Dual Inference for Machine Learning

Yingce Xia[1,*], **Jiang Bian**[2], **Tao Qin**[2], **Nenghai Yu**[1] and **Tie-Yan Liu**[2]
[1]University of Science and Technology of China, Hefei, Anhui, China
[2]Microsoft Reasearch Asia, Beijing, China
yingce.xia@gmail.com, {jiabia,taoqin,tie-yan.liu}@microsoft.com, ynh@ustc.edu.cn

Abstract

Recent years have witnessed the rapid development of machine learning in solving artificial intelligence (AI) tasks in many domains, including translation, speech, image, etc. Within these domains, AI tasks are usually not independent. As a specific type of relationship, structural duality does exist between many pairs of AI tasks, such as translation from one language to another vs. its opposite direction, speech recognition vs. speech synthetization, image classification vs. image generation, etc. The importance of such duality has been magnified by some recent studies, which revealed that it can boost the learning of two tasks in the dual form. However, there has been little investigation on how to leverage this invaluable relationship into the inference stage of AI tasks. In this paper, we propose a general framework of *dual inference* which can take advantage of both existing models from two dual tasks, without re-training, to conduct inference for one individual task. Empirical studies on three pairs of specific dual tasks, including machine translation, sentiment analysis, and image processing have illustrated that dual inference can significantly improve the performance of each of individual tasks.

1 Introduction

Due to the splendid power of mining big data, machine learning algorithms have played a critical role in solving artificial intelligence (AI) tasks in many practical domains, including machine translation, speech analysis, image processing, etc. Various particular AI tasks, though different in their goals and formations, are usually not independent and yield diverse relationships between each other within each domain. Among them, structural duality emerges as one of the important relationships. Two AI tasks are of structure duality if the goal of one task is to learn a function mapping from space $\mathcal{X}$ to $\mathcal{Y}$, the other's goal is to learn a reverse mapping from $\mathcal{Y}$ and $\mathcal{X}$. Note that we could call either of these two tasks as the primal task and the other as the dual one.

Duality does exist between many pairs of AI tasks in real-world. For example, machine translation [Wu *et al.*, 2016] from one language to another, e.g., English-to-French, and that of the opposite direction, e.g. French-to-English, form up a typical dual form; speech recognition [Graves *et al.*, 2013; Amodei *et al.*, 2016] and speech generation/synthesis [Oord *et al.*, 2016] constitute a duality in the domain of speech processing; besides, a specific pair of dual tasks in face attribute manipulation can be comprised of the task of removing glass from face and that of wearing glass to face [Shen and Liu, 2016]. Beyond, advanced deep learning algorithms can formulate those pairs of tasks without explicit duality, such as image classification [He *et al.*, 2016b; 2016c] and conditional image generation [van den Oord *et al.*, 2016b; 2016a], into the dual form.

As common in real-world applications, the duality can provide vital knowledge for enhancing learning tasks. A recent study [He *et al.*, 2016a] has magnified its importance by introducing a new *dual learning* framework to boost the learning of two tasks in the dual form. In particular, by leveraging unlabeled data, this work exploited the structure duality in machine translation to design a two-player game with a closed-loop feedback system as a dual Neural Machine Translation (dual-NMT) algorithm. [Xia *et al.*, 2017] explored duality for supervised learning, whose idea is to increase the probabilistic consistency of the two dual tasks. Another effort [Shen and Liu, 2016] attempted to combine duality with adversarial training to improve the performance of face attribute manipulation.

To the best of our knowledge, all existing studies regarding the duality focus on utilizing it to boost the training process so as to obtain more powerful models. However, there has been little investigation on how to leverage this invaluable relationship into the inference stage of AI tasks. Intuitively, we have high confidence to judge y is a good output for the input x in the primal task, if x is a good output for y in the dual task. Therefore, in this paper, we propose a general framework of *dual inference* which can take advantage of both existing models from two dual tasks, without re-training, to conduct inference for each individual task.

To better illustrate the high effectiveness of dual inference, we apply it into dual AI tasks in three particular domains:

[*]This work was done when the first author was an intern at Microsoft Research Asia.

