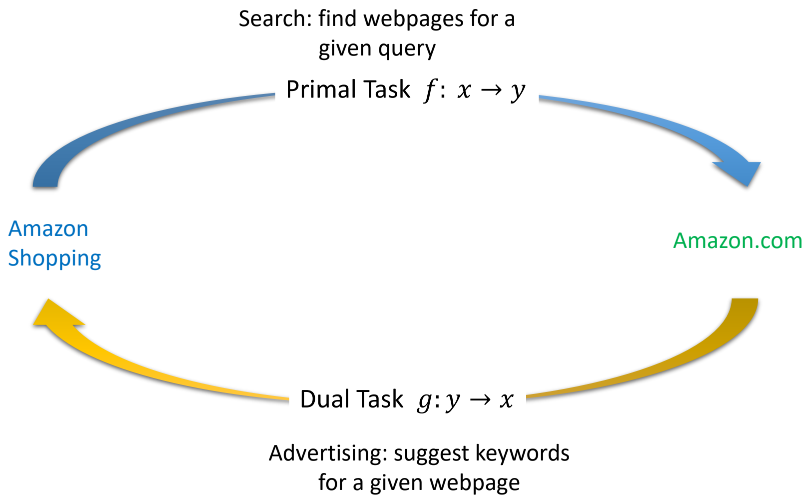
1. Duality in Speech Processing.



1. Duality in Question Answering and Generation.



1. Duality in Search and Advertising.



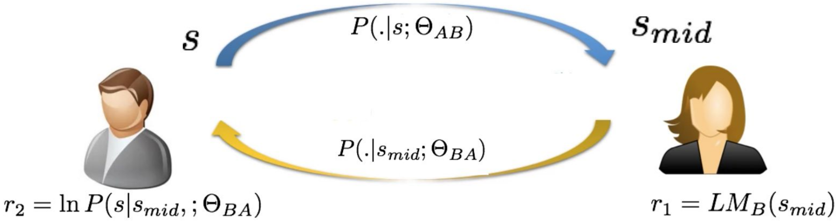
Structural duality is very common in artificial intelligence

AI Tasks	$X \rightarrow Y$	$Y \rightarrow X$
Machine translation	Translation from EN to CH	Translation from CH to EN
Speech processing	Speech recognition	Text to speech
Image understanding	Image captioning	Image generation
Conversation	Question answering	Question generation
Search engine	Query-document matching	Query/keyword suggestion

Currently most machine learning algorithms do not exploit structure duality for training and inference.

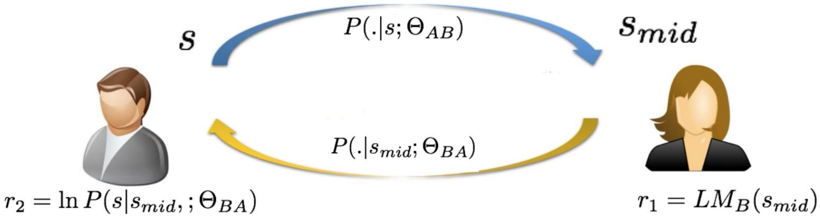
1. A new learning framework that leverages the symmetric (primal-dual) structure of AI tasks to obtain effective feedback or regularization signals to enhance the learning/inference process.
2. If you don't have enough labeled data for training, can we use unlabeled data?
3. Dual Unsupervised Learning can leverage structural duality to learn from unlabeled data.

1. Let us to define (He et al. 2016)
   - D_A Corpus of language A.
   - D_B Corpus of language B.
   - $P(.|s, \theta_{AB})$ translation model from A to B.
   - $P(.|s, \theta_{BA})$ translation model from B to A.
   - $LM_A(.)$ learned language model of A.
   - $LM_B(.)$ learned language model of B.



Dual learning (Algorithm)

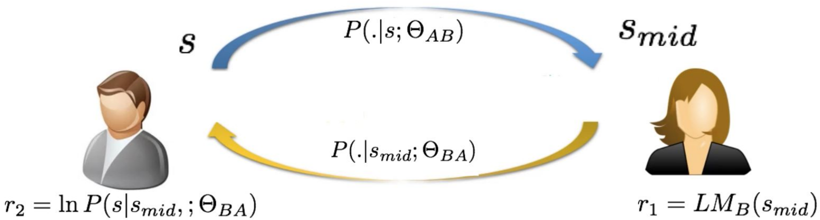
1. We have



2. Generate K translated sentences $s_{mid,1}, s_{mid,2}, \ldots, s_{mid,K}$ from $P(.|s, \theta_{AB})$
3. Compute intermediate rewards $r_{1,1}, r_{1,2}, \ldots, r_{1,K}$ from $LM_B(s_{mid,K})$ for each sentence as $r_{1,k} = LM_B(s_{mid,k})$

