A Supplemental Material

A.1 Typical Examples from CNN/Daily Mail Dataset

We show other typical examples here. We compare our best model (end-to-end-trained with inconsistency loss) with other state-of-the-art methods.

Original article: Jamie Robbins (above) handed himself into the police a year after he stole £1,600 from a convenience store in Birmingham with an accomplice because he preferred life in jail. A robber handed himself into police a year after he raided a convenience store saying he preferred life in jail. **Jamie Robbins stole £1,600 from the Select 'n' Save shop in Kings Norton, Birmingham, with an accomplice on January 13, 2014. The case was not solved until the 35-year-old walked into a police station earlier this year** and said: 'Please arrest me.' **Robbins was jailed for four-and-a-half years after pleading guilty to robbery and possession of a bladed article** at Birmingham Crown Court on Thursday. Ayoub Khan, defending, said Robbins had felt he could not cope outside of prison and wanted to go back behind bars. But Mr Khan added that the day before his court appearance Robbins discovered his girlfriend was pregnant - and no longer wanted to be given a custodial sentence. 'There has been an enormous change in direction of his life. He now has purpose in his life,' he said. Judge Murray Creed, sentencing, described the events as 'unusual' as the inquiry had stopped. He said: 'But for your surrender it does not look as though matters would have recommenced. This was a group attack, planned, and there was a weapon and disguises.' Robbins, of no fixed address, gave detectives full details of the robbery. He explained how he and another masked man had waited outside the store until a woman and her small child had finished their shopping. Robbins then burst into the shop with a machete and his accomplice, and demanded the shopkeeper hand over the safe. Judge Murray Creed, sentencing, described the events as 'unusual' at the hearing at Birmingham Crown Court (above) He threatened the owner with the knife and the pair escaped with £1,600, which they split between them. Robbins even told police the getaway driver did not get his cut because he had not waited in the car as had been agreed. Christopher Lester, prosecuting, said: 'The day before the incident, Robbins was called and asked if he wanted to be in a robbery.' 'The next day he met two other men and the defendant took the weapon.' Mr Lee added that the original police investigation had not been successful and Robbins was only convicted because of the evidence he gave in interview.
Reference summary: Jamie Robbins raided a Select 'n' Save shop in Birmingham last January. Case was not solved until he walked into a police station and confessed. 35-year-old was sentenced to four-and-a-half years on Thursday.
DeepRL: Jamie Robbins stole £1,600 from a convenience store in Kings Norton, Birmingham, with an accomplice on January 13, 2014. *He stole himself into police a year after he raided convenience store.* He was jailed for four-and-a-half years after pleading guilty to robbery and possession of a bladed article at Birmingham Crown Court. *Khan,* the case was not solved until the 35-year-old walked into a police station earlier this year. Jamie Robbins had felt he could not cope outside of prison and wanted to go back.
GAN: Jamie Robbins stole £1,600 from the Select 'n' Save shop in Kings Norton, Birmingham. ~~(The case was not solved ...)~~ He was jailed for four-and-a-half years after pleading guilty to robbery and possession of a bladed article at Birmingham Crown Court on Thursday. Robbins discovered his girlfriend was pregnant and no longer wanted to be given a custodial sentence.
Pointer-generator: Jamie Robbins stole £1,600 from the Select 'n' Save shop in Birmingham. *He stole £1,600 from a convenience store in Birmingham* with accomplice because he preferred life in jail. Robbins was jailed for four-and-a-half years after pleading guilty to robbery and possession of a bladed article at Birmingham Crown Court .
Our unified model (with inconsistency loss): Jamie Robbins stole £1,600 from the Select 'n' Save shop in Kings Norton, Birmingham, with an accomplice on January 13, 2014. The case was not solved until the 35-year-old walked into a police station earlier this year. Robbins was jailed for four-and-a-half years after pleading guilty to robbery and possession of a bladed article.

Figure 1: Typical Comparison. Our model attended at the most important information (blue bold font) matching well with the reference summary; while other state-of-the-art methods miss (green strikethrough font) or generate repeated/wrong information (red italic font).

