Layer-Wise Cross-View Decoding for Sequence-to-Sequence Learning

Fenglin Liu[1]*, Xuancheng Ren[2]*, Guangxiang Zhao[3], Xu Sun[2,3]
[1]School of ECE, Peking University
[2]MOE Key Laboratory of Computational Linguistics, School of EECS, Peking University
[3]Center for Data Science, Peking University
{fenglinliu98, renxc, zhaoguangxiang, xusun}@pku.edu.cn

Abstract

In sequence-to-sequence learning, the decoder relies on the attention mechanism to efficiently extract information from the encoder. While it is common practice to draw information from only the last encoder layer, recent work has proposed to use representations from different encoder layers for diversified levels of information. Nonetheless, the decoder still obtains only a single view of the source sequences, which might lead to insufficient training of the encoder layer stack due to the *hierarchy bypassing* problem. In this work, we propose *layer-wise cross-view decoding*, where for each decoder layer, together with the representations from the last encoder layer, which serve as a global view, those from other encoder layers are supplemented for a stereoscopic view of the source sequences. Systematic experiments show that we successfully address the hierarchy bypassing problem and substantially improve the performance of sequence-to-sequence learning with deep representations on diverse tasks.

1 Introduction

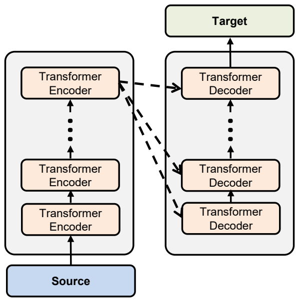
In recent years, encoder-decoder based models (Bahdanau et al., 2015; Vaswani et al., 2017; Xu et al., 2015) have become the fundamental instrument for sequence-to-sequence learning, especially in tasks that involves natural language. The attention mechanism (Luong et al., 2015; Bahdanau et al., 2015) proves essential for the encoder-decoder based models to efficiently draw useful source information from the encoder.

As illustrated in Figure 1, for those representations output by different encoder layers, the common practice for the decoder is to draw information only from the last encoder layer, which is regarded as a global and comprehensive view of the source sequence but short of precise and finer details. The impact of such architectural choice is demonstrated in Table 1. For this German-to-English translation example, the target sentence generated by the system reveals two common shortcomings in neural sequence-to-sequence models: 1) the generated text is unfaithful to the source (e.g., *three* compared to *dreizehn "thirteen"*); and 2) repeated texts are generated (e.g., *at the age of three*) (See et al., 2017). They can both be attributed to the lack of detailed and accurate information.

Several studies have tried to narrow the information gap by introducing deep representations from the encoder layers. For example, fusing the representations from different encoder layers (Dou et al., 2018; Wang et al., 2018, 2019) or providing representations from different encoder layers to different decoder layers (Bapna et al., 2018; Domhan, 2018; He et al., 2018). However, mixed conclusions have been reached by such work, either with slight improvements (Bapna et al., 2018) or obvious degradation (Domhan, 2018; He et al., 2018). In this work, we identify the *hierarchy bypassing* problem that affects all previous related efforts, hindering the efficient training of the encoder layer stack and weakening the hierarchy and diversity of representations from different encoder layers, which is the premise for sequence-to-sequence learning with deep representations.

We propose a novel approach called *layer-wise cross-view decoding*, where for each decoder layer, together with the global view from the last encoder layer, another purposeful view of the source sequence is also supplemented (*soft integration*). Moreover, the layer-wise cross-view training is devised to continue the conventional training so that the representational ability of the multi-layer encoder is directly inherited (*continued learning*). The effect of our approach is systematically investigated on typical strategies

* Equal contribution.



(a) Conventional Transformer

Figure 1: Illustration of the Transformer (Vaswani et al., 2017).

Source	aber ich hatte das große glück , ihn sehr früh , mit *dreizehn* jahren , kennenzulernen und so war ich schon zu meiner schulzeit auf seinen kursen .
Target	but i had the good fortune to meet him at a very young age , when i was *thirteen* , and so i always attended his courses while i was at school .
Base	but i was very lucky to meet him *at the age of three at the age of three* , and so at the time , i was on his classes .
Ours	but i was very lucky to meet him at the age of *thirteen*, and so at the time while i was at school , i always on his courses .

Table 1: Examples of the target sentence translated by different methods. Base stands for the conventional Transformer-Base (Vaswani et al., 2017) which generates unfaithful and repeated text.

of routing source representations (see Figure 2), which shows that the proposed approach successfully addresses the hierarchy bypassing problem, requires minimal parameter increase, and improves the performance of sequence-to-sequence learning with deep representations.

Specifically, through our experiments on the machine translation with two strong baselines, i.e., Transformer (Vaswani et al., 2017) and DynamicConv (Dou et al., 2019), which are the previous state-of-the-art models, we find that one of the cross-view variant, i.e., granularity consistent attention (GCA) (see Figure 2 (a)), promotes the performance substantially. We speculate the reason is that the GCA builds connections between the corresponding layers in the encoder and the decoder, so that the first decoder layer pays attention to the global information, i.e., *coarse-grained* representations, of the source sequence, which is instrumental in language modeling, while the last decoder layer pays attention to the *fine-grained* representations of the source sequence, which is helpful to generate words that are more precise. Notably, a similar connection pattern without the global view is used in computer vision for biomedical image segmentation with success (Ronneberger et al., 2015). It is also considered by Domhan (2018) in natural language processing but with opposite results, i.e., performance degradation. The reason is that a *direct transfer* of such pattern is not viable for sequence-to-sequence learning and further considerations are required for successful training as we show.