Dual learning (Algorithm)

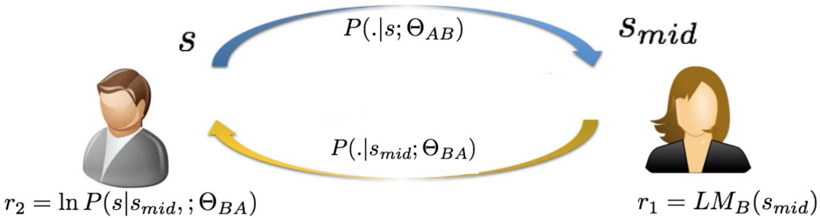
1. We have



2. Compute communication rewards $r_{2,1}, r_{2,2}, \ldots, r_{2,K}$ for each sentence as $r_{2,k} = \ln P(s|s_{mid},;\theta_{BA})$
3. Set the total reward of kth sentence as $r_k = \alpha r_{1,k} + (1-\alpha) r_{2,k}$

Dual learning (Algorithm)

1. We have



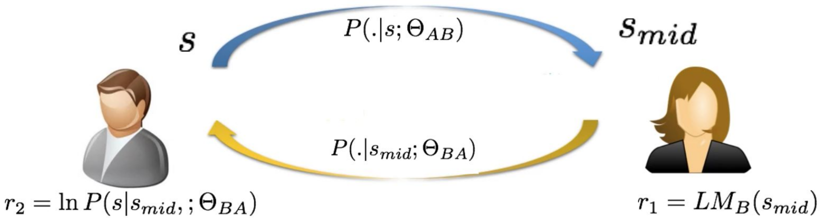
2. Compute the stochastic gradient of θ_{AB} and θ_{BA}

$$\nabla_{\theta_{AB}}\mathbb{E}[r] = \frac{1}{K}\sum_{k=1}^{K}\alpha\nabla_{AB}\ln P(s_{mid,k}|s,\theta_{AB})$$

$$\nabla_{\theta_{BA}}\mathbb{E}[r] = \frac{1}{K}\sum_{k=1}^{K}(1-\alpha)\nabla_{BA}\ln P(s_{mid,k}|s,\theta_{BA})$$

Dual learning (Algorithm)

1. We have



2. Update the mode parameters θ_{AB} and θ_{BA}

$$\theta_{AB} \leftarrow \theta_{AB} + \gamma_1 \nabla_{\theta_{AB}} \mathbb{E}[r]$$
$$\theta_{BA} \leftarrow \theta_{BA} + \gamma_2 \nabla_{\theta_{BA}} \mathbb{E}[r]$$

Dual learning algorithm (pseudo code))

Algorithm 1 The dual-learning algorithm

```
1: Input: Monolingual corpora D_A and D_B, initial translation models Θ_AB and Θ_BA, language
   models LM_A and LM_B, α, beam search size K, learning rates γ_{1,t}, γ_{2,t} .
2: repeat
3:     t = t + 1.
4:     Sample sentence s_A and s_B from D_A and D_B respectively.
5:     Set s = s_A.                               ▷ Model update for the game beginning from A.
6:     Generate K sentences s_{mid,1}, ..., s_{mid,K} using beam search according to translation model
   P(.|s; Θ_AB).
7:     for k = 1, ..., K do
8:         Set the language-model reward for the kth sampled sentence as r_{1,k} = LM_B(s_{mid,k}).
9:         Set the communication reward for the kth sampled sentence as r_{2,k} =
   log P(s|s_{mid,k}; Θ_BA).
10:        Set the total reward of the kth sample as r_k = α r_{1,k} + (1 − α) r_{2,k}.
11:    end for
12:    Compute the stochastic gradient of Θ_AB:
```

$$\nabla_{\Theta_{AB}} \hat{E}[r] = \frac{1}{K} \sum_{k=1}^{K} [r_k \nabla_{\Theta_{AB}} \log P(s_{mid,k} | s; \Theta_{AB})].$$

