Exploring Human-like Learning Capabilities of Neural Machine Translation

Ph.D. Thesis Proposal

Dušan Variš
Faculty of Mathematics and Physics, Charles University,
Malostranské náměstí 25
118 00 Prague, Czech Republic
varis@ufal.mff.cuni.cz

Abstract

Neural networks became the dominant approach to solving many statistical modeling problems, reaching or surpassing human-level performance on several tasks. Even though they were originally inspired by the neural interactions inside the biological brain, their learning process is very different from that of a human. For this reason, there are several aspects of human-like learning that the current state-of-the-art networks cannot capture. This thesis proposal focuses on describing crucial problems of a sequential learning in the neural natural language processing. Furthermore, we look into possible approaches to solving these problems and the results of the research we've done so far. Lastly, we provide a proposal for directing our future thesis work.

1 Introduction

In recent years, deep learning became the state-of-the-art approach to solving various natural language processing (NLP) tasks (Graves and Jaitly, 2014; Xu et al., 2015; Anderson et al., 2018) including machine translation (MT, Bahdanau et al., 2014; Vaswani et al., 2017) slowly getting close to human-level performance on several of these tasks.

Being inspired by the interactions between the real biological neurons, it would seem logical that neural networks (NN) should be an important component when building a general purpose artificial intelligence (AI) systems. On the other hand, these similarities to human brain vanish when we take a closer look at their learning process. Using gradient-based methods, NN often require large numbers of training examples or various regularization techniques to learn proper generalizations. These methods are usually successful when the network tries to learn a single specific task, however, the difference from the human-like learning becomes more apparent when we consider learning multiple heterogeneous tasks, especially in a continual manner, i.e. learning one task after another. While being quite simple for humans, NN manifest difficulties with such learning scenarios.

Generally speaking, there are three main problems related to the continual multi-task learning in NN:

1. **Catastrophic forgetting.** A problem of overwriting (forgetting) network weights optimized for the initial task when training on a new task.

2. **Knowledge composition.** A problem of solving a complex task by breaking it down to simpler problems that the network can already solve and combining the knowledge about solving these tasks to solve the original task.

3. **Low-resource learning.** An ability to learn new task from a very small number of examples possibly by leveraging the knowledge about similar patterns from previously learned tasks.

Furthermore, if we consider the multi-modal aspects of the human condition we must also include the problem of multi-modal **representation learning**. When processing inputs from different modalities (e.g. language and vision), it is natural for human brain to properly connect concepts that are related across these modalities. This is usually achieved with NN by jointly learning tasks using inputs from these modalities. One problem is that it is not always clear whether the network learned a unified concept across the modalities of if two independent task solvers simply live "next to each other". Another problem is that this can be more challenging in the continual learning scenarios.

In this text, we focus on these problems in the context of single-source and multi-source MT models and their unsupervised (or semi-supervised) initialization. Our aim is to investigate methods for effective transfer of knowledge about previously learned tasks without losing this prior knowledge. If possible, we also plan to adapt our findings to the multi-modal (text with images as an input) translation scenario.

We also present our experimental results so far. We show our initial results in the task of automatic post-editing (APE) together with a discussion of the task challenges. Then, we discuss our previous work on the multi-modal machine translation (MMMT) task. Also, we present the results of our experiments tackling the problem of catastrophic forgetting during unsupervised neural machine translation (NMT) initialization. Furthermore, we discuss our unpublished results in the area of visual object representation (VOR) grounding in the context of image captioning (IC).

Lastly, we propose our plans for the future work. While discussing alternative approaches to the problems we already investigated, we will focus mainly on knowledge composition (discussed in more detail in Section 2.2). We plan to explore the effects of the current state-of-the-art techniques on the multi-source and multi-lingual translation and whether they are complementary with the methods related to catastrophic forgetting.

This proposal is structured as follows. In Section 2 we describe the problems of continual learning in more detail together with the current state-of-the-art. In Section 3, we describe our experiment methodology. Section 4 summarizes our experiments so far. We describe our future research plans in more detail in Section 5.

2 Related Work

There has been an extensive work on problems related to the continual learning in neural networks, often related to the findings in the field of cognitive psychology. In this section, we take a closer look at the recent research related to the problems listed in the previous section.

2.1 Catastrophic Forgetting

One of the key requirements when developing a general AI is the ability to solve multiple tasks learned in a sequential manner (Legg and Hutter, 2007). However, it has been observed that the current state-of-the-art networks suffer from *catastrophic forgetting* (French, 1999; McCloskey and Cohen, 1989). This occurs when a neural network trained to solve one task (e.g. task A) is used to learn a new task (e.g. task B). Without any constraints, the network adjusts its weights to solve task B while forgetting its original weight configuration for task A. However in practice, if we train the network jointly to solve both tasks A and B, we can still find a weight configuration that is performing well for both tasks (Luong et al., 2016).

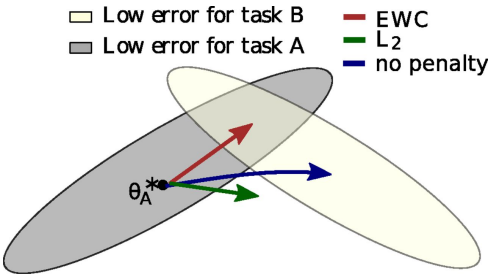

Figure 1: Illustration of catastrophic forgetting during continual training. Even though there exists an overlap of weight configurations with low error for both task A and B, optimization only for task B without any weight restriction can lead to leaving the low error weight configuration for the initial task A. The figure was taken from Kirkpatrick et al. (2017).

Learning a task requires searching for a configuration network weights which results in low-error performance on that particular task, usually defined by a loss function relevant to the task. In general, many weight configurations can result in a same performance (Hecht-Nielsen, 1992; Sussmann, 1992). Learning a to solve multiple tasks therefore requires finding a intersection between the low-error weight configurations for both task A and task B. In continual learning, this usually means shifting original solution θ_A for task A to a low-error area of the task B. If we assume no access to the original data for task A, optimization for task B without any constraints can lead to completely missing the desired low-error weight space intersection as illustrated by Figure 1. Network weight regularization is therefore required.