Overall, our contributions are as follows:

- We propose layer-wise cross-view decoding to efficiently use the diverse-grained representations from the multi-layer encoder. For a decoder layer, together with the global view from the last encoder layer, another purposeful view is supplemented for a stereoscopic view of the source sequences.
- The global view of the source sequence in cross-view decoding mitigates the hierarchy bypassing problem and the proposed continued learning transfers the representational ability of the conventional training seamlessly into cross-view training.
- Experiments and analyses demonstrate that our approach works for representative models and tasks in natural language processing, i.e., Transformer and DynamicConv on the machine translation task, consistently and substantially improves the performance of sequence-to-sequence learning with deep representations.

2 Approach

We first briefly review the conventional encoder-decoder model and then introduce the proposed cross-view decoding realized on the Transformer model (Vaswani et al., 2017).

2.1 Conventional Single-View Decoding

In the conventional encoder-decoder model, the encoder encodes the source sequence S_0 with N stacked encoder layers, defined as

$$\mathrm{S}_i = f_{\text{encoder}}\left(\mathrm{S}_{i-1}\right). \tag{1}$$

By repeating the same process for N times, the encoder outputs the representations $\mathbf{S} = \{\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_N\}$ of the source sequence. Transformer (Vaswani et al., 2017) proposes Multi-Head Attention (MHA) and Feed-Forward Network (FFN) to implement the encoder layers:

$$f_{\text{encoder}}\left(\mathrm{x}\right) = \mathrm{FFN}\left(\mathrm{MHA}\left(\mathrm{x,x,x}\right)\right). \tag{2}$$

Then, the conventional decoder generates the target sequence T_N with N stacked decoder layers. Typically, only the representations from the final encoder layer S_N serve as the bridge between the source and the target, via the attention mechanism. Unlike the encoder, the decoder has an extra objective that handles the intermediate representations from the encoder, defined as

$$\mathrm{T}_i = f_{\text{decoder}}\left(\mathrm{T}_{i-1}, \mathrm{S}_N\right). \tag{3}$$

For Transformer, the decoder achieves the extra objective by an extra multi-head attention:

$$f_{\text{decoder}}\left(\mathrm{x,y}\right) = \mathrm{FFN}(\mathrm{MHA}(\mathrm{MHA}(\mathrm{x,x,x}), \mathrm{y}, \mathrm{y})), \tag{4}$$

where x and y stand for the query and the key/value, respectively.

Figure 1 shows a conceptual illustration of the conventional Transformer model, focusing on the attention patterns. As we can see, it only makes use of the coarse-grained representations at the last encoder layer in decoding, which may fail to capture accurate source information that is better framed by representations from first encoder layers, i.e., fine-grained representations. The effect is more obvious for longer source text (see Sec. 4.4), because it is difficult for the model to find the correct source information in the coarse-grained representations, which can make the generated target sentence unfaithful to the original sentence.

2.2 Hierarchy Bypassing

Existing works (Domhan, 2018; Bapna et al., 2018) have tried to narrow the information gap by substituting the representations from the last encoder layer by representations from other encoder layers. Taking the generality into consideration, this kind of decoding can be formalized as

$$\mathrm{T}_i = f_{\text{decoder}}\left(\mathrm{T}_{i-1}, g_i(\mathbf{S})\right), \tag{5}$$

where $\mathbf{S} = \{\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_N\}$ represents the source representations of different granularity from the N encoder layers. The function $g_i(\cdot)$ stands for the layer-wise strategy. For example, supposing the decoding is conducted in a coarse-to-fine manner, that is, the first layer need high-level information for language planning and the last layer need low-level information for detail realization, one may hypothesize that the decoder layers should draw information from the encoder layers inversely to their positions, i.e., $\mathrm{T}_i = f_{\text{decoder}}\left(\mathrm{T}_{i-1}, \mathrm{S}_{N-i+1}\right)$. As a result, the first/last decoder layer adopts the information from the last/first encoder layer.

However, such a *direct replacement* may be problematic, leading to the *hierarchy bypassing* problem. To explain the problem, take the previous inversely-connected model as an example. Intuitively, there is a direct connection between the first encoder layer and the last decoder layer shorter than any other connection between the encoder layers and the decoder layers, leading to the problem that the gradient information can pass through the remaining layers of the encoder directly to the first encoder layer. As a result, the representational ability of the encoder is impaired, which in turn leads to performance degradation.