Original article (truncated):

McDonald's says it's simplifying its grilled chicken recipe to remove ingredients people might not recognize, marking the latest sign the company is rethinking its menu to keep up with changing tastes. The company **says it expects the new 'Artisan Grilled Chicken' to be in its more than 14,300 U.S. stores by the end of next week, in products including a new sandwich, as well as existing sandwiches, wraps and salads. It says the biggest change is the removal of sodium phosphates**, which it said was used to keep the chicken moist, in favor of vegetable starch. **The new recipe also does not use maltodextrin, which McDonald's said is generally used as a sugar to increase browning or as a carrier for seasoning.** Jessica Foust, director of culinary innovation at McDonald's, said the changes were made because customers said they want 'simple, clean ingredients' they are familiar with. 'Maltodextrin is just not something they have in their pantries,' Foust said. Easy to follow: McDonald's is simplifying its grilled chicken recipe to remove ingredients people might not recognize. Its 'Artisan Grilled Chicken' will be included in a new sandwich and existing menu items. The change comes as McDonald's fights to hold onto customers amid the growing popularity of places like Chipotle that position themselves as more wholesome alternatives to traditional fast-food. Already, that 'clean label' trend has prompted numerous restaurant chains and packaged food makers to reformulate products, even while standing by the safety and quality of their previous recipes. Subway also introduced a new grilled chicken recipe earlier this year that it said had no artificial flavors or preservatives. And *Panera Bread has said it plans to purge artificial colors, flavors and preservatives from its food by 2016.* As ingredient quality becomes a more powerful marketing advantage, executives at McDonald's have also been trying to freshen up the company's image and shake perceptions that it serves junk food. TV ads and signs in stores, for instance, are playing up the fact that Egg McMuffins are made with freshly cracked eggs. And last month, McDonald's announced it would start asking suppliers to curb the use of antibiotics in raising chickens. (...)

Reference summary:

The company says it expects the new 'Artisan Grilled Chicken' to be in its more than 14,300 U.S. stores by the end of next week. It says the biggest change is the removal of sodium phosphates, which it said was used to keep the chicken moist, in favor of vegetable starch. The new recipe also does not use maltodextrin, which McDonald's said is generally used as a sugar to increase browning or as a carrier for seasoning.

DeepRL:

McDonald says it expects new 'Artisan Grilled Chicken' to be in its more than 14,300 U.S. stores by the end of next week, in products including a new sandwich, as existing sandwiches, wraps and salads. It says the biggest change is the removal of sodium phosphates. The new recipe also does not use maltodextrin, which it says it is used to keep the chicken moist, in favor of vegetable starch. *The McDonald's says it will start asking suppliers to curb the use of antibiotics in raising chickens.*

GAN:

The company says it expects the new 'Artisan Grilled Chicken' to be in its more than 14,300 U.S. stores by the end of next week, in products including a new sandwich, as well as existing sandwiches, wraps and salads. The new recipe also does not use maltodextrin, which *McDonald's said it plans to purge artificial colors, flavors and preservatives from its food by 2016.*

Pointer-generator:

McDonald's says it expects the new 'Artisan Grilled Chicken' to be in its more than 14,300 U.S. stores by the end of next week. The company says the changes were made because customers said they want 'simple, clean ingredients' they are familiar with. *McDonald's said it plans to purge artificial colors, flavors and preservatives from its food by 2016.*

Our unified model (with inconsistency loss):

McDonald's says it expects the new 'Artisan Grilled Chicken' to be in its more than 14,300 U.S. stores by the end of next week, in products including a new sandwich, as well as existing sandwiches, wraps and salads. It says the biggest change is the removal of sodium phosphates. The new recipe also does not use maltodextrin, which McDonald's said is generally used as a sugar to increase browning or as a carrier for seasoning.

Figure 2: Typical Comparison. Our model generates a coherent summary that contains the most important information (blue bold font) matching well with the reference summary; while other state-of-the-art methods make mistakes or generate less important information (red italic font).

Original article (truncated):