While joint task learning (or multi-task learning, Luong et al., 2016; Hashimoto et al., 2017) avoids catastrophic forgetting, it assumes that data for both task A and B will be available at the same time which may not be true in practice. Another option is storing network weights from the previous task separately and learning weight configura-

tion for the new task by fine-tuning the previous model. This however, can lead to a weaker generalization since the model forgets (possibly relevant) knowledge from the previous task by overfitting to the new task. In addition, the storage requirements needed to keep weight configurations for each task can be limiting. Hence, we are interested in methods that aim to restrict the network in a way that it does not completely override its knowledge about the previously learned tasks.[1]

The main idea behind restricting weights of the network is to force the network to minimize updates to weights that are important for the task A when learning task B. Pasunuru and Bansal (2019) achieve this by introducing two additional weight constraints. They force each weight matrix of the network to be *block-sparse*, using only a subset of the matrix weight for computation. After learning network weights θ_A for task A the network learns weight shift ψ_B such as $\theta_B = \theta_A + \psi_B$. To avoid forgetting task A, they again apply block-sparsity constraint on ψ_B and add a regularization term that forces orthogonality between θ_A and ψ_B. This helps the network to learn θ_B that performs well on both tasks.

Instead of separating weights for each task explicitly by the block-sparsity constraint, Kirkpatrick et al. (2017) introduce a regularization called elastic weight consolidation (EWC). They penalize the difference between the original and updated network weight, each scaled by its importance with respect to the original task. They derive this weight importance from a curvature of the loss function near the local minimum at the end of the training of the given task. To approximate this this, they use diagonal of a Fisher information matrix (FIM, MacKay, 1992) assuming that network weights are mutually independent. Computing FIM requires expected value of a gradient over all possible outputs for each data point, which is intractable. In practice empirical FIM is used as an approximation instead.

Another way of estimating importance of the network weights is path integral (PI, (Zenke et al., 2017)), focusing on their influence on the loss surface over the whole optimization path. They define a surrogate loss using a similar regularization term as Kirkpatrick et al. (2017), however, they increment weight importance during each mini-batch update, accumulating it throughout the whole task training. Furthermore, they define the weight importance as a ratio of its contribution to the drop in loss and the size of its change after the update.

Recently, a generalization of EWC and PI, theoretically grounded in KL-divergence has been proposed (Chaudhry et al., 2018). Still using FIM, they introduce an efficient way to store and update the FIM diagonal using a moving average. Additionally, they propose a methodology for evaluating continual learning capabilities, measuring forgetfulness and intransigence of a model across the examined sequence of tasks.

Alternatively, methods focusing on regularization of the network activations (outputs) can by used instead (Li and Hoiem, 2016; Rebuffi et al., 2016). Although they offer a better flexibility when updating network weights, they can become memory inefficient with the increasing number of network activations.

2.2 Knowledge Composition

Knowledge composition and problem decomposition are key aspects of human learning. For example, a popular hypothesis in cognitive linguistics postulates that humans are able to produce potentially infinite number of sentences while only learning a finite set of production rules (Chomsky, 1965). On the other hand, current neural networks are generally very data-hungry, requiring huge number of training examples, which makes model scaling troublesome.

Although increasing the number of network weights, given enough training data, leads to a better prediction accuracy (Sutskever et al., 2014; Amodei et al., 2016; Devlin et al., 2019), it still requires execution of the whole network together with more training iterations. The resulting computation complexity growth is quadratic with respect to the number of network weights. On the other hand, the human brain usually requires only few of its regions to be active at the same time (Ramezani et al., 2014). There are also hypotheses that a modularity of these biological neural connections leads to a more optimized energy costs (Clune et al., 2013; Legenstein and Maass, 2002) and helps countering catastrophic forgetting (Sporns and Betzel, 2015).

The main approach to modularity in deep learning focuses on replacing a single layer with a set

[1] By referring to knowledge we mostly mean information stored in the network weight configuration.

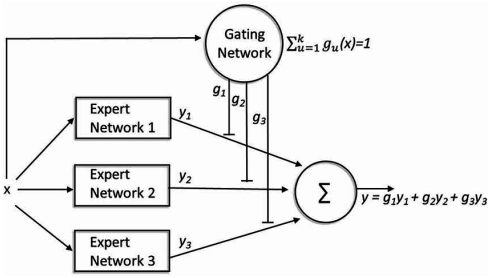

Figure 2: Illustration of a version of mixture of experts network. Each Expert Network learns a separate transformation of the input x and their outputs are combined using a weighted sum. The weights are produced by a Gating Network, also conditioned by the input x. The figure was taken from Bock and Fine (2014).

of smaller blocks (of the same type) and learning to compose them based on their learned specialization. This is usually accomplished using mixtures of experts (Jacobs et al., 1991; Jordan and Jacobs, 1994; Eigen et al., 2014) shown in Figure 2. The model learns separate functions for each "expert" block and a gating mechanism conditioned by the layer input, combining them usually through weighted sum of their outputs. These blocks can range from simple linear layers (Eigen et al., 2014) to whole encoder/decoder architectures similar to ensembles (Garmash and Monz, 2016).

Although it allows better computation parallelism, it still requires execution of all experts within the layer. This can be addressed by introducing sparsity, choosing only top-k scored experts (Shazeer et al., 2017). Another option is to learn an activation-dependent policy using reinforcement learning that decides which expert blocks to use for the computation (Bengio et al., 2016). However, these methods are prone to module collapse, lacking diversity when choosing experts during training often due to self-reinforcing of the favored modules by training them more rapidly (Kirsch et al., 2018). To counter this, some form of regularization needs to be enforced during the training.

Addressing these issues, Kirsch et al. (2018) suggest using stochastic selection of a subset of modules instead of mixtures. They treat the subset choice as a latent variable requiring summation over all possible subsets to generate output distribution. To avoid computational explosion during training, they use a generalized Expectation-Maximisation (EM, (Neal and Hinton, 1998)). In the estimation step, they sample a small number of module subsets based on the current *module probability* distribution and then choose the best candidate subset by maximizing *output probability* of the current training label. In the maximization step, they use the best candidate module subset to compute the loss on several data points and update both *module probability* and *output probability* distribution by adjusting the network weights.