(1) *Neural Machine Translation (NMT)*: Translation from a source language into a target language naturally yields a dual task of inverse translation from the target to the source. NMT, emerging as widely-used approach, employs a Recurrent Neural Network based encoder-decoder framework to model the probability of a sentence in target language conditioned on the sentence in source one.
(2) *Sentiment Analysis*: Sentiment classification, aiming at predict the sentiment label of sentences, is a popular primal task in the domain of sentiment analysis. The corresponding dual task, is indeed sentence generation, whose objective is to automatically generate sentences based on a pre-designed sentiment.
(3) *Image Processing*: Image classification, the goal of which is to predict the label of an image, is one of major AI tasks in the domain of image processing. The dual task of image classification is obviously image generation, which is an emerging AI task to automatically generate images based on category labels.

Empirical studies on dual tasks under these three specific domains have shown that dual inference can significantly improve the inference performance of each of individual tasks. We would like to point out that such improvement are achieved without changing/re-training the original primal and dual models. Moreover, we provide theoretical discussions to provide better understanding on dual inference.

2 Dual Inference Framework

As structural duality is popular and yields important knowledge is many AI applications, in this section, we propose a general framework of *dual inference* to leverage existing models of both dual tasks for better inference for each individual task.

Assume there are two tasks in the dual form in a particular AI application. We use $f : \mathcal{X} \mapsto \mathcal{Y}$ to denote the model for the primal task which is a mapping from space $\mathcal{X}$ to space $\mathcal{Y}$, and use $g : \mathcal{Y} \mapsto \mathcal{X}$ to denote the model for the dual task[1]. The loss functions corresponding to f and g are represented as $\ell_f(x, y)$ and $\ell_g(x, y)$, respectively, which are mappings from the product space $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$.

There are various potential rules to formulate duality into the dual inference framework. In this paper, we employ a most natural and straightforward approach, which first combines the loss functions of the primal task and the dual task and then selects the output that can minimize the combined loss as the inference result. More formally, we have following dual inference equations for the primal and dual task, respectively:

$$f_{\text{dual}}(x) = \arg\min_{y' \in \mathcal{Y}} \{\alpha \ell_f(x, y') + (1 - \alpha)\ell_g(x, y')\}, \quad (1)$$

$$g_{\text{dual}}(y) = \arg\min_{x' \in \mathcal{X}} \{\beta \ell_g(x', y) + (1 - \beta)\ell_f(x', y)\}, \quad (2)$$

where α and β are hyperparameters to balance the tradeoff between two losses, and they will be tuned based on performance on a validation set. Note that we do not re-train or make any change on the models of both primal and dual tasks.

Most of inference rules, currently widely-used in machine learning tasks, can be described as below.

$$f(x) = \arg\min_{y' \in \mathcal{Y}} \ell_f(x, y'); \;\; g(y) = \arg\min_{x' \in \mathcal{X}} \ell_g(x', y). \quad (3)$$

which are extreme cases in dual inference by setting α and β to one. From this perspective, dual inference can be viewed as a more general inference framework.

Note that, a branch of inference rules using multiple models correspond to the ensemble [Opitz and Maclin, 1999] methods. However, dual inference yields a crucial difference from the ensemble. In particular, all models in an ensemble framework follow the same mapping direction, thus they can only serve for either the primal or the dual task, whilst the two models applied in the dual inference framework serve for both the primal and the dual tasks with opposite mapping directions, simultaneously.

To gain better understanding of dual inference, in the following, we apply this new framework into dual AI tasks in three particular domains and conduct corresponding empirical studies to examine the effectiveness of dual inference.

3 Neural Machine Translation

Structural duality apparently exists in the scenario of machine translation. Specifically, translation from a source language into a target language naturally yields a dual task of inverse translation from the target to the source. As a state-of-the-art approach, neural machine translation (NMT) is a deep learning based end to end approach for machine translation. NMT models the conditional probability $P(y|x; \theta)$ of a sentence y in target language given a sentence x in source language, and the parameter θ is learned based on the training data consisting of a set of bilingual sentence pairs. During the typical inference step, given a source sentence x, NMT finds the target sentence y with largest conditional probability $P(y|x; \theta)$ as the translation of x. Since the number of candidate target sentences is exponentially large, it usually employs beam search to find a reasonable target y more efficiently.