More specifically, the phenomenon can be explained by the following analysis. We denote the loss function in the conventional model as $\mathcal{E}$ and the $\mathbf{e}_i$ and $\mathbf{d}_i$ denote the parameters of the i_{th} encoder layer and the i_{th} decoder layer, respectively. We take the most significant N_{th} encoder layer as an example. From the chain rule of backpropagation (LeCun et al., 1989), we can get the gradient of the N_{th} encoder layer in the conventional model and similarly for models with the replacement:

$$\left(\frac{\partial \mathcal{E}}{\partial \mathbf{e}_N}\right)_{\text{conventional}} = \sum_{i=1}^{N} \frac{\partial \mathcal{E}}{\partial \mathbf{d}_i}\frac{\partial \mathbf{d}_i}{\partial \mathbf{e}_i}, \tag{6}$$

$$\left(\frac{\partial \mathcal{E}}{\partial \mathbf{e}_N}\right)_{\text{replaced}} = \frac{\partial \mathcal{E}}{\partial \mathbf{d}_1}\frac{\partial \mathbf{d}_1}{\partial \mathbf{e}_i}. \tag{7}$$

As we can see, the gradient information of $\left(\partial\mathcal{E}/\partial\mathbf{e}_N\right)_{\text{conventional}}$ is significantly richer than that of $\left(\partial\mathcal{E}/\partial\mathbf{e}_N\right)_{\text{replaced}}$. It means the top encoder layer of the conventional model is better suited to the final output and the top encoder layer of the inversely-connected model is shorted in the sense that its representational ability is not adequately expressed and the representation hierarchy is bypassed in this connection pattern. This negative effect is confirmed in our experiments (see Sec. 4.2) and shown by previous efforts (Domhan, 2018).

2.3 Cross-View Decoding

To construct a proper way of using deep representations in sequence-to-sequence learning and encourage the decoder to make full use of the source sequence information from the global and local perspectives, we enhance the decoding process with multiple views of the source sequences with different granularity, such that the expressive power of the model could be fully taken ad-

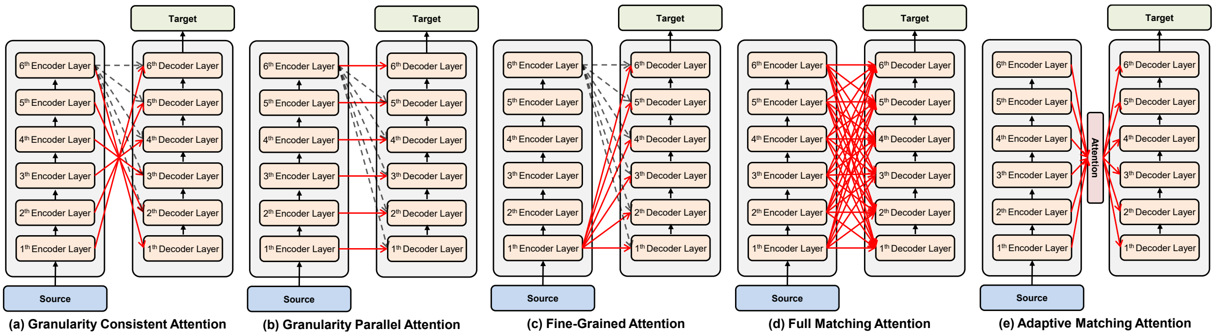


Figure 2: We present the proposal on Transformer with various strategies for routing the source representations: (a) Granularity Consistent Attention; (b) Granularity Parallel Attention; (c) Fine-Grained Attention; (d) Full Matching Attention; (e) Adaptive Matching Attention. The dashed lines represent the original attention to the last encoder layer and we omit them in (e) for clarity.

vantage of. Particularly, we propose to conduct layer-wise cross-view decoding (see Figure 2), for each decoder layer, a different combination of the source views is considered, encouraging the decoder to make adequate, efficient use of the source sequence information from the global and local perspectives. Different from existing work in using deep representations, the focus of our proposal is on crossed views, that is, two different views of the source information is provided at the same time for a decoder layer.

2.3.1 Soft Integration

To properly train the encoder stack and address the hierarchy bypassing problem, the final encoder layer needs comprehensive error signals. The simple way to achieve that is to incorporate the additional source views on top of the representations of the final encoder layer. The decoder of the final model is updated as follows:

$$\mathrm{T}_i = f_{\text{decoder}}\left(\mathrm{T}_{i-1}, \mathrm{LN}(g_i(\mathbf{S}) + \mathrm{S}_N)\right), \qquad (8)$$

where LN stands for Layer Normalization (Ba et al., 2016). The layer normalization is needed to keep the scale of the representations for stability in deep neural networks and allow the soft integration of two different source views into the model.

2.3.2 Continued Learning

For the conventional encoder-decoder model, once trained, its encoder should be capable of describing the source sequence in different granularity. Such ability is crucial to the cross-view decoding. However, while applying soft integration alone maintains the richness of gradient sources, the last encoder layers could still underfit due to the co-adaption of the first encoder layer. Due to the augmentative nature of the proposal, the model structure can be extended seamlessly upon the original model. Thus, we propose to conduct continued training based on the trained conventional model. In practice, the conventional model is first trained normally, and then, our cross-view decoding is applied to further fine-tune the attention structure. Although the structural changes of attention structures pose challenges for continued learning, the model can adapt successfully in reality due to the inherent augmentative design of approach.

2.3.3 Cross-View Decoding Strategies

In the multi-layer setting, each encoder layer adds a level of abstraction and is believed to produce representations that are more and more coarse-grained describing global context. However, in the decoder, the process is more complex, since each decoder layer receives both the source information and target information, which opens up the question how to properly incorporate different source view, i.e., the function $g_i(\cdot)$ in Eq. (5) and Eq. (8). Extending the related literature (Domhan, 2018; Bapna et al., 2018), we systematically investigate diverse strategies of routing the source views.