2.3 Low-Resource Learning

In this section, we address the low-resource learning from the MT perspective. Most of the time, the amount of available bilingual data for a given pair of languages is much lower then the amount of their respective monolingual resources. The data sparsity becomes even more evident when we increase the number of languages (e.g. tri-lingual, quad-lingual data, etc.).

Transfer learning addresses data sparsity by reusing models trained on similar task that was trained on a larger dataset. The model can be then fine-tuned for a new task requiring only a smaller amount of data. In the context of NMT, the idea is usually to estimate network weights on a high-resource language pair and fine-tune them on the low-resource data (Kocmi and Bojar, 2018).

Other approaches suggest sharing either the encoder between multiple languages (Dong et al., 2015) or having single shared attention mechanism between different combinations of languages (Firat et al., 2017). However, these are trained jointly for all languages not exploring the possibilities of continual training.

Another branch of research focuses on using monolingual corpora to improve bilingual machine translation. Currently, the most popular method is augmenting the bilingual training corpus using synthetic data created by translation of the monolingual corpora (Sennrich et al., 2016a). Although the synthetic sentences often contain translation mistakes, they help with generalization leading to improvements on the validation data.

Alternatively, the NMT encoder and decoder can be initialized with a pre-trained monolingual language model (LM) and fine-tuned it on the bilingual corpus (Ramachandran et al., 2017). To address over-fitting (which arguably might be a result of catastrophic forgetting of the original LM task) they combine the MT loss with the original LM losses used during pre-training.

Conneau and Lample (2019) expand on the idea of monolingual pre-training by introducing a cross-lingual language model. They train their model jointly on all languages at once, sampling examples from each language according to the size of their respective monolingual corpus. They demonstrate its usefulness for initialization of various NLP tasks including machine translation.

An extreme case of data sparsity is unsupervised machine translation, where there is no bilingual data available. Recent solutions utilize only monolingual corpora while exploiting the dual nature of machine translation (Artetxe et al., 2018b; Lample et al., 2017). They show promising results by initializing the NMT embeddings using a pre-trained cross-lingual embedding space (Artetxe et al., 2018a) and introducing additional de-noising and reconstruction losses. Artetxe et al. (2019) propose a refined version of the system also including phrase-based MT and incremental back-translation.

Generally, similar techniques as the ones listed above are also explored in the multi-source translation scenarios. Synthetic data is widely used in automatic post-editing (Junczys-Dowmunt and Grundkiewicz, 2016; Negri et al., 2018) and using additional incomplete, text-only data is beneficial in multi-modal translation (Grönroos et al., 2018). Additionally, Nishimura et al. (2018) proposed a robust multi-source MT able to translate multi-lingual input even if one of the expected language input is not present.

An open problem in multi-modal MT is the effective use of the visual modality. Current state-of-the-art systems overly rely on the textual modality only ignoring input images to a degree (Elliott, 2018). Several methods have been proposed including various attention combination strategies (Libovický and Helcl, 2017; Libovický et al., 2018), enhancing textual representations with visual embeddings (Caglayan et al., 2018) or including additional objectives to regularize them using the visual modality (Elliott and Kádár, 2017).

2.4 Cross-modal Representation Learning

The ongoing research on multi-modal representation learning, inspired by human perception, aims at comprehensively representing information from different sensory inputs. In practice, these sensory inputs are processed by specialized parts of the neural network model. However, these parts are usually either trained on different tasks or the model lacks constraints forcing the network to learn better semantically correlated representations between the modalities. In literature, the latter is often referred to as the heterogeneity gap (Guo et al., 2019).

Current research is mostly focused on combining vision with textual modality, grounding natural language in its visual counterpart. Such grounding is motivated by different occurrences of object and the relations between them found in real world and in the written text (e.g. in text, people are *murdered* more often than they are *hugged*, Gordon and Van Durme, 2013). Main focus is therefore on reducing this divergence between text and reality.

In general, we can separate textual grounding into two categories: word-level and sentence-level grounding. Word-level methods either learn grounded representations jointly from multiple sources (Hill et al., 2016) or learn to combine independently learned representations from different modalities (Silberer and Lapata, 2014; Collell et al., 2017). Usually, word embeddings trained with a standard objective function such as skip-gram (Mikolov et al., 2013b) or continuous bag-of-words (Mikolov et al., 2013a) are grounded in the related visual features (Lazaridou et al., 2015) or visual context (Zablocki et al., 2018) using additional training objective.

The sentence-level grounding is mostly accomplished by introducing an additional objective for predicting visual features associated with the current sentence using cross-modal projection (Chrupała et al., 2015; Kiela et al., 2018; Elliott and Kádár, 2017). Collell and Moens (2018) argue that such a projection, however, does not preserve proper structure of the original embedding space. This constraint on the sentence embeddings can be relaxed by using an intermediate grounded space (Bordes et al., 2019).

In multi-modal MT, the most common method of incorporating visual information is by conditioning target-side decoder on the intermediate visual features. Usually, this conditioning is using an attention mechanism (Xu et al., 2015; Caglayan et al., 2016a). These mediated features can either come from the output of an intermediate layer of a image recognition network (Caglayan et al., 2016b, 2017) or from the representations of objects detected by an object detection network

(Grönroos et al., 2018). Still, it seems that current state-of-the-art systems only exploit distributional similarity in the visual feature space, generating outputs closest to the training examples (Madhyastha et al., 2018).

3 Methodology

In this section, we describe in a closer detail the tasks we focus on with respect to the MT-related continual learning research. In general, we make a distinction between two sets of continual learning:

- **Language-based.** The difference between tasks is in their language or languages. Our goal is training a single multi-task model with performance similar to its single-task counterparts.