Due to the natural existence in NMT, structural duality has been exploited into the learning process of NMT [He *et al.*, 2016a]. However, there has been little investigation on how to leverage duality into the inference stage. Hence, we will examine how to apply dual inference into NMT in the following of this section.

3.1 Dual Inference for NMT

Let f denote the machine translation model from language X to Y and let g denote that from Y to X. Following the widely used work [Bahdanau *et al.*, 2015], the loss functions used for inference in two directions, represented as ℓ_f and ℓ_g respectively, are specialized as negative log-likelihood in machine translation. Mathematically,

$$\ell_f(x, y) = -\log P(y|x; f), \ell_g(x, y) = -\log P(x|y; g), \quad (4)$$

The dual inference for the primal task of neural machine translation is shown as follows:

[1]There are some tasks like generating images from a given label. In this case, usually the mapping function g needs additional inputs, e.g., random vectors. With a little confusion, we still say that g maps the points in $\mathcal{Y}$ to $\mathcal{X}$.

1. Translate source x with beam search by model f and get K candidates $\hat{y}_i$ $i \in [K]$; (K is beam size)
2. $i^* = \arg\min_{i\in[K]} \alpha\ell_f(x, \hat{y}_i) + (1-\alpha)\ell_g(x, \hat{y}_i)$, where ℓ_f and ℓ_g are defined in Eqn.(4)
3. Return $\hat{y}_{i^*}$ as the translation of x.

Since the primal task and the dual one are of equal position in machine translation, the dual inference algorithm for the dual task can be defined in the same way, and we only show the algorithm for the primal task due to the limited space.

3.2 Experimental Setup

In this paper, we conduct empirical studies on the translation between English↔German (briefly, En↔De) and that between English↔French (briefly, En↔Fr). To train the primal and the dual translation models for each language pair, we rely on the same datasets as those used in [Jean *et al.*, 2015; He *et al.*, 2016a]. To be more concrete, the bilingual training data are part of WMT'14, consisting of 4.5M for En↔De and 12M for En↔Fr sentences pairs, respectively. We concatenate *newstest2012* and *newstest2013* as the validation sets and use *newstest2014* as the test sets[2]. We employ two training methods, as described below, to obtain the NMT models:

- **RNNSearch** represents the standard sequence-to-sequence training method as introduced in [Bahdanau *et al.*, 2015; Jean *et al.*, 2015].
- **dual-NMT** denotes the dual learning as that proposed in [He *et al.*, 2016a].

The translation qualities are evaluated by tokenized case-sensitive BLEU [Papineni *et al.*, 2002] scores calculated by[3] *multi-bleu.pl*. The larger BLEU is, the better the translation quality is.

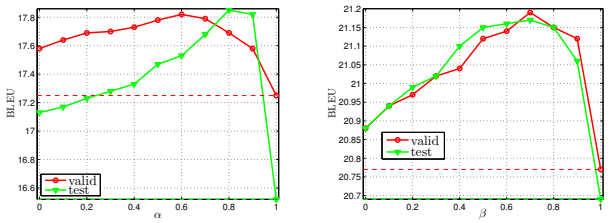
3.3 Results of NMT

Figure 1 shows the BLEU scores by dual inference, under varying parameter settings, of En↔De translation models trained by RNNSearch. (Figures for En↔Fr are omitted due to space limitation.) In this figure, the red curves represent the results on the validation sets while the green ones denote those on the test sets. From this figure, we can see that dual inference can outperform the standard inference rule, i.e., $\alpha = 1$ or $\beta = 1$, in a wide value range of α and β.

Table 1 compares the BLEU scores by various inference methods of two different types of translation models. In this table, column "Standard" represents the results obtained by standard inference rule, where α and β are set as 1, column "Dual" denotes the results of dual inference, and column "Δ" indicates the increase in BLEU. From this table, we can observe that the dual inference can give rise to significant improvement on most of tasks, especially on the En→De. Furthermore, we can also find that the improvements on dual-NMT models are smaller compared to those without dual-training, due to that dual-NMT already integrate duality knowledge into the translation model.