Granularity Consistent Attention (GCA): Figure 2 (a) illustrates the connection pattern. In this strategy, each decoding step is seen as a realization process, where an abstract, coarse-grained idea turns into concrete, fine-grained words gradually through the layers. Hence, the granularity consistent attention keeps the granularity of source views in order with the decoder layers. Therefore, the GCA can be defined as $g_i(\mathbf{S}) = \mathrm{S}_{N-i+1}$.

Granularity Parallel Attention (GPA): Opposite to the GCA strategy that keeps the granularity order in attention, the source granularity order in attention is reversed. In this strategy, the decoding is regarded the same as the encoding, where each layer abstracts together the source and the target sequences, hence GPA, defined as $g_i(\mathbf{S}) = \mathrm{S}_i$.

Fine-Grained Attention (FGA): Opposite to the conventional Transformer (see Figure 1) that only draws information from the final encoder layer, i.e., coarse-grained representations, we experiment with adding only the information of the first layer of encoder, i.e., fine-grained representations. The strategy is named as the Fine-Grained Attention (FGA), which is defined as $g_i(\mathbf{S}) = \mathrm{S}_1$.

Full Matching Attention (FMA): We also consider routing all source views into each decoder layer, i.e., Full Matching Attention (FMA). In implementation, we combine them by linear transformations, and define the FMA strategy as $g_i(\mathbf{S}) = \sum_{j=1}^{N} (W_{ij}\mathrm{S}_j + b_{ij})$. Note that, it is layer-wise in that each decoder layer uses different linear transformation parameterized by W_{ij} and b_{ij}.

Adaptive Matching Attention (AMA): Finally, we apply an attention mechanism to help inject information of various granularity levels adaptively, i.e., Adaptive Matching Attention (AMA). We build an independent vector for each encoder layer to predict the attention weight. We define the AMA as $g_i(\mathbf{S}) = \sum_{j=1}^{N} \alpha_{ij}\mathrm{S}_j$, where α_{ij} is computed by the attention mechanism and $\sum_{j=1}^{N} \alpha_{ij} = 1$.

It should be noted that, in our experiments, we find that all variants are viable in practice and can promote the performance substantially. However, our experiments also show that GCA performs the best, since it could be more in line with the characteristics of the multi-layer encoding and decoding process, while other strategies may route unneeded information leading to learning difficulties.

3 Experiments

Our main experiments focus on machine translation (Bahdanau et al., 2015; Vaswani et al., 2017), which is arguably the most important sequence-to-sequence learning task in natural language processing. We report results using the granularity consistent attention (GCA) for cross-view decoding, which is the best performed strategy in our preliminary experiments.

Implementation As the proposal only relates to the injection of different mix of source representations and is augmentative to the existing models, we keep the inner structure of the baselines untouched and preserve the original settings. For soft integration, we initialize the original attention structures with the parameters of the re-implemented baseline models. For continued learning, we further fine-tune the full model with the number of training steps used to re-implement the baseline model. We experiment on Transformer (Vaswani et al., 2017) and DynamicConv (Wu et al., 2019). Especially, DynamicConv (Wu et al., 2019) established a state-of-the-art in WMT EN-DE and EN-FR translation tasks in comparable settings, i.e., without much larger extra dataset for training as Edunov et al. (2018). For the re-implementation of the DynamicConv, we use the configuration of six blocks for both encoder and decoder. For detailed introduction to the implementation, please refer to Appendix A.

Automatic Evaluation Results We report results on three benchmarks, including two large WMT-2014 datasets, i.e., English-German (EN-DE) and English-French (EN-FR), and a small IWSLT-2014 dataset, i.e., German-English (DE-EN). As shown in Table 2, for three datasets, our approach outperforms all the baselines. Based on the Transformer, we promote the baseline by 0.9, 1.5 and 1.2 BLEU score for the EN-DE, EN-FR and DE-EN, respectively. More encouragingly, based on the DynamicConv, which is the previous state-of-the-art, our approach sets a new state-of-the-art performance on three datasets, achieving 29.8, 43.5 and 36.2 BLEU score on EN-DE, EN-FR and DE-EN respectively. We conduct significance test on DE-EN and the results are statistically significant (t-test with $p < 0.01$). The improvements on various datasets demonstrate the effectiveness of the proposal.

Human Evaluation Results We further conduct targeted human evaluation in terms of the faithfulness and fluency of the translated sentences. Specifically, we randomly selected 100 sentences from the IWSLT DE-EN dataset and recruited 5 annotators with sufficient language skills. Each annotator is asked to compare the performance of our approach with the baseline model as Transformer. As shown in Table 3, our approach enjoys obvious advantage in terms of the two aspects, meaning the translated sentences contain fewer factual errors and repeated segments.

Methods	EN-DE	EN-FR	DE-EN
Transformer (Vaswani et al., 2017)	28.4	41.0	-
Layer-wise Coordination (He et al., 2018)	29.0	-	35.1
Deep Representations (Dou et al., 2018)	29.2	-	-
Evolved Transformer (So et al., 2019)	29.8	41.3	-
DynamicConv (Wu et al., 2019)	29.7	43.2	35.2
Transformer (re-implementation)	29.0	41.1	34.7
w/ Cross-View Dec.	**29.9 (+0.9)**	42.6 (+1.5)	35.9 (+1.2)
DynamicConv (re-implementation)	29.3	43.1	35.2
w/ Cross-View Dec.	29.8 (+0.5)	**43.5 (+0.4)**	**36.2 (+1.0)**

Table 2: Experimental results of machine translation in terms of BLEU. As a whole, the proposed cross-view decoding with granularity consistent attention significantly improves the baselines.