- **Complexity-based.** We make a distinction between tasks based on the complexity of the conditioning of the probability distribution the model is trained to estimate. For instance, consider MT vs. automatic post-editing. In MT, the output is conditioned on the source sentence only whereas in post-editing, the model has to consider both the source and the candidate translation from the baseline MT model. The conditioning of the post-editing task is thus more complex. It is important to note that this enlarged source information *need not be useful* in practice; the post-editing model can operate as an independent MT model, fully ignoring the baseline MT it received in the input. However, the model setup is nevertheless more complex, because an untrained network *does not know* that the baseline MT is not relevant

From the language-based perspective, our main concern is transfer learning from a high-resource task A to a low-resource task B. Generally, we are interested in the relationship between avoiding catastrophic forgetting and network regularization, e.g. how much remembering task A can help reduce over-fitting on the small training data for task B. Additionally, we would like to explore possibilities of an *implicit* knowledge decomposition, i.e. how the neural network represents various aspects of language (e.g. syntax, semantics) and whether is this knowledge reused (e.g. exploiting similar properties of two and more languages). Contrary to transfer learning, we aim to be able to perform well for both task A and task B. We also assume that data available during training on task A are no longer available during the task B training.

In the complexity-based experiments, we aim to study possibilities of an *explicit* knowledge decomposition defined beforehand, e.g. how to leverage monolingual knowledge (learned from data) about languages when learning the mapping between them. Again, we are interested in the high-resource to low-resource scenarios because the amount of available parallel data generally decreases with the increasing number of languages involved.

We give a brief description of the examined tasks and the relationships between them in the rest of this section. We list the tasks based on the perceived complexity, from least complex to most. If not stated otherwise, the following description should be general and independent of the underlying network architecture.

Language modelling. The goal is to learn a probability distribution over the sentences in a language based on the monolingual training corpus. Given that each sentence can be represented as a sequence of N tokens $\mathbf{y} = \{y_0, .., y_{N-1}\}$, we aim to minimize the cross-entropy between the model probability distribution p_θ and the true distribution p^* given the training data D.

$$H(p_\theta, p^*) = \sum_{\mathbf{y} \in D} p_\theta(\mathbf{y}) p^*(\mathbf{y}) \tag{1}$$

In practice, due to intractability of the probability distribution over all possible sentences, p_θ is factorized over the sentence tokens, for example:

$$p_\theta(\mathbf{y}) = \prod_{i=0}^{N-1} p_\theta(y_i | \mathbf{y}_{<\mathbf{i}}) \tag{2}$$

Equation 2 assumes that probability of each token only depends on the previous tokens in the sentence. In practice, prediction of the token sequence $\mathbf{y}$ is based on intermediate representations $\mathbf{h} = \{h_0, .., h_{N-1}\}$ containing information about the previous predictions. Although such factorization assumes independence on the following tokens, it allows a simple auto-regressive (left-to-right) sequence generation. Figure 3 shows a schematic of such prediction model.

Other ways of factorization have been also suggested, e.g. masked language models for sequence

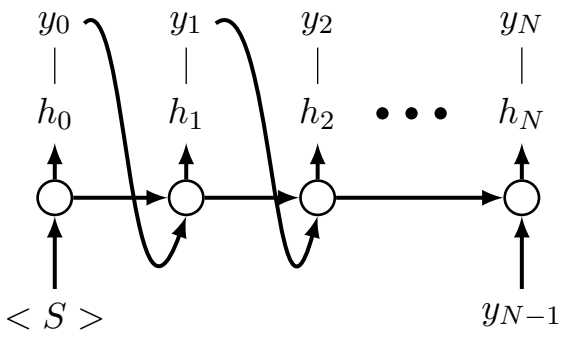

Figure 3: Simplified schematic of the auto-regressive decoding. At each decoding step, a current input is encoded by the input layer and then processed by the network (a circle node) using previous state h_{i-1}, producing next hidden state h_i. The hidden state h_i is further processed, producing y_i which is passed as the next input to the decoder. Although the actual production of the hidden state h_i may vary between network architectures, the underlying principle (conditioning on the previous outputs) remains the same.

classification (Devlin et al., 2019) or sequence models incorporating syntax (Zhang et al., 2016). Their main difference is in the method of modeling the hidden state sequence $\mathbf{h}$.

Machine translation. Similar to language modeling, the goal is to generate target-language sentence $\mathbf{y} = \{y_0, .., y_{N-1}\}$ given a source-language sentence $\mathbf{x} = \{x_0, .., x_{M-1}\}$ by learning the following conditional probability:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \prod_{i=0}^{N-1} p_\theta(y_i|\mathbf{y_{<i}}, \mathbf{x}) \tag{3}$$

Similarly to previous paragraph, Equation 3 is already the most common factorization, producing one target token at a time. In practice, instead of using input sequence $\mathbf{x}$, we condition the output generation on the hidden states $\mathbf{h} = \{h_0, .., h_{N-1}\}$ produced by sentence encoder. We assume that these hidden states should capture similar information as the hidden states learned during language modeling, thus making knowledge between these two tasks transferable.

The output sequence $\mathbf{y}$ is then produced by a sentence decoder. With some simplifications, the decoder is a language model explicitly conditioned on the hidden states $\mathbf{h}$. Therefore, we can also assume that the weights of the decoder and the language model can be transferred. Given these assumptions, we can further modify Equation 3 as:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \left(\prod_{i=0}^{N-1} p_{\theta_{dec}}(y_i|\mathbf{y_{<i}}, \mathbf{h})\right) p_{\theta_{enc}}(\mathbf{h}|\mathbf{x})) \tag{4}$$

where $\theta_{enc} \cup \theta_{dec} \subseteq \theta$ are the respective encoder and decoder weights.

Based on these assumptions, we define an explicit decomposition within the MT task as learning structure of a language (for both source and target language) and finding relationship between these structures. Whether monolingual structure learning should be modelled by a single multilingual network or multiple language-specific networks is currently an open research question.