The **sheriff who had former New England Patriots player Aaron Hernandez in custody for more than 18 months said** Tuesday that **he's a master manipulator and will probably do fine in prison now that he has been sentenced to life for murder. Bristol County Sheriff** Thomas Hodgson **said Hernandez knows how to use his charm and manipulate better than anyone he has ever seen**, adding that the former football star is generally affable and polite and would try to use those qualities to get what he wanted at the Bristol County House of Corrections. 'He would make every effort to get extra sandwiches,' Hodgson said. 'He would just try to convince the officers to give him more than what they otherwise could get.' Scroll down for video. Former New England Patriots NFL football player Aaron Hernandez during his trial for murder (file photo) Staff members were directed to treat Hernandez like any other inmate, Hodgson said. **Hernandez was convicted April 15 of the 2013 killing of Odin Lloyd, who was dating the sister of Hernandez's fiancee.** He will ultimately end up at the maximum-security Souza-Baranowski state prison. A lawyer for Hernandez did not immediately return messages seeking comment on the sheriff's remarks. After the verdict was delivered, staff from Bristol County brought Hernandez to a state prison not far from Gillette Stadium, where Hernandez used to play football in front of tens of thousands of cheering fans. There, he told them, 'I'll miss you guys, but they got it wrong,' Hodgson said. 'He didn't really have much of a change in his demeanor. He pretty much still had a swagger in his step.' Hodgson said the 25-year-old Hernandez has an ability to compartmentalize things and lock out the negative. 'He doesn't really look at it as jail,' Hodgson said. 'It's more like training camp.' Hernandez got into trouble from time to time in jail in Bristol County. He was accused of threatening to kill a prison guard and his family, and he got into a fight with another inmate. (...)

Reference summary:

Sheriff Thomas Hodgson who had Aaron Hernandez in custody for 18 months is opening up about the football star. He claims Hernandez is a master manipulator who would charm prison guards and be polite to get what he wants. Hodgson added that he thinks Hernandez will be fine behind bars given his demeanor and attitude. Hernandez was convicted April 15 of the 2013 killing of Odin Lloyd and received a life sentence.

DeepRL:

Hernandez was convicted April 15 of the 2013 killing of Odin Lloyd, who was dating the sister of Hernandez's fiancee. The sheriff Hodgson said he's a master manipulator and will probably do fine in prison. *Hernandez was accused of threatening to kill a prison guard and his family and he got into trouble from time to time in jail.* He has been sentenced to life for murder. *He will ultimately end up at the maximum-security Souza-Baranowski state prison.*

GAN:

Bristol County Sheriff Thomas Hodgson says he's a master manipulator and will probably do fine in prison now that he has been sentenced to life for murder. Hernandez was convicted April 15 of the 2013 killing of Odin Lloyd, who was dating the sister of Hernandez's fiancee. *He will ultimately end up at the maximum-security Souza-Baranowski state prison not far from Gillette Stadium.*

Pointer-generator:

Sheriff Thomas Hodgson said Hernandez knows how to use his charm and manipulate better than anyone he has ever seen. *Staff members were directed to treat Hernandez like any other inmate, Hodgson said.* Hernandez was convicted April 15 of the 2013 killing of Odin Lloyd, who was dating the sister of Hernandez's fiancee. *He will ultimately end up at the maximum-security Souza-Baranowski state prison.*

Our unified model (with inconsistency loss):

Sheriff who had former New England Patriots player Aaron Hernandez in custody for more than 18 months said he's a master manipulator and will probably do fine in prison now that he has been sentenced to life for murder. Bristol County Sheriff said Hernandez knows how to use his charm and manipulate better than anyone he has ever seen. Hernandez was convicted April 15 of the 2013 killing of Odin Lloyd, who was dating the sister of Hernandez's fiancee.

Figure 3: Typical Comparison. Our model attends at the most important information (blue bold font) matching well with the reference summary; while other state-of-the-art methods generate less important information (red italic font).

Original article:

Formula One star Jenson Button completed the London Marathon with an impressive time of two hours, 52 minutes and 30 seconds. The 35-year-old McLaren driver told the BBC he was 'chuffed to bits' with his performance, but joked he was disappointed to finish behind former Olympic rower James Cracknell. He added, 'There was such an amazing atmosphere out there.' Formula One driver Jenson Button celebrates completing the London Marathon 2015 in under three hours. **Button praised the 'amazing atmosphere' and collective spirit of the runners at the event.** Button was running on behalf of the charity Cancer Research and was really buoyed by the good will feelings about the event. 'People have put in so much effort and it's actually quite emotional it was really cool,' he said, before praising the collective spirit of the runners. 'To come here and do a completely different sport with 36,000 other people... it's really really emotional.' Button was running on behalf of Cancer Research and described his experience as 'really really emotional'.

Reference summary:

McLaren driver Jenson Button ran the London Marathon 2015. He finished with a time of 2 hours 52 minutes and 30 seconds. Button praised the 'amazing atmosphere' and collective spirit of the runners at the event.