Metric	Loss (%)	Tie (%)	Win (%)
Faithful	20	34	**46**
Fluent	18	45	**37**

Table 3: Results of human evaluation on the IWSLT DE-EN dataset in terms of faithfulness and fluency.

Methods	Transformer	DynamicConv
Baseline	34.7	35.2
Cross-View Dec. w/ GCA	**35.9**	**36.2**
Cross-View Dec. w/ GPA	35.2	35.5
Cross-View Dec. w/ FGA	35.0	35.5
Cross-View Dec. w/ FMA	35.4	35.7
Cross-View Dec. w/ AMA	35.4	35.8

Table 4: Results of different cross-view strategies on the IWSLT DE-EN dataset in terms of BLEU.

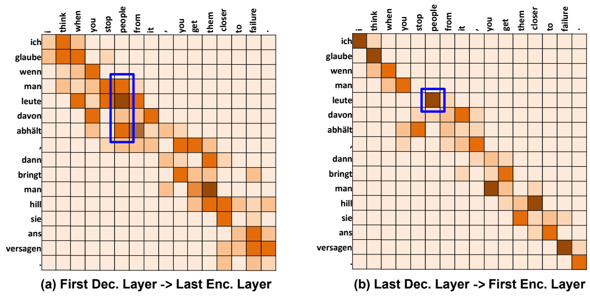


Figure 3: Visualization of the attention distribution of the proposed granularity consistent attention (GCA).

4 Analysis

In this section, we conduct analysis from different perspectives to better understand our approach. Unless otherwise specified, we use the Transformer model and the GCA strategy.

4.1 Cross-View Decoding Strategies

As reported in Table 4, all the strategies can improve the performance and GCA shows the best improvement. In comparison, GPA could not provide the fine-grained representations for the last decoder layer; FGA only provides a fixed level of source views; FMA combines all views and may introduce redundant noise, while AMA has a hard time in learning proper weights for different views. The results demonstrate the GCA strategy could be more in line with the characteristics of the multi-layer encoding and decoding process and the information is provided more efficiently.

This explanation can also be supported by the attention distribution instanced in Figure 3, where we show the granularity consistent attention on a sentence. For the first decoder layer attending to the last encoder layer, the attention distribution is fairly disperse, extracting information for language modeling. For example, as shown in the blue boxes in Figure 3, with the target input word *stop*, the source words with semantic-related roles are attended, i.e., *abhält* "*to stop*", *leute* "*people*" and *man* "*you*". In turn, for the last decoder layer attending to the first encoder layer, the attention is focused directly on the word to be generated, e.g., the source word *leute* when generating *people*.

4.2 Ablation Study

In this section, we conduct the ablation study to investigate the contribution of soft integration and continued learning, which are used to mitigate the hierarchy bypassing problem. The results are reported in Table 5. As we can see, directly replacing the conventional context attention is harmful to the accuracy similarly to existing work (Domhan, 2018). If the direct replacement is conducted with continued learning, that is, the model first learns with conventional context attention before GCA, we can observe an improvement that is substantial

Methods	BLEU
Baseline (Transformer)	34.7
w/ GCA (Direct Replacement)	33.6
w/ GCA + Soft Integration	34.0
w/ GCA + Continued Learning	34.9
w/ GCA + Soft Integration + Continued Learning (i.e., Cross-View Dec.)	**35.9**

Table 5: Ablation results on the IWSLT DE-EN dataset. The direct replacement has a negative accuracy impact.

over the vanilla direct replacement, albeit minimal to the baseline, meaning the single-view approach cannot realize the potential of multi-level source representations. Complementing GCA with the global cross-view has a positive impact, especially when combined with continued learning, reaching an overall improvement of 1.2 over the baseline, indicating the joint advantages of the two proposed methods in addressing the hierarchy bypassing problem. We also check if the accuracy improvement is simply due to longer training that comes with continued learning. If the conventional models are allowed the same more training time, their results are not bettered meaningfully. The increase is 0.0, 0.1, and 0.3 in terms of BLEU for DE-EN, EN-DE, and EN-FR, respectively.

4.3 Hierarchy Bypassing

The hierarchy bypassing problem can be better understood through analyzing the granularity of the source representations. In Figure 4, we show the cosine similarity within the source sequence of the first and the last encoder layer. For the conventional transformer, i.e., Eq. (3), we can see that the last encoder layer embodies diffused similarities, indicating more coarse-grained representations, where related information is better associated. However, with direct replacement as in Eq. (5), the similarities in the last encoder layer remain relatively concentrated, suggesting impaired learning of representation hierarchy and causing the hierarchy bypassing problem. In contrast, with the proposed cross-view decoding, i.e., Eq. (8), the hierarchy bypassing problem is effectively relieved, enabling the performance improvement of sequence-to-sequence learning with deep hierarchical and diverse representations.