Multi-source translation. We generalized MT task to k sources of input in the following way. Let us assume K different input sources $\mathbf{X} = \{\mathbf{x^{(j)}}|0 \leq j < K\}$ and an output sequence $\mathbf{y}$. Given that at least one of the inputs is in language L_1, the output is in language L_2 and $L_1 \neq L_2$, multi-source MT estimates the following distribution:

$$p_\theta(\mathbf{y}|\mathbf{X}) = \prod_{i=0}^{N-1} p_\theta(y_i|\mathbf{y_{<i}}, \mathbf{X}) \tag{5}$$

Generally, each input $\mathbf{x^{(j)}}$ can be processed by a specialized encoder, producing a hidden representation $\mathbf{h^{(j)}}$. Let us define $\mathbf{H} = \{\mathbf{h^{(j)}}|0 \leq j < K\}$. In practice, we usually assume a mutual independence between the inputs $\mathbf{X}$, leading to a more effective training process. Thanks to this assumption, we can process each input $\mathbf{x^{(j)}}$ with a separate encoder with weights $\theta_{enc}^{(j)}$, leading to the following factorization:

$$p_\theta = \left(\prod_{i=0}^{N-1} p_{\theta_{dec}}(y_i|\mathbf{y_{<i}}, \mathbf{H})\right) \prod_{j=0}^{K-1} p_{\theta_{enc}^{(j)}}(\mathbf{h^{(j)}}|\mathbf{x^{(j)}})) \tag{6}$$

Equation 6 leads again to an explicit decomposition where, for example, a single-source MT system can be adapted to multi-source MT reusing weights from the original task of bilingual MT and an additional encoder for the new input. Also note that the inputs do not have to be explicitly textual, for example, this decomposition can be applied to multi-modal translation initializing the multi-modal system with an image captioning model (visual input, textual output) and a language model as a source language encoder, etc.

3.1 Evaluation

Since human evaluation is costly and not exactly reproducible, it is not suitable for the development. Multiple automatic evaluation metrics were proposed for each of the considered tasks. The quality of a language model is usually measured by the perplexity of its sentence-level probability distribution with respect to the evaluation data. For MT evaluation in general, metrics based on n-gram precision such as BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005) are commonly used.

Forgetting and *Intransigence* metrics were recently proposed to quantify continual learning capabilities of a model (Chaudhry et al., 2018; Kim et al., 2019). They describe *Forgetting* as a measure of how much knowledge of previous tasks is preserved by a model and *Intransigence* as a measure of its inability to learn new tasks. Currently, due to the nature of the studied tasks these metrics are accuracy-based.

4 Experiments

In this section, we present both our published and unpublished experimental results.

Our experiments so far focused mostly on monolingual (language modeling, MT initialization) and bilingual tasks (MT, multi-modal MT, post-editing). In the next section, we will describe our plans for a follow-up in a more complex settings.

4.1 Automatic Post-Editing

In this section, we describe our submission to automatic post-editing (APE) task at WMT17 (Variš and Bojar, 2017). The goal of this task is to develop a system that corrects machine translation errors produced by an unknown MT system. Specifically, given a source language sentence and a sentence translated by MT, the APE system should produce a translated sentence of the same or higher translation quality.

In our submission, we compared different input processing strategies, and network architectures. Additionally, we also tried improving our system using synthetic post-editing data (Junczys-Dowmunt and Grundkiewicz, 2016). For input processing, we compared single-encoder architecture with a multi-encoder architecture. A multi-encoder architecture uses separate encoder for source language sentence and for its translation. Although these weights can be shared between the encoders, we only investigated the setup with separate sets of weights. A single-encoder architecture concatenates source language sentence and its translation and treats it as a single input sequence (Niehues et al., 2016). Although this results in a fewer network weights, the complexity of the input sequence increases.

Regarding architectures, we examined a recurrent neural network (Sutskever et al., 2014) and a recurrent-over-convolutional network (Lee et al., 2017), both combined with attention mechanism (Bahdanau et al., 2014). We used subword tokenization (Sennrich et al., 2016b) in combination with the first architecture, however, recurrent-over-convolutional architecture was designed to work without any explicit segmentation, processing the streams character by character.

In the end, we found that the addition of the additional synthetic data had the biggest impact on the system performance. The results also suggested that using character-level architecture benefits the post-editing task more than the subword tokenization. Based on our manual examination of the post-edited sentences, we concluded that evaluation by the automated metrics such as BLEU (Papineni et al., 2002) might not be suitable for this task.

4.2 Multi-Modal Machine Translation

This section describes our submission to multi-modal translation task at WMT18 (Helcl et al., 2018). The task focuses on translating a textual input given an additional visual input to help resolve possible ambiguities. The visual input is an image and the translated sentence is a caption describing the image.

We used a self-attentive network (Vaswani et al., 2017) in our the submission experiments instead of the recurrent neural network (Sutskever et al., 2014). We used a multi-source approach, combining textual representations produced by the encoder and intermediate visual features generated by a pre-trained image-classification network (He et al., 2016). We combine them by first attending over the textual representations and then over the visual features.

Furthermore, we included additional training objective called *Imagination* (Elliott and Kádár, 2017). Our aim was to learn grounded textual representation by approximating visual features using

related textual representations.

Although the suggested methods did improve our system over the purely textual baseline, the biggest performance gain was again achieved by preparing additional synthetic data. We used the MSCOCO (Lin et al., 2014) image captioning dataset and applied back-translation (Sennrich et al., 2016a) on the captions to generate synthetic source training sentences. We also added textual-only data from various bilingual corpora using perplexity-based filtering, only including examples similar to the caption translation domain.

Aside from comparing our systems using standard metrics for automatic evaluation of MT, we also provide results of the adversarial evaluation suggested by Elliott (2018). We report that explicit integration of the visual features into the self-attentive model improves overall performance, confirmed by both automatic metric and adversarial evaluation.

4.3 Unsupervised MT Pre-Training

In this section, we discuss our published results on the use of *Elastic Weight Consolidation* (EWC) in machine translation (Variš and Bojar, 2019). The main goal of these initial experiments was to tackle regularization capabilities of the EWC in the low-resource translation scenario.