[2]Data from *http://www.statmt.org/wmt14/translation-task.html*

[3]https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl



(a) En→De (b) De→En

Figure 1: BLEU scores by dual inference of En↔De models trained by RNNSearch for varying settings of α / β

Model	Task	Standard	Dual	Δ
RNNSearch	En→De	16.54	17.53	0.99
	De→En	20.69	21.17	0.48
	En→Fr	29.92	30.45	0.53
	Fr→En	27.49	27.86	0.37
dual-NMT	En→De	18.49	18.96	0.47
	De→En	22.14	22.37	0.23
	En→Fr	32.06	32.26	0.20
	Fr→En	29.78	30.34	0.56

Table 1: BLEU scores with varying inference models and varying training methods for machine translation

One may notice that in Figure 1(a), the BLEU scores at $\alpha = 0$ (corresponding to the dual model) outperform those at $\alpha = 1$ (corresponding to the primal model). This does not mean that the dual model is better than the primal model for the primal task. The reason is that, when $\alpha = 0$, we first use the primal model to generate several candidate translations (through beam search), and then re-rank the candidates using the dual model. That is, $\alpha = 0$ actually uses both the primal and dual models. Therefore, it is possible that $\alpha = 0$ outperforms $\alpha = 1$ (using the primal model only).

Case study

In the following, we leverage a specific example in De→En translation to illustrate how dual inference can improve the performance of NMT. Let y denote the German sentence, x_1 and x_2 denote 2 out of 12 candidates generated by beam search, x^* represent the ground-truth translation. Then, let f denote the En→De model and g denote the De→En model. The examples and the loss of corresponding sentence pairs are shown in Table 2.

y	*Das System schafft die Gefahr , die es bekämpft .*
x_1	*The system is in danger of being tackled .*
x_2	*The system creates the risk that it is fighting against .*
x^*	*The system creates the threat that it is fighting against .*
$\ell_g(x_1, y) = 5.15;$ $\ell_g(x_2, y) = 7.92;$ $\ell_f(x_1, y) = 34.05;$ $\ell_f(x_2, y) = 12.28;$	

Table 2: Examples & loss in the case study

After using dual inference, the losses will be

$$\beta\ell_g(x_1, y) + (1-\beta)\ell_f(x_1, y) = 13.82, \text{where } \beta = 0.7;$$
$$\beta\ell_g(x_2, y) + (1-\beta)\ell_f(x_2, y) = 9.23, \text{where } \beta = 0.7;$$

From these results, we can find that, standard inference rule prefers x_1 which is, however, quite a bad translation; in contrast, dual inference, by considering the probabilities of both directions, tends to boost a better translation.

Discussions about NMT with reconstruction

An intuitive explanation of the magic of dual inference lies in that it leverages the reconstruction ability, which indicates that a matched pair (x, y) should not only get smaller loss for primal task, but also for dual task. The similar idea is also used for NMT by [Tu *et al.*, 2017]. In that work, for any sentence pair (x, y), the training objective is re-designed to minimize $\ell_{\text{rec}} = \log \mathbb{P}(y|x, \theta) + \log \mathbb{P}(x|s, \gamma)$, where θ and γ are the parameters to learn, and s is a hidden representation related to x and y. The inference phase in that work consists of first using model θ to generate K candidates and then selecting the sentence that can minimize ℓ_{rec}. The reconstructor γ is only served for θ without capability to make translations by itself. On the contrary, in the dual inference framework, the translation models of two directions can both make translations and can be trained individually. Furthermore, dual inference can improve the performance for both f and g, which makes it quite different from the work in [Tu *et al.*, 2017].