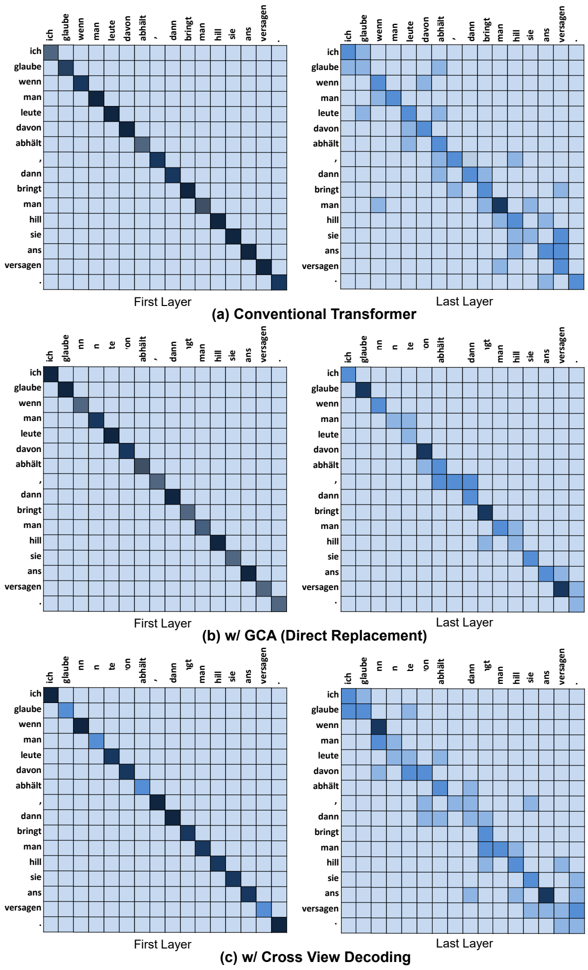


Figure 4: Visualization of the cosine similarity within the source sequence at the first encoder layer (left) and last encoder layer (right).

4.4 Length Analysis

In order to analyze the performance of our approach on sentences of different lengths, following in Bahdanau et al. (2015) and Dou et al. (2018), we group sentences of similar lengths together and calculate the BLEU score for each group. As shown in Figure 5, our approach is superior to the baseline in all length segments on DE-EN, EN-DE and EN-FR datasets. It is worth noticing that the proposal is always better than the baseline, and the longer the sentences, the more the improvements. Intuitively, it is hard for the global representation from the final encoder layer to retain all the detailed input information, especially for longer sentences. However, in conventional encoder-decoder model, the decoder is only equipped with a single view of the source sequence, which causes a dilemma that although both global and local information are important, only one can be used. In contrast, we can avoid the dilemma by adopting

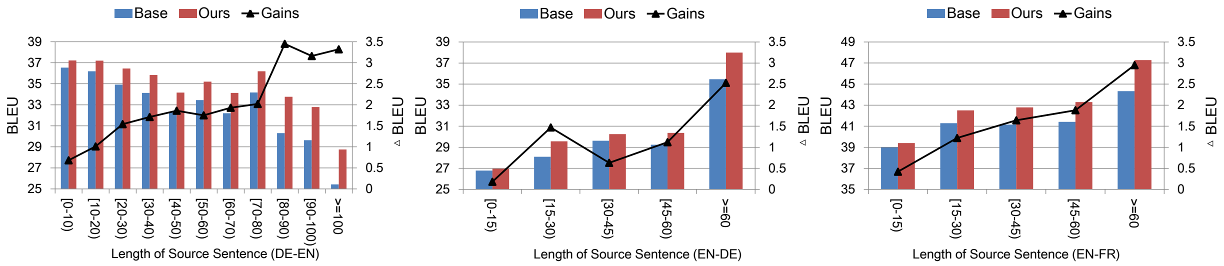


Figure 5: BLEU scores on the test sets with respect to various input sentence lengths. The gains in different length of sentences are shown with the polyline and the right y-axis.

Methods	Abstractive Summarization			Image Captioning	
	ROUGE-1	ROUGE-2	ROUGE-LCS	SPICE	CIDEr
Transformer (re-implementation)	39.3	17.3	36.2	21.2	124.9
w/ Cross-View Dec.	**39.9 (+0.6)**	**18.0 (+0.7)**	**36.8 (+0.6)**	**22.6 (+1.4)**	**129.4 (+4.5)**

Table 6: Results of text summarization and image captioning.

cross-view decoding, which injects fine-grained representations to the decoder and keeps the original global representation at the same time.

4.5 Generalization Analysis

It is interesting to see whether our approach works for other sequence-to-sequence tasks. To this end, based on GCA, we further conduct experiments on the CNN-Daily Mail dataset for abstractive summarization (Hermann et al., 2015) and the COCO dataset for image captioning (Chen et al., 2015). The improvements are all statistically significant (t-test with $p < 0.01$). For more details, please refer to Appendix A.

Abstractive Summarization Table 6 shows the results in terms of ROUGE-1, -2 and -LCS (Lin, 2004). The proposal achieves an advantage over the baseline, which indicates that our approach generalizes well to tasks with much longer source sequences, which are around 400 words and is effective in dealing with longer sequences. In fact, in such scenarios, where the summary should be concise but also accurate in detail, the GCA naturally streamlines the decoding process.