Given two tasks, A and B, with their respective data $D_A \cup D_B = D$, the goal of multi-task learning is to maximize likelihood of the network weights given the whole data D. Kirkpatrick et al. (2017) show that this can be factored to a likelihood $p(D_B|\theta)$ and a prior knowledge about the previously learned task $p(\theta|D_A)$ using Bayes theorem:

$$p(\theta|D) = \frac{p(D_B|\theta)p(\theta|D_A)}{p(D_B)} \quad (7)$$

Equation 7 assumes using identical network for each task. We show in our work that under specific assumptions about network weight independence, this equation can be generalized for network optimization given parts of the network were trained for different tasks. In particular, we apply a modified EWC regularization on the unsupervised MT pre-training. Given two languages, L_1 and L_2, we first train separate language models using available monolingual data. Next, we initialize MT encoder and decoder using L_1 and L_2 language model respectively and fine-tune the MT model with the bilingual data. We use EWC to avoid forgetting the original language modelling tasks, however, we only investigate the regularization capabilities of the consolidation method.

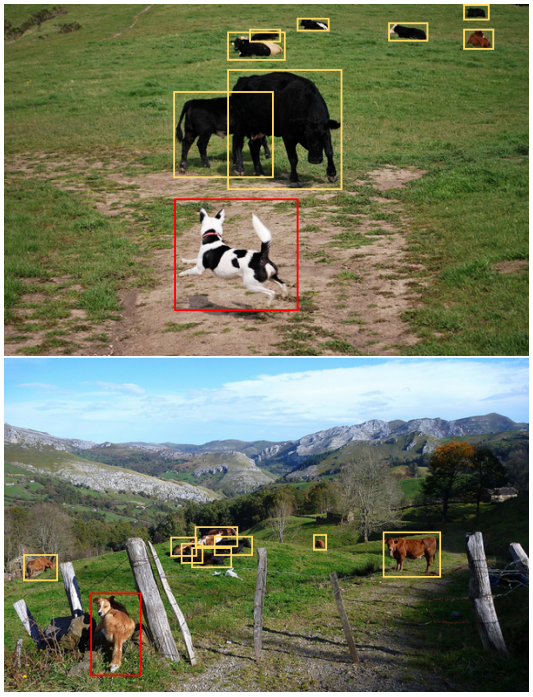

Figure 4: Example of two images where an identical sets of objects are detected even though the actions related to the objects differ (e.g. upper: dog jumping, lower: dog lying down).

We found minor improvements when using EWC regularization only during MT decoder pre-training. We also compared our method with another approach using the original language model loss as a form of regularization. While having similar results when applied to the decoder, our method has faster wall-time convergence possibly due to less arithmetic operations need during network updates. We suspect that the poor performance when regularizing encoder is due to a task mismatch. The pre-trained language model is an auto-regressive decoder, however, our MT self-attentive encoder models both left and right-side context.

4.4 Unpublished Results

Our most recent work was focused on multi-modal representation learning. We build upon the work of Anderson et al. (2018) suggesting using visual object representations (VOR) associated with objects detected within a picture by a Faster-RCNN object detection (OD) network (Ren et al., 2015). We study the methods of grounding these VORs in the word embedding space of their correspond-

ing labels and their effects on the image captioning (IC) task.

There were several proposals for using explicit object detections for IC instead of abstract features extracted from the intermediate layers of image processing network (Wang et al., 2018; Yin and Ordonez, 2017). However, these works treat the set of objects detected within a picture as a bag-of-objects (BOO), learning explicit embeddings of the object labels instead of directly using VORs produced by the OD network. We argue that BOO representation of an image is not expressive enough since two different images can have identical representation. Some level of disambiguation can be gained adding encoded information about the bounding box size and location of the objects, however, it still does not capture details about the actions associated with the objects as illustrated in Figure 4.

During our initial experiments we found that even though word embeddings learned from monolingual corpora and VOR extracted by an OD network exhibit similar properties, they occupy different parts of the vector space. Therefore, we tried grounding VOR in the word embeddings of their respective labels during IC training by adding additional *cluster loss* and *perceptual loss* suggested by Bordes et al. (2019) to the training objective.

Results of our experiments suggest that even though the system learns to cluster VORs properly, it does not lead to any significant IC improvements. One of the reasons might be in the limitations of our label set. We only used labels defined in the MSCOCO dataset (containing around 80 object classes) while Ren et al. (2015) used an OD system fine-tuned on Visual Genome (Krishna et al., 2017) which contains around 1600 object classes.

Furthermore, since VORs should already contain information about the context of an object, grounding them in context-less word embeddings might be too restrictive. We expect that using e.g. hidden states produced by the decoder (language model) might be more effective since they should also encode information about the label context.

5 Future Plans

So far, we have studied each of the continual learning problems mainly in isolation. Our following research will try to combine introduced techniques that counter these problems. However, our main focus will be on the problem decomposition and knowledge composition. Other problems such as catastrophic forgetting and low-resource learning will be studied only on the side.

The core of our research will be based around the self-attentive network architectures (Vaswani et al., 2017). These architectures are currently state-of-the-art not only in MT but also language modeling (Devlin et al., 2019) and they seem to be most efficient in terms of training data needed, as observed for summarization by (Çano and Bojar, 2019)).

Recently published results also showed cross-lingual capabilities of these models and their application to MT (Lample et al., 2017). Furthermore, studies of the multi-head attention mechanism, a key component of the self-attentive networks demonstrate the capacity for specialization. For example, it was shown in NMT that the attention heads can abstract linguistically-interpretable structures with different levels of contribution (Voita et al., 2019). However, deeper understanding of these capabilities is required.

In a standard *k-headed* self-attentive layer, each attention head computes a dot-product on a learned transformation of the input vector space. During training, it is hypothesised that these transformations can capture different attributes of the structure of the modeled data (Mareček and Rosa, 2019), providing a mechanism for information decomposition. However, the currently studied architectures are quite constrained having a fixed set of attention heads regardless of the input.