Model	Standard	Dual	Δ
$\mathcal{M}_{\text{w2v}}$	10.10	8.31	1.79
$\mathcal{M}_{\text{LM}}$	7.76	7.15	0.61

Table 3: Comparison of classification error (%)

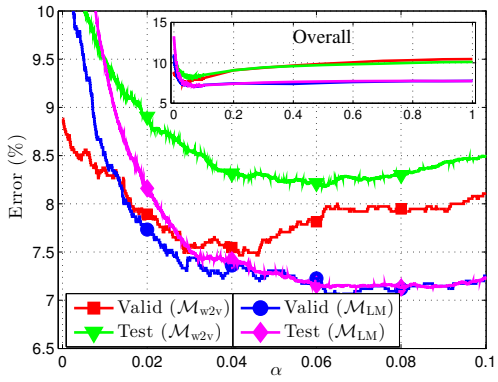


Figure 2: IMDB classification error with using dual inference

4 Sentiment Analysis

Although not very explicit, there exists duality in the domain of sentiment analysis. Particularly, sentiment classification and sentence generation comprise two AI tasks in the dual form. On one hand, the goal of the primal task, i.e. sentiment classification, is to classify the polarity of given natural language sentences. The dual task, on the other hand, aims at automatically generating sentences with the certain polarity class of sentiment.

The widely-used approach for sentiment classification usually takes advantage of LSTM based RNN [Dai and Le, 2015], in which a sentence is encoded word by word such that it is eventually transformed into a hidden representation, then the hidden representation is used as the input to a fully-connected neural networks to predict a polarity label.

Meanwhile, a typical sentence generation approach, inspired by [Wang and Cho, 2015], is designed as follows: two sentiment labels are first projected into a certain size of sentiment embedding, which will then become the input of LSTM cells. More formally, let x denote the sentence (with the t-th word denoted as x_t) and y denote the sentiment label. The LSTM cell takes $W^e_w E_w x_{t-1} + W^e_s E_s y$ and h_{t-1} as inputs and outputs h_t, where E represents the embedding matrix, W denotes the connection between the embedding layer and LSTM cells, h_{t-1} and h_t denote the hidden states at timestep $t-1$ and t. Sentences will be generated word-by-word and the probability that a specific word x_t is generated is proportional to $\exp(W^d_w E_w x_{t-1} + W^d_s E_s y + W_h h_t)$. Note that W's and the E's are the parameters to learn.

4.1 Dual Inference for Sentiment Analysis

Let f denote the sentiment classification model and g denote the sentence generation model. Let ℓ_f and ℓ_g represent the negative log-likelihood. Still,

$$\ell_f(x, y) = -\log P(y|x; f), \ell_g(x, y) = -\log P(x|y; g), \quad (5)$$

where x and y are sentence and sentiment label respectively. For the sentiment classification task, we can replace the ℓ_f and ℓ_g in Eqn.(1) with those in (5). For the sentence generation task, we still follow the similar dual inference approach as in NMT, by using which K candidates will be generated first and the best one will be selected jointly by the sentiment classifier and sentence generator.

4.2 Experimental Setup

To conduct our empirical studies, we use the IMDB movie review dataset [Maas *et al.*, 2011][4], which consists of 25K training sentences and 25K test sentences. We split 3750 sentences from the training as the validation set. During learning, we set 500 dimension embedding size and 1024 dimension hidden node size. As proposed by [Dai and Le, 2015], there are two ways to initialize the classifiers:

(1) $\mathcal{M}_{\text{w2v}}$: The embedding matrix is initialized by a pre-trained word embedding matrix;

(2) $\mathcal{M}_{\text{LM}}$: The classifier is initialized by a pre-trained language model.

Note that when training $\mathcal{M}_{\text{LM}}$, we can leverage a great amount of unlabeled data, by using which can drastically reduce the error of sentiment classification.

4.3 Results of Sentiment Classification

Table 3 compares the accuracy of sentiment classification by using standard inference rule against using dual inference. From this table, we can find that dual inference can result in better performance for both classifiers with separate initialization methods.

Moreover, Figure 2 demonstrates the validation/test curves of the sentiment classifier with using dual inference. The small figure is the valid/test curve over a wider range $[0, 1]$, while we zoom in the $[0, 0.1]$ region in the larger figure. From

[4] http://ai.stanford.edu/~amaas/data/sentiment/

this figure, we can find that the dual inference can give rise to better classification accuracy in a wide ranges than standard inference rules.