Image Captioning This task combines image understanding and language generation and is a cross-modal setting compared to machine translation and text summarization. The source sequence contains non-ordered region-of-interest features (Anderson et al., 2018). Table 6 reports the results on test set in terms of SPICE (Anderson et al., 2016) and CIDEr (Vedantam et al., 2015), which are specifically designed to evaluate image captioning systems. As we can see, our approach further improves the performance of baseline to 22.6 SPICE score and 129.4 CIDEr score, which is competitive with the state-of-the-art models (Huang et al., 2019; Yang et al., 2019). The results suggest that the proposal could be extended to a wide range of sequence generation tasks with various kinds of source representations.

4.6 Case Study

In Figure 6, we list some examples from DE-EN and COCO to analyze how our approach improves the baseline. The examples show that our approach does not alter the structure of the output sentence significantly compared to the baselines. The reason is that our approach can be seen as an extension for fine-tuning existing models. However, our approach can capture more detailed information about the source sequence. For example, in machine translation, our approach enables the model to use words that are more precise, including verb forms, and singulars/plurals, especially when the baseline is unable to choose a proper word to continue the sentence, e.g., repetition. For the image captioning, our approach helps the model to generate more detailed captions in colors (e.g., *pink* umbrella) and attributes (e.g., *rainy* street), for each object.

5 Related Work

Sequence-to-Sequence Learning In recent years, many deep neural systems have been proposed for sequence to sequence learning.

IWSLT DE-EN Translation Task		COCO Image Captioning Task	
Source	**Target**	**Source**	**Target**
stellen sie sich vor , dass jedes einzelne von menschen gesteuert wird .	**Reference**: and imagine each of them then controlled by people . **Baseline**: imagine that every single one of people are being controlled . **Ours**: imagine that every single person is controlled by people .		**Reference**: a lady walking in the rain carrying a pink umbrella. **Baseline**: a woman walking down a street with an umbrella. **Ours**: a woman walking down a rainy street with a pink umbrella.
aber ich glaube wirklich , dass nach uns noch etwas kommt .	**Reference**: but i think that there really is something coming along after us . **Baseline**: but i really believe that that is coming after us . **Ours**: but i really believe that something is coming after us .		**Reference**: a man wearing a white shirt and black tie. **Baseline**: a man wearing a white tie and shirt. **Ours**: a man wearing a white shirt and black tie with chairs nearby.

Figure 6: Examples of the target sentences generated by different methods. The color Blue denotes the examples when our model generates better target sentences than the baseline, while Red denotes unfavorable results.

The commonly-used approaches (Bahdanau et al., 2015; Vaswani et al., 2017) depend on the encoder-decoder framework to map a source sequence to a target sequence, such as in machine translation. The encoder network computes intermediate representations for the source sequence and the decoder network defines a probability distribution over target sentences given that intermediate representation. Specifically, to allow a more efficient use of the source sequences, a series of attention methods (Vaswani et al., 2017; Xu et al., 2015; Luong et al., 2015) have been proposed to directly provide the decoder with source information. Especially, the recent advent of fully-attentive models, e.g., Transformer (Vaswani et al., 2017), in which no recurrence is required, has been proposed and successfully applied to multiple tasks, e.g., neural machine translation. The work on attention reveals that attention is efficient, necessary, and powerful at combining information from diverse sources. Despite their dominance in the last few years, little work has been done discussing the effect of the connection between the encoder and the decoder in sequence-to-sequence learning.

Using Deep Representations In natural language processing, several efforts (Peters et al., 2018; Shen et al., 2018; Wang et al., 2018; Dou et al., 2018; Domhan, 2018; Bapna et al., 2018; Dou et al., 2019; Wang et al., 2019; Li et al., 2020) have investigated strategies to make the best use of deep representations among layers, e.g., using linear combination (Peters et al., 2018), dense connection (Shen et al., 2018) and hierarchical layer aggregation (Dou et al., 2018). However, most of them (Wang et al., 2018; Dou et al., 2018, 2019; Wang et al., 2019; Li et al., 2020) focused on the information within the encoder or the decoder, and excluded the effect of the information flow from the encoder to the decoder, which is the focus of this work. Some of them (He et al., 2018; Domhan, 2018; Bapna et al., 2018) considered individual source-to-target attention routing strategy in non-comparable contexts and we provide a unified, systematic overview of such efforts and frame them under the problem of narrowing decoder-encoder information gap. More importantly, they still adopted the single-view decoding approach and faced the hierarchy bypassing problem analyzed in this work, leading to unpromising results in common settings. The cross-view decoding approach is unique to previous work and substantially improve the performance of sequence-to-sequence learning using deep representations.

6 Conclusions

In this work, we focus on enhancing the information transfer between the encoder and the decoder for sequence-to-sequence learning, through injecting diverse source representations into the generation process. Different from the single view approach in existing work, we propose the *layer-wise cross-view decoding* approach to route source representations of different granularity to different decoder layers, together with a global view preventing the regression of last encoder layers due to the *hierarchy bypassing* problem. The cross-view decoding also builds upon continued training to mine the expressive power of encoder layer stack in learned conventional single-view models. Out of several source-target routing strategies, we find that the granularity consistent attention (GCA) strategy for context attention shows the best improvements with the proposed layer-wise cross-view decoding. Experiments on the machine translation task verify the effectiveness of our approach. In particular, it outperforms the DynamicConv model which is the previous state-of-the-art in machine translation.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *ECCV*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and VQA. In *CVPR*.