We propose a modification to the multi-head attention mechanism inspired by a work on modular networks (Kirsch et al., 2018) called modular self-attentive network. Similarly to the previous work, we will use a pool of modules and a controller mechanism to choose a subset of network modules depending on the layer input. In our case, a module will be single attention head. We plan on using a single controller with a single set of attention heads. We suggest focusing mainly on the following two controller schemes: *global* scheme should distribute the modules between layers based solely on the network input, *local* scheme chooses a set of modules for each layer separately based on the output of the previous layer. Figure 5 shows comparison between the original self-attentive network architecture and our proposed modifications.

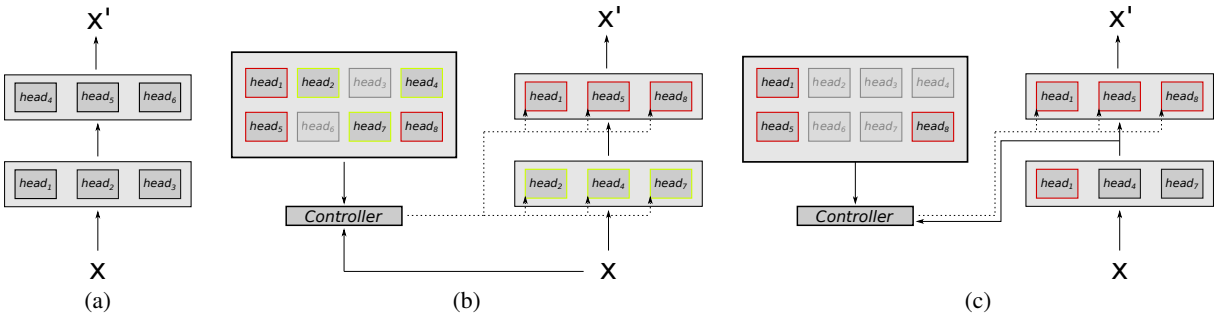

Figure 5: Comparison of the original self-attentive network architecture with the proposed modular modification. (a) Original network with two layers and three attention heads per layer, (b) modular network with the *global* head assignment (conditioned on the network input), (c) modular network with the *local* head assignment (conditioned on the output of the previous layer). The *local* variant allows module sharing between the network layers. We omit the feed-forward layers of the self-attentive network for simplicity.

Initial experiments will compare the performance of the state-of-the-art self-attentive networks with their modular variations. We will examine both proposed schemes. Besides hoping for better results, we are interested in the analysis of the emergent network structure. Since the choice of the network modules is conditioned by its input, we can for example study the self-attention modules using clustering methods. We will focus mostly on the MT task in this phase, namely German-English and Czech-English translation separate or multi-source.

As a next step, based on the results of these initial experiments, we will continue with the multilingual experiments. We will study only the network variation with the most promising results. Our focus will be on machine translation with identical target-side language and language modeling using training objectives suggested by Lample et al. (2017). In these experiments, we will be interested in the module specialization with respect to the languages and their linguistic closeness. Given the available monolingual and bilingual corpora, we plan to study mostly English, Czech and German. If the results are promising, we may also include Dutch-English MT. We plan to compare both joint training and the continual training. In the latter case, techniques that counter catastrophic forgetting (e.g. EWC) might be necessary.

In the last set of experiments, we will study the multi-task scenario. Similarly to cross-lingual initialization (Lample et al., 2017), we plan to reuse multi-lingual language models trained in the previous set of experiments for MT encoder/decoder initialization similar to our EWC experiment. We will then expand the idea to multi-source translation by either using a bi-lingual MT model trained for multiple language pairs or combining single MT model with an additional encoder initialized by a language model trained on the other source language.

We also plan several alternative experiment options in case some of the proposed settings prove to be ineffective. First, we can try other modular network approaches and apply them mainly to multi-source translation, for example the mixture-of-experts methods (Shazeer et al., 2017). We would be mainly studying different scope definitions of experts (e.g. layer-wise, encoder wise, etc.). Second, we can examine to which extent module-specialization and per-weight regularization methods (e.g. EWC) complement each other. Since the per-weight regularization assigns "importance" values to network weights, we can, for example, study which weights (or modules if taking an average importance) are considered important for a certain task or language. This can be further studied jointly with module specialization. Third, we can also apply our findings from shared representation space learning and see how the modular network approach benefits from such vector space constraints.

6 Conclusion

In this proposal, we described main problems of continual learning in the current state-of-the-art neural networks which is a key requirement for building general AI systems. We summarized related work with respect to each problem and pro-

vided details about the methodology related to our experiments. We also provided a brief overview of our experimental results so far, both published and unpublished.

In our future work, our main focus will be on two aspects of continual learning: task decomposition and knowledge composition. We proposed a modification to an existing state-of-the-art architecture introducing modularity to improve the model performance. Additionally, we plan to study possible contributions of the modified architecture to the model analysis. We structure our plans into several steps together with alternatives in case the intermediate results of the original proposal will stray from our expectations.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, Jie Chen, Jingdong Chen, Zhijie Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Ke Ding, Niandong Du, Erich Elsen, Jesse Engel, Weiwei Fang, Linxi Fan, Christopher Fougner, Liang Gao, Caixia Gong, Awni Hannun, Tony Han, Lappi Johannes, Bing Jiang, Cai Ju, Billy Jun, Patrick LeGresley, Libby Lin, Junjie Liu, Yang Liu, Weigao Li, Xiangang Li, Dongpeng Ma, Sharan Narang, Andrew Ng, Sherjil Ozair, Yiping Peng, Ryan Prenger, Sheng Qian, Zongfeng Quan, Jonathan Raiman, Vinay Rao, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Kavya Srinet, Anuroop Sriram, Haiyuan Tang, Liliang Tang, Chong Wang, Jidong Wang, Kaifu Wang, Yi Wang, Zhijian Wang, Zhiqian Wang, Shuang Wu, Likai Wei, Bo Xiao, Wen Xie, Yan Xie, Dani Yogatama, Bin Yuan, Jun Zhan, and Zhenyao Zhu. 2016. Deep speech 2 : End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 173–182, New York, New York, USA. PMLR.

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 Visual Question Answering. In *CVPR*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, and Doina Precup. 2016. Conditional computation in neural networks for faster models.