4.4 Results of Sentence Generation

To illustrate the effectiveness of dual inference on boosting the performance of sentence generation, we show some example sentences generated based on standard inference rules and dual inference respectively in Table 4.

From this table, we can find that standard inference rules tend to generate sentences of high-frequency in the training data, such as "the plot is...", as well as those with strong sentiment but using quite common style phrases, such as "it was so bad". We hypothesize the main reason of which is that the sentence generation model g follows the language modeling approach. Meanwhile, we highlight those sentences boosted by dual task, i.e. the classifier f, in Table 4. We can observe that, in order to achieve lower classification error in dual inference, the classifier f not only prefers to phrases with strong sentiment like "i love this movie" but also favors those covering more delicate aspects and modes, such as "I give it 2 out of 10". All these examples have obviously illustrated the strength of dual inference in improving sentence generation.

[Standard] *Positive*	*this movie is one of the funniest movies i have ever seen. the acting is great, the plot is simple. it is one of the best movies i've seen in a long time.*
[Dual] *Positive*	***i love this movie. i watched it over and over again*** *and i have to say that it is one of the best movies i've seen in a long time. the plot is simple, the acting is great. if you are looking for a good movie, go to see this movie.*
[Standard] *Negative*	*when i first saw this movie, i thought it was going to be funny, but it didn't. it was so bad, i didn't think it was going to be funny. the only thing i can say about this movie is that it is so bad that it's not funny.*
[Dual] *Negative*	***i give it 2 out of 10*** *because , it's the worst movie i have ever seen . the only thing i can say about this movie is that it is so bad that* ***it makes no sense at all . don't waste your time .***

Table 4: Sentences generated by standard / dual inference

5 Image Processing

An important duality relationship bridges between two major tasks, i.e. image classification and image generation. In particular, image classification aims at predicting the semantic category label of an image. Most recently, image classification modeling, driven by deep learning, has attracted worldwide research efforts, and some state-of-the-art work includes NIN [Lin *et al.*, 2014], DSN [Lee *et al.*, 2015], ResNet [He *et al.*, 2016b] and WRN [Zagoruyko and Komodakis, 2016]. On the other hand, the corresponding dual task, image generation, targets automatically generating images based on category labels. More formally, let x denote an N-pixel image with the ith pixel x_i, y denote the category label. The inherent problem of an image generator is indeed to model the image distribution, which is $\prod_{i=1}^{N} P(x_i|x_{<i}, y)$. PixelCNN++ [Salimans *et al.*, 2017] is a representative model for image generation and achieves state-of-the-art performance. Note that, the detailed dual inference approaches for above two tasks are similar to those for sentiment analysis tasks as described above. Therefore, we skip the details of them due to the limited space.

5.1 Experimental Setup

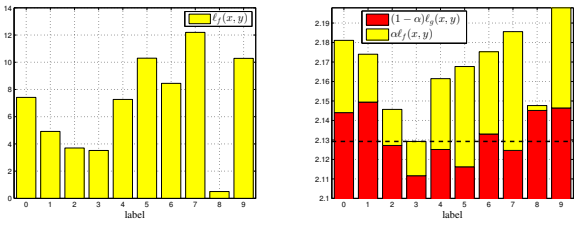
In this paper, we use CIFAR-10 dataset in the experiments regarding image processing. We split 5k images away from the training data as the validation set. Two state-of-the-art classifiers, including the 110 layer ResNet [He *et al.*, 2016b] and the WRN-40-10 proposed in [Zagoruyko and Komodakis, 2016], are used to verify the effects of dual inference on their performance. Meanwhile, we choose PixelCNN++ [Salimans *et al.*, 2017] as the approach to model image generation. Note that PixelCNN++ enables to take a one-hot label as the input and output an image with respect to the given category. Details of above three models can be referred in the corresponding literature.