Lei Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Ankur Bapna, Mia Xu Chen, Orhan Firat, Yuan Cao, and Yonghui Wu. 2018. Training deeper neural machine translation models with transparent attention. In *EMNLP*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Tobias Domhan. 2018. How much attention do you need? A granular analysis of neural machine translation architectures. In *ACL*.

Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Shuming Shi, and Tong Zhang. 2018. Exploiting deep representations for neural machine translation. In *EMNLP*.

Zi-Yi Dou, Zhaopeng Tu, Xing Wang, Longyue Wang, Shuming Shi, and Tong Zhang. 2019. Dynamic layer aggregation for neural machine translation with routing-by-agreement. In *AAAI*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *EMNLP*.

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *NMT@ACL 2018*.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *NeurIPS*.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. 2019. Attention on attention for image captioning. In *ICCV*.

Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *ACL*.

Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*.

Jian Li, Xing Wang, Baosong Yang, Shuming Shi, Michael R. Lyu, and Zhaopeng Tu. 2020. Neuron interaction based representation composition for neural machine translation. In *AAAI*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL-HLT*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *WMT*, pages 1–9.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Yanyao Shen, Xu Tan, Di He, Tao Qin, and Tie-Yan Liu. 2018. Dense information flow for neural machine translation. In *NAACL-HLT*.

David R. So, Quoc V. Le, and Chen Liang. 2019. The evolved transformer. In *ICML*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *CVPR*.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *ACL*.

Qiang Wang, Fuxue Li, Tong Xiao, Yanyang Li, Yinqiao Li, and Jingbo Zhu. 2018. Multi-layer representation fusion for neural machine translation. In *COLING*.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *ICLR*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*.

A Task and Implementation Details

A.1 Machine Translation

Datasets We report results on three benchmarks, including two large WMT-2014 datasets, i.e., English-German (EN-DE) and English-French (EN-FR), and a small IWSLT-2014 datasets, i.e., German-English (DE-EN). Follow common practice (Vaswani et al., 2017; Wu et al., 2019), for EN-DE dataset, we use newstest2013 for development and newstest2014 for testing. For EN-FR, we validate on newstest2012+2013 and test on newstest2014. For fair comparisons, following Wu et al. (2019) and Vaswani et al. (2017), for WMT EN-DE and EN-FR we measure case-sensitive tokenized BLEU (`multi-bleu.pl`) against the reference translations. For IWSLT DE-EN, the BLEU is case-insensitive, and since the target language is English, the results are also valid. For WMT EN-DE only, we apply compound splitting similar to Wu et al. (2019) and Vaswani et al. (2017). It is unclear how some of the compared methods calculated the BLEU scores, but results provided by us are comparable.

Implementation We use the *fairseq* (Ott et al., 2019) for both our re-implementation of baselines and baselines with proposal. For the experiments with Transformer on the two WMT datasets that are much larger, we use the Transformer-Big configuration and train on 8 GPUs. For the experiments on the IWSLT DE-EN dataset, we use the Transformer-Base configuration and train on a single GPU, as it is relatively small. For WMT EN-DE and EN-FR datasets, we also accumulate the gradients for 16 batches before applying an update (Ott et al., 2018), except for Transformer on EN-FR where we do not accumulate gradients. For all datasets, we report single model performance by averaging the last 10 checkpoints. Besides, we use beam search of size 4 and length penalty of 0.6 for EN-DE and EN-FR, and use beam search of size 5 for DE-EN.

A.2 Abstractive Summarization

Dataset We train the models on the CNN-Daily Mail dataset (Hermann et al., 2015), which contains online news articles (781 tokens on average) paired with multi-sentence summaries (56 tokens on average). Following See et al. (2017), we truncate each source sentence to 400 words and each target sentence to 100 words. ROUGE-1, -2 and

-LCS (Lin, 2004) are used to evaluate the performance of models. The dataset is able to test the ability of our approach to deal with longer texts.

Implementation We use the default setting provided by *OpenNMT* (Klein et al., 2017). For the experiments with Transformer, we use the Transformer-Base configuration and train on a single GPU. When generating summaries, we follow standard practice in tuning the maximum output length, disallowing repeating the same trigram, and applying a stepwise length penalty (Paulus et al., 2018; Fan et al., 2018).

A.3 Image Captioning

Dataset The task combines image understanding and language generation and is a cross-modal task compared to machine translation and text summarization. We conduct experiments on the popular COCO dataset (Chen et al., 2015). We use the publicly-available splits in Karpathy and Li (2015) for evaluation. There are 5,000 images each in validation set and test set. Following common practice, we replace caption words that occur less than 5 times in the training set with the generic unknown word token UNK, resulting in a vocabulary of 9,487 words. We report results with the help of the evaluation toolkit (Chen et al., 2015), which includes the commonly-used metrics SPICE (Anderson et al., 2016) and CIDEr (Vedantam et al., 2015), which are customized for evaluating image captioning systems, based on scene-graph and n-gram matching, respectively, which are more consistent with human judgment.

Implementation For experiments on the image captioning dataset, we use the Transformer-base model and train on a single GPU. For fair comparisons, we use the RCNN-based image features provided by Anderson et al. (2018). During inference, we apply beam search with beam size = 5.