Andrew S. Bock and Ione Fine. 2014. Anatomical and functional plasticity in early blind individuals and the mixture of experts architecture. *Frontiers in Human Neuroscience*, 8:971.

Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and patrick Gallinari. 2019. Incorporating Visual Semantics into Sentence Representations within a Grounded Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 696–707, Hong Kong, China. Association for Computational Linguistics.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. Lium-cvc submissions for wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016a. Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation*, pages 627–633, Berlin, Germany. Association for Computational Linguistics.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 597–602, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016b. Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Erion Çano and Ondej Bojar. 2019. Efficiency metrics for data-driven models: A text summarization case study. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG 2019)*, pages 229–239, Stroudsubrgh, PA, USA. Association for Computational Linguistics.

Arslan Chaudhry, Puneet Kumar Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 556–572. Springer.

Noam Chomsky. 1965. Aspects of the theory of syntax. *Journal of Philosophy*, 64(2):67–74.

Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. Learning Language through Pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 112–118, Beijing, China. Association for Computational Linguistics.

Jeff Clune, Jean-Baptiste Mouret, and Hod Lipson. 2013. The evolutionary origins of modularity. *Proceedings of the Royal Society B: Biological Sciences*, 280(1755):20122863.

Guillem Collell and Marie-Francine Moens. 2018. Do Neural Network Cross-Modal Mappings Really Bridge Modalities? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 462–468, Melbourne, Australia. Association for Computational Linguistics.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. 2017. Imagined Visual Representations As Multimodal Embeddings. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 4378–4384. AAAI Press.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2014. Learning factored representations in a deep mixture of experts. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T. Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Comput. Speech Lang.*, 45(C):236252.

Robert M. French. 1999. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3:128–135.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418, Osaka, Japan. The COLING 2016 Organizing Committee.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on International Conference*

on Machine Learning - Volume 32, ICML14, page II1764II1772. JMLR.org.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels. Association for Computational Linguistics.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep Multimodal Representation Learning: A Survey. *IEEE Access*, 7:63373–63394.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark. Association for Computational Linguistics.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Robert Hecht-Nielsen. 1992. *Theory of the Backpropagation Neural Network*, page 6593. Harcourt Brace & Co., USA.

Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):7987.

Michael I. Jordan and Robert A. Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural Comput.*, 6(2):181214.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning Visually Grounded Sentence Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 408–418, New Orleans, Louisiana. Association for Computational Linguistics.

Dahyun Kim, Jihwan Bae, Yeonsik Jo, and Jonghyun Choi. 2019. Incremental learning with maximum entropy regularization: Rethinking forgetting and intransigence. *CoRR*, abs/1902.00829.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114 13:3521–3526.

Louis Kirsch, Julius Kunze, and David Barber. 2018. Modular networks: Learning to decompose neural computation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2408–2418. Curran Associates, Inc.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Robert A. Legenstein and Wolfgang Maass. 2002. Neural circuits for pattern recognition with small total wire length. *Theoretical Computer Science*, 287(1):239 – 249. Natural Computing.

Shane Legg and Marcus Hutter. 2007. Universal intelligence: A definition of machine intelligence. *Minds Mach.*, 17(4):391444.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *European Conference on Computer Vision*, pages 614–629. Springer.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

David J. C. MacKay. 1992. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2018. End-to-end image captioning exploits distributional similarity in multimodal space. In *Proceedings of the British Machine Vision Conference (BMVC)*.

David Mareček and Rudolf Rosa. 2019. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy. Association for Computational Linguistics.

Michael McCloskey and Neil J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24:104–169.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Radford M. Neal and Geoffrey E. Hinton. 1998. A view of the em algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, volume 89 of *NATO ASI Series*, pages 355–368. Springer Netherlands.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1828–1836, Osaka, Japan. The COLING 2016 Organizing Committee.

Yuta Nishimura, Katsuhito Sudoh, Graham Neubig, and Satoshi Nakamura. 2018. Multi-source neural machine translation with missing data. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 92–99, Melbourne, Australia. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2019. Continual and multi-task architecture search. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1911–1922. Association for Computational Linguistics.

Prajit Ramachandran, Peter J. Liu, and Quoc V. Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language*

Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 383–391.

Mahdi Ramezani, Kristopher Marble, Heather Trang, Ingrid Johnsrude, and Purang Abolmaesumi. 2014. Joint sparse representation of brain activity patterns in multi-task fmri data. *IEEE transactions on medical imaging*, 34.

Sylvestre-Alvise Rebuffi, Alexander I Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2016. icarl: Incremental classifier and representation learning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 91–99, Cambridge, MA, USA. MIT Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Carina Silberer and Mirella Lapata. 2014. Learning Grounded Meaning Representations with Autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland. Association for Computational Linguistics.

Olaf Sporns and Richard Betzel. 2015. Modular brain networks. *Annual Review of Psychology*, 67:613–640.

Héctor J. Sussmann. 1992. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Dušan Variš and Ondřej Bojar. 2017. CUNI system for WMT17 automatic post-editing task. In *Proceedings of the Second Conference on Machine Translation*, pages 661–666, Copenhagen, Denmark. Association for Computational Linguistics.

Dušan Variš and Ondřej Bojar. 2019. Unsupervised pretraining for neural machine translation using elastic weight consolidation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 130–135, Florence, Italy. Association for Computational Linguistics.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Josiah Wang, Pranava Swaroop Madhyastha, and Lucia Specia. 2018. Object Counts! Bringing Explicit Detections Back into Image Captioning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2180–2193, New Orleans, Louisiana. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Xuwang Yin and Vicente Ordonez. 2017. Obj2Text: Generating Visually Descriptive Language from Object Layouts. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 177–187, Copenhagen, Denmark. Association for Computational Linguistics.

Éloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. 2018. Learning Multi-Modal Word Representation Grounded in Visual Context.

In *Association for the Advancement of Artificial Intelligence (AAAI)*, New Orleans, United States.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 3987–3995.

Xingxing Zhang, Liang Lu, and Mirella Lapata. 2016. Top-down tree long short-term memory networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 310–320, San Diego, California. Association for Computational Linguistics.