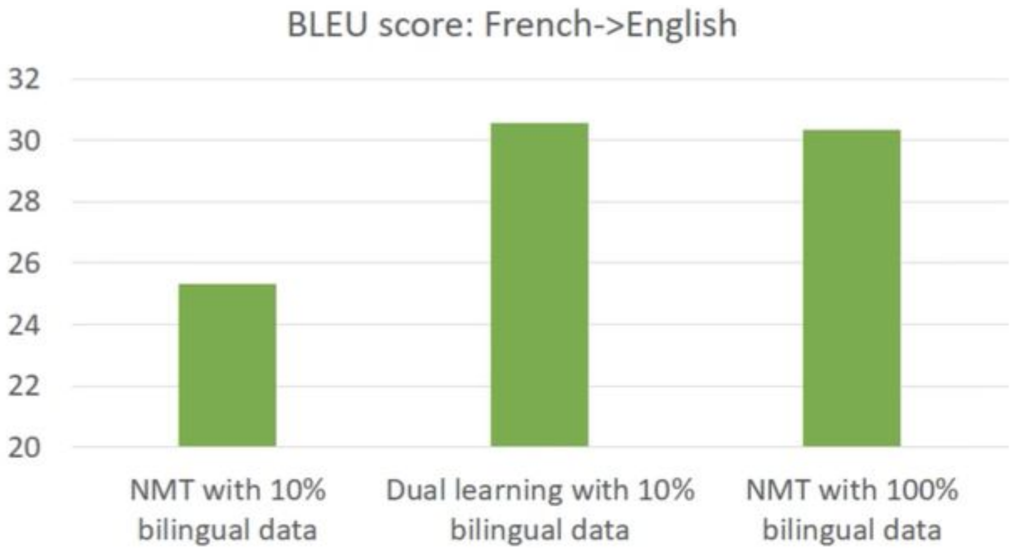
```
13:    Compute the stochastic gradient of Θ_BA:
```

$$\nabla_{\Theta_{BA}} \hat{E}[r] = \frac{1}{K} \sum_{k=1}^{K} [(1-\alpha) \nabla_{\Theta_{BA}} \log P(s | s_{mid,k}; \Theta_{BA})].$$

```
14:    Model updates:
```

$$\Theta_{AB} \leftarrow \Theta_{AB} + \gamma_{1,t} \nabla_{\Theta_{AB}} \hat{E}[r], \Theta_{BA} \leftarrow \Theta_{BA} + \gamma_{2,t} \nabla_{\Theta_{BA}} \hat{E}[r].$$

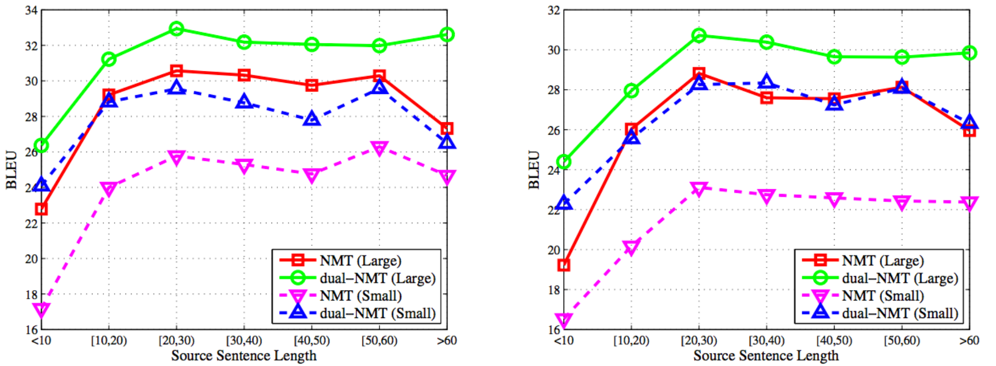
```
15:    Set s = s_B.                               ▷ Model update for the game beginning from B.
16:    Go through line 6 to line 14 symmetrically.
17: until convergence
```


BLEU score: French->English
32
30
28
26
24
22
20
NMT with 10% bilingual data
Dual learning with 10% bilingual data
NMT with 100% bilingual data

1. Reconstruction performance (BLEU: geometric mean of *n*-gram precision)

	En→Fr→En (L)	Fr→En→Fr (L)	En→Fr→En (S)	Fr→En→Fr (S)
NMT	39.92	45.05	28.28	32.63
pseudo-NMT	38.15	45.41	30.07	34.54
dual-NMT	**51.84**	**54.65**	**48.94**	**50.38**

1. For different source sentence length (Improvement is significant for long sentences)



(a) En→Fr (b) Fr→En

1. Reconstruction examples

	Translation-back-translation results before dual-NMT training	Translation-back-translation results after dual-NMT training
Source (En)	The majority of the growth in the years to come will come from its liquefied natural gas schemes in Australia.	
En→Fr	La plus grande partie de la crois- -sance des années à venir viendra de ses systèmes de gaz naturel liquéfié en Australie .	La majorité de la croissance dans les années à venir viendra de ses régimes de gaz naturel liquéfié en Australie .
En→Fr→En	Most of the growth of future years will come from its liquefied natural gas systems in Australia .	The majority of growth in the coming years will come from its liquefied natural gas systems in Australia .
Source (Fr)	Il précise que " les deux cas identifiés en mai 2013 restent donc les deux seuls cas confirmés en France à ce jour " .	
Fr→En	He noted that " the two cases identified in May 2013 therefore remain the only two two confirmed cases in France to date " .	He states that " the two cases identified in May 2013 remain the only two confirmed cases in France to date "
Fr→En→Fr	Il a noté que " les deux cas identifiésen mai 2013 demeurent donc les deux seuls deux deux cas confirmés en France à ce jour "	Il précise que " les deux cas identifiés en mai 2013 restent les seuls deux cas confirmés en France à ce jour ".