5.2 Results of Image Classification

Table 5 shows the error rate of two image classifiers with varying inference methods. From this table, we can find that, despite that ResNet-110 and WRN-40-10 are strong models in image classification as they have reached quite low errors, dual inference can still improve their performances, which indeed emphasizes the strength of dual inference.

Model	Standard	Dual	Δ
ResNet-110	6.46	5.98	0.48
WRN-40-10	3.86	3.68	0.18

Table 5: Error rates (%) of with varying inference methods.



(a) Standard Inference (b) Dual Inference

Figure 3: Classification result comparison between different inference methods on an example image. In the legends, x refers to the image, and y refers to the label.

Figure 3 compares the classification result between different inference methods on an example image. The selected input image in this experiment belongs to category 3. From this figure, we can find that, by using standard inference rules, this image will be misclassified into category 8. However, dual inference can make the right decision since it considers the

conditional probability of the image conditioned on different labels during inference process.

5.3 Results of Image Generation

Figure 4 compares the results of image generation by using standard inference rule and that by applying dual inference. Each row in this figure corresponds to a specific category of generated images. The left five columns contain the results by using standard inference rule, which picks the top-5 images with minimal test negative log-likelihood, while the right five includes the results by applying dual inference.

Figure 4: Image generated by standard / dual inference

From this figure, we can find that, dual inference tends to generate much clearer images, especially for the category of plane, dog, and ship. An interesting observation is that the standard inference (i.e., the vanilla inference) often prefers to generating blurred images, while the dual inference is capable of filtering out such images.

6 Discussions

We observe that the performance of dual inference does not highly depend on the model structures of the two dual tasks: (1) In NMT, the network structures of f and g are the same, i.e., bidirectional GRUs; (2) In sentiment analysis, the classifier (i.e., LSTM+sigmoid) and the sentence generator (i.e., LSTM+softmax) share some basic structures; (3) In image processing, the classifier (i.e., ResNet) and the image generator (i.e., PixelCNN++) are quite different. Dual inference works well on all these three situations.

The experiments in the above three sections show that dual inference can give rise to significant improvements for both of dual tasks. In the remaining part of this section, we provide some simple theoretical discussions for dual inference.

Let φ_f denote $1-\ell_f$ and let φ_g denote $1-\ell_g$. We make two assumptions for theoretical analysis: (i) $\mathcal{Y} = \{1, 2, \cdots, c\}$ where $c \geq 2$; (ii) For any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\varphi_f(x, y) \in [0, 1]$ and $\varphi_g(x, y) \in [0, 1]$. We further define φ as $\alpha\varphi_f + (1-\alpha)\varphi_g$. The margin $\rho(x, y)$ is defined as $\varphi(x, y) - \max_{y' \neq y} \varphi(x, y')$. Thus, φ misclassifies (x, y) if $\rho_{(x,y)} \leq 0$. We assume that training and test samples are drawn i.i.d. according to some distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and denote by $\mathcal{S} = ((x_1, y_1), ..., (x_m, y_m))$ a training set of size m drawn i.i.d according to $\mathcal{D}$. For any $\rho > 0$, the generalization error $R(\varphi)$ and its empirical margin error $\hat{R}_{S,\rho}$ are defined as follows:

$$\begin{aligned} R(\varphi) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}\{\rho_\varphi(x, y) \leq 0\}]; \\ \hat{R}_{S,\rho} &= (1/m)\textstyle\sum_{i=1}^{m}[\mathbf{1}\{\rho_\varphi(x_i, y_i) \leq \rho\}]. \end{aligned} \tag{6}$$

Following [Kuznetsov *et al.*, 2014], for any family of hypothesis G mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, we define $\Pi_1(G)$ as

$$\Pi_1(G) = \{x \mapsto h(x, y) : y \in \mathcal{Y}, h \in G\}. \tag{7}$$

Let $\mathcal{H}_f$ and $\mathcal{H}_g$ denote the two hypothesis spaces of φ_f and φ_g. Let $\mathfrak{R}_m(\cdot)$ denote the Rademacher complexity [Bartlett and Mendelson, 2002]. We leverage the Theorem 1 in [Kuznetsov *et al.*, 2014], further optimize it under our settings, and obtain the following theorem:

Theorem 1. *Fix $\rho > 0$, for any $\delta > 0$, with probability at least $1-\delta$ over the choice of a training set S of size m drawn i.i.d. according to $\mathcal{D}$, the following inequality holds:*

$$R(\varphi) \leq \hat{R}_{S,\rho}(\varphi) + \frac{8c}{\rho}\Big(\alpha\mathfrak{R}_m(\Pi_1(\mathcal{H}_f)) + (1-\alpha)\mathfrak{R}_m(\Pi_1(\mathcal{H}_g))\Big)$$
$$+\frac{1}{\rho}\sqrt{\frac{2}{m}} + \sqrt{\frac{1}{2m}\log\Big(\lceil\frac{4}{\rho^2}\log(\frac{mc^2\rho^2}{2})\rceil + 1\Big) + \frac{1}{2m}\log\frac{1}{\delta}}.$$

The above theorem shows that the generalization bound of dual inference is related to the Rademacher complexities of both $\mathcal{H}_f$ and $\mathcal{H}_g$, and the hyper-parameter α also plays an important role in the bound.

7 Conclusion

In this work, we have proposed a general framework of *dual inference* which enables dual tasks to boost each other in the inference stage according to the valuable structure duality widely-existed in AI applications. Empirical studies on three pairs of specific dual tasks have revealed that dual inference can efficiently improve the performance of both tasks.

In the future, we will explore better dual inference rules, which can more efficiently explore the power of two dual models. Moreover, we will enrich the theoretical analysis for dual inference. Finally, we will investigate how to feed the signal provided by the dual inference back to dual learning (i.e., dual training) so as to induce more powerful learning paradigms.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61371192), the Key Laboratory Foundation of the Chinese Academy of Sciences (CXJJ-17S044) and the Fundamental Research Funds for the Central Universities (WK2100330002).

References

[Amodei *et al.*, 2016] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *33rd International Conference on Machine Learning*, 2016.

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[Bartlett and Mendelson, 2002] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(Nov):463–482, 2002.

[Dai and Le, 2015] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*, pages 3079–3087, 2015.

[Graves *et al.*, 2013] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[He *et al.*, 2016a] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances In Neural Information Processing Systems*, pages 820–828, 2016.

[He *et al.*, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[He *et al.*, 2016c] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[Jean *et al.*, 2015] Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *ACL*, 2015.

[Kuznetsov *et al.*, 2014] Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Multi-class deep boosting. In *Advances in Neural Information Processing Systems*, pages 2501–2509, 2014.

[Lee *et al.*, 2015] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, volume 2, page 6, 2015.

[Lin *et al.*, 2014] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *International Conference on Learning Representations*, 2014.

[Maas *et al.*, 2011] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *ACL*, pages 142–150, 2011.

[Oord *et al.*, 2016] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[Opitz and Maclin, 1999] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *JAIR*, 11:169–198, 1999.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[Salimans *et al.*, 2017] Tim Salimans, Andrej Karpathy, Xi Chen, Diederik P. Kingma, and Yaroslav Bulatov. Pixelcnn++: A pixelcnn implementation with discretized logistic mixture likelihood and other modifications. In *International Conference on Learning Representations*, 2017.

[Shen and Liu, 2016] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. *arXiv preprint arXiv:1612.05363*, 2016.

[Tu *et al.*, 2017] Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. Neural machine translation with reconstruction. In *AAAI*, 2017.

[van den Oord *et al.*, 2016a] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pages 4790–4798, 2016.

[van den Oord *et al.*, 2016b] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *33rd International Conference on Machine Learning*, 2016.

[Wang and Cho, 2015] Tian Wang and Kyunghyun Cho. Larger-context language modelling. *arXiv preprint arXiv:1511.03729*, 2015.

[Wu *et al.*, 2016] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[Xia *et al.*, 2017] Yingce Xia, Tao Qin, Wei Chen, Bian Jiang, Nenghai Yu, and Tie-Yan Liu. Dual supervised learning. In *ICML*, 2017.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *27th British Machine Vision Conference*, 2016.