DeepRL:

~~(Jenson Button completed the London Marathon with ...)~~ Formula One star Button was 'chuffed to bits' with his performance, but joked he was disappointed to finish behind former Olympic rower James Cracknell. The 35-year-old was running on behalf of the charity Cancer Research. *McLaren driver said he was 'awesome' and his performance.* ~~(Button praised the 'amazing atmosphere' and collective spirit ...)~~

GAN:

Jenson Button completed the London Marathon with an impressive time of two hours, 52 minutes and 30 seconds. The 35-year-old Mclaren driver told the BBC he was 'chuffed to bits' with his performance, but joked he was disappointed to finish behind former Olympic rower James Cracknell. ~~(Button praised the 'amazing atmosphere' and collective spirit ...)~~

Pointer-generator:

Jenson Button completed the London Marathon 2015 in under three hours. The 35-year-old told the BBC he was 'chuffed to bits' with his performance. But joked he was disappointed to finish behind former Olympic rower James Cracknell. ~~(Button praised the 'amazing atmosphere' and collective spirit ...)~~

Our unified model (with inconsistency loss):

Formula one star Jenson Button completed the London Marathon with an impressive time of two hours, 52 minutes and 30 seconds. The 35-year-old driver told the BBC he was 'chuffed to bits' with his performance. Button praised the 'amazing atmosphere' and collective spirit of the runners at the event.

Figure 4: Typical Comparison. Our model generates a coherent summary that contains the most important information (blue bold font) matching well with the reference summary; while other state-of-the-art methods make mistakes (red italic font) or miss important information (green strikethrough font).

Original article (truncated):

Tottenham manager Mauricio Pochettino **has likened England hot-shot Harry Kane to Gabriel Batistuta** and insisted he will be there to help him when things do not go to plan. **Kane moved to within two strikes of the 30-goal barrier for the season** after scoring on his full international debut some 79 seconds after coming on against Lithuania at Wembley last Friday night. *England boss Roy Hodgson* **is mindful of keeping expectations in check for Kane**, who has enjoyed a meteoric rise after finally breaking into the Spurs first-team following loan spells at the likes of Leicester and Norwich. Tottenham manager Mauricio Pochettino has compared Harry Kane to Gabriel Batistuta. Batistuta was an Argentina international who also played for Italian club Roma. Pochettino has said he will be there to help Kane when things are not going to plan for him. Pochettino (right) played alongside Batistuta with the Argentina international team. Former Argentina international Pochettino knows lows can come just as quickly as highs, so will be ready to offer Kane advice whenever it is needed. 'Harry Kane has skills that are similar to different players, including Batistuta. The challenge is doing it all again next season,' said Pochettino, who faced England in the 2002 World Cup finals in the same team as the renowned Argentina forward. 'Now it is up to him. Harry knows the way to improve his ability. He is in a very good moment to keep working hard and develop his game.' 'You need to wait for the right moment to give a player the responsibility because sometimes we can see a player is ready, but not quite complete to assume the responsibility and keep the same level.' 'Big players always have a good first season - like (Wayne) Rooney, after his first season a lot of people maybe had doubts about the next one, but this is football.' 'It is not easy when the player goes down, this is our challenge, but not only with Harry, with all players.' 'There may be people who doubt him about next season, but the people that know him understand, like us, know that Harry can keep the level he has shown.' **England boss Hodgson wants Kane to continue his international development with the Under 21s this summer**, after helping Gareth Southgate's team qualify unbeaten for the June tournament in Portugal. (...)

Reference summary:

Harry Kane is approaching 30 goals for the season in all competitions. Mauricio Pochettino says Kane is similar to Gabriel Batistuta. Roy Hodgson wants Kane to play for the England Under 21s in the summer.

DeepRL:

Kane moved to within two strikes of the 30-goal barrier for the season. *England boss Mauricio Pochettino* has likened England hot-shot Kane to Gabriel Batistuta. Tottenham manager Mauricio Pochettino has said he will be there to help him when things do not go to plan. *Tottenham boss Roy Hodgson* is mindful of keeping expectations in check for Kane. ~~Roy Hodgson wants Kane to continue his international development with the Under 21s this summer.~~

GAN:

Tottenham manager Mauricio Pochettino has likened England hot-shot Harry Kane to Gabriel Batistuta. England boss Roy Hodgson is mindful of keeping expectations in check for Kane. Kane has enjoyed a meteoric rise after finally breaking into the spurs first-team. ~~Roy Hodgson wants Kane to continue his international development with the Under 21s this summer.~~

Pointer-generator:

Tottenham manager Mauricio Pochettino has likened England hot-shot Harry Kane to Gabriel Batistuta and insisted he will be there to help him when things are not going to plan. Kane moved to within two strikes of the 30-goal barrier for the season after scoring on his full international debut some 79 seconds after coming on against Lithuania at Wembley last Friday night. England boss Roy Hodgson is mindful of keeping expectations in check for Kane, who has enjoyed a meteoric rise. ~~Roy Hodgson wants Kane to continue his international development with the Under 21s this summer.~~

Our unified model (with inconsistency loss):

Mauricio Pochettino likened England hot-shot Harry Kane to Gabriel Batistuta. Kane moved to within two strikes of the 30-goal barrier for the season. England boss Roy Hodgson is mindful of keeping expectations in check for Kane. England boss Hodgson wants Kane to continue his international development with the Under 21s this summer.

Figure 5: Typical Comparison. Our model attends the most important information (blue bold font) matching well with the reference summary; while other state-of-the-art methods make mistakes (red italic font) or miss important information (green strikethrough font).

A.2 Typical Examples from Non-news Articles

To further evaluate the generalization of our summarization model, we give some examples using non-news articles as inputs. In the following, we test our model with some introductions of the papers. We use **blue bold font** to highlight the fragments that our summary attends in the article.

Sine the model is trained on CNN/Daily Mail dataset, which contains all news articles, our model tends to focus on the first two paragraphs of the article while the introduction of a paper usually places the importance at the end. This problem can be solved by finetuning our model on the data of the new domain. For instance, for the following examples, we can finetune our model on the paired data of introductions and abstracts of the papers.

Original introduction (Vinyals et al., 2015):

Being able to automatically describe the content of an image using properly formed **English sentences** is a very challenging task, but it could have great impact, for instance by helping visually impaired people better understand the content of images on the web. **This task is significantly harder, for example, than the well-studied image classification or object recognition tasks**, which **have been a main focus in the computer vision community [27]**. Indeed, a description must capture not only the objects contained in an image, but it also must express how these objects relate to each other as well as their attributes and the activities they are involved in. Moreover, the above semantic knowledge has to be expressed in a natural language like English, which means that a language model is needed in addition to visual understanding.

Most previous attempts have proposed to stitch together existing solutions of the above sub-problems, in order to go from an image to its description [6, 16]. In contrast, we would like to present in this work a single joint model that takes an image I as input, and is trained to maximize the likelihood $p(S|I)$ of producing a target sequence of words $S = \{S_1, S_2, ...\}$ where each word S_t comes from a given dictionary, that describes the image adequately.

The main inspiration of our work comes from recent advances in machine translation, where the task is to transform a sentence S written in a source language, into its translation T in the target language, by maximizing $p(T|S)$. For many years, machine translation was also achieved by a series of separate tasks (translating words individually, aligning words, reordering, etc), but recent work has shown that translation can be done in a much simpler way using Recurrent Neural Networks (RNNs) [3, 2, 30] and still reach state-of-the-art performance. An encoder RNN reads the source sentence and transforms it into a rich fixed-length vector representation, which in turn in used as the initial hidden state of a decoder RNN that generates the target sentence.

Here, we propose to follow this elegant recipe, replacing the encoder RNN by a deep convolution neural network (CNN). Over the last few years it has been convincingly shown that CNNs can produce a rich representation of the input image by embedding it to a fixed-length vector, such that this representation can be used for a variety of vision tasks [28]. Hence, it is natural to use a CNN as an image encoder, by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN decoder that generates sentences (see Fig. 1). We call this model the Neural Image Caption, or NIC.

Our contributions are as follows. First, we present an end-to-end system for the problem. It is a neural net which is fully trainable using stochastic gradient descent. Second, our model combines state-of-art sub-networks for vision and language models. These can be pre-trained on larger corpora and thus can take advantage of additional data. Finally, it yields significantly better performance compared to state-of-the-art approaches; for instance, on the Pascal dataset, NIC yielded a BLEU score of 59, to be compared to the current state-of-the-art of 25, while human performance reaches 69. On Flickr30k, we improve from 56 to 66, and on SBU, from 19 to 28.

Original abstract (Vinyals et al., 2015):

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In this paper, we present a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. Our model is often quite accurate, which we verify both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 score (the higher the better) on the Pascal dataset is 25, our approach yields 59, to be compared to human performance around 69. We also show BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, we achieve a BLEU-4 of 27.7, which is the current state-of-the-art.