Dual Supervised Learning

1. Given m training pairs $\{(x_1, y_2), \ldots, (x_m, y_m)\}$ sampled from the space $\mathcal{X} \times \mathcal{Y}$.
2. Learn the bi-directional relationship of (x, y), in two independent supervised learning tasks (primal f and dual g) (Xia et al. 2017):

$$\min_{\theta_{xy}} \frac{1}{m} \sum_{i}^{m} \mathrm{L}_1 \left(f(x_i; \theta_{xy}), y_i\right)$$

$$\min_{\theta_{yx}} \frac{1}{m} \sum_{i}^{m} \mathrm{L}_2 \left(g(y_i; \theta_{yx}), x_i\right)$$

3. If the learned primal and dual models are perfect, for all x and y, we should have

$$P(x)P(y|x; \theta_{xy}) = P(y)P(x|y; \theta_{yx})$$

1. Incorporate joint distribution matching in supervised learning

$$\min_{\theta_{xy}} \frac{1}{m} \sum_{i}^{m} \mathrm{L}_1 \left(f(x_i; \theta_{xy}), y_i \right)$$

$$\min_{\theta_{yx}} \frac{1}{m} \sum_{i}^{m} \mathrm{L}_2 \left(g(y_i; \theta_{yx}), x_i \right)$$

$$P(x)P(y|x; \theta_{xy}) = P(y)P(x|y; \theta_{yx})$$

2. Empirical marginal distributions $\hat{P}(x)$ and $\hat{P}(y)$

$$\mathrm{L}_{duality} = \left(\log \hat{P}(x) + \log \hat{P}(y|x; \theta_{xy}) \right) - \left(\log \hat{P}(y) + \log \hat{P}(x|y; \theta_{yx}) \right)$$

Algorithm 1 Dual Supervise Learning Algorithm

Input: Marginal distributions $\hat{P}(x_i)$ and $\hat{P}(y_i)$ for any $i \in [n]$; Lagrange parameters λ_{xy} and λ_{yx}; optimizers Opt_1 and Opt_2;

repeat

Get a minibatch of m pairs $\{(x_j, y_j)\}_{j=1}^m$;

Calculate the gradients as follows:

$$\begin{aligned} G_f = \nabla_{\theta_{xy}}(1/m)\textstyle\sum_{j=1}^m \big[\ell_1(f(x_j;\theta_{xy}), y_j) \\ + \lambda_{xy}\ell_{\text{duality}}(x_j, y_j; \theta_{xy}, \theta_{yx})\big]; \\ G_g = \nabla_{\theta_{yx}}(1/m)\textstyle\sum_{j=1}^m \big[\ell_2(g(y_j;\theta_{yx}), x_j) \\ + \lambda_{yx}\ell_{\text{duality}}(x_j, y_j; \theta_{xy}, \theta_{yx})\big]; \end{aligned} \tag{4}$$

Update the parameters of f and g:

$\theta_{xy} \leftarrow Opt_1(\theta_{xy}, G_f)$, $\theta_{yx} \leftarrow Opt_2(\theta_{yx}, G_g)$.

until models converged

Supervised Dual learning Algorithm results

Tasks	RNNSearch	DSL	Δ
En→Fr	29.92	31.99	2.07
Fr→En	27.49	28.35	0.86
En→De	16.54	17.91	1.37
De→En	20.69	20.81	0.12
En→Zh (MT08)	15.45	15.87	0.42
Zh→En (MT08)	31.67	33.59	1.92
En→Zh (MT12)	15.05	16.10	1.05
Zh→En (MT12)	30.54	32.00	1.46

Reading

Readings

1. Read the survey paper (Khoshvishkaie and Beigy 2020).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *International Conference on Learning Representations.*

He, Di et al. (2016). “Dual Learning for Machine Translation”. In: *Advances in Neural Information Processing Systems*, pp. 820–828.

Khoshvishkaie, Ali Akbar and Hamid Beigy (2020). “Deep Learning: Methods and Applications”. In: *The CSI Journal on Computing Science and Information Technology* 17.2, pp. 33–44. URL: https://jcsit.ir/article/86.

Xia, Yingce et al. (2017). “Dual Supervised Learning”. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70, pp. 3789–3798.

Questions?