Our unified model (with inconsistency loss):

English sentences have been a main focus in the computer vision community [27].

This task is significantly harder, for example, than the well-studied image classification or object recognition tasks.

Most previous attempts have proposed to stitch together existing solutions of the above sub-problems, in order to go from an image to its description.

Figure 6: Our model attends at the first two paragraphs (blue bold font). It mainly describes the task of image captioning and the method of previous works.

Original introduction (Gehring et al., 2017):

Sequence to sequence learning has been successful in many tasks such as machine translation, speech recognition (Sutskever et al., 2014; Chorowski et al., 2015) and text summarization (Rush et al., 2015; Nallapati et al., 2016; Shen et al., 2016) amongst others. **The dominant approach to date encodes the input sequence with a series of bi-directional recurrent neural networks (RNN) and generates a variable length output** with another set of decoder RNNs, both of which interface via a soft-attention mechanism (Bahdanau et al., 2014; Luong et al., 2015). In machine translation, this architecture has been demonstrated to outperform traditional phrase-based models by large margins (Sennrich et al., 2016b; Zhou et al., 2016; Wu et al., 2016).

Convolutional neural networks are less common for sequence modeling, despite several advantages (Waibel et al., 1989; LeCun & Bengio, 1995). Compared to recurrent layers, convolutions create representations for fixed size contexts, however, **the effective context size of the network can easily be made larger by stacking several layers on top of each other. This allows to precisely control the maximum length of dependencies to be modeled. Convolutional networks do not depend on the computations of the previous time step and** therefore **allow parallelization over every element in a sequence.** This contrasts with RNNs which maintain a hidden state of the entire past that prevents parallel computation within a sequence.

Multi-layer convolutional neural networks create hierarchical representations over the input sequence in which nearby input elements interact at lower layers while distant elements interact at higher layers. Hierarchical structure provides a shorter path to capture long-range dependencies compared to the chain structure modeled by recurrent networks, e.g. we can obtain a feature representation capturing relationships within a window of n words by applying only O(nk) convolutional operations for kernels of width k, compared to a linear number O(n) for recurrent neural networks. Inputs to a convolutional network are fed through a constant number of kernels and non-linearities, whereas recurrent networks apply up to n operations and non-linearities to the first word and only a single set of operations to the last word. Fixing the number of nonlinearities applied to the inputs also eases learning.

Recent work has applied convolutional neural networks to sequence modeling such as Bradbury et al. (2016) who introduce recurrent pooling between a succession of convolutional layers or Kalchbrenner et al. (2016) who tackle neural translation without attention. However, none of these approaches has been demonstrated improvements over state of the art results on large benchmark datasets. Gated convolutions have been previously explored for machine translation by Meng et al. (2015) but their evaluation was restricted to a small dataset and the model was used in tandem with a traditional count-based model. Architectures which are partially convolutional have shown strong performance on larger tasks but their decoder is still recurrent (Gehring et al., 2016).

In this paper we propose an architecture for sequence to sequence modeling that is entirely convolutional. Our model is equipped with gated linear units (Dauphin et al., 2016) and residual connections (He et al., 2015a). We also use attention in every decoder layer and demonstrate that each attention layer only adds a negligible amount of overhead. The combination of these choices enables us to tackle large scale problems.

We evaluate our approach on several large datasets for machine translation as well as summarization and compare to the current best architectures reported in the literature. On WMT16 English-Romanian translation we achieve a new state of the art, outperforming the previous best result by 1.9 BLEU. On WMT14 English-German we outperform the strong LSTM setup of Wu et al. (2016) by 0.5 BLEU and on WMT14 English-French we outperform the likelihood trained system of Wu et al. (2016) by 1.6 BLEU. Furthermore, our model can translate unseen sentences at an order of magnitude faster speed than Wu et al. (2016) on GPU and CPU hardware.

Original abstract (Gehring et al., 2017):

The prevalent approach to sequence to sequence learning maps an input sequence to a variable length output sequence via recurrent neural networks. We introduce an architecture based entirely on convolutional neural networks. Compared to recurrent models, computations over all elements can be fully parallelized during training to better exploit the GPU hardware and optimization is easier since the number of non-linearities is fixed and independent of the input length. Our use of gated linear units eases gradient propagation and we equip each decoder layer with a separate attention module. We outperform the accuracy of the deep LSTM setup of Wu et al. (2016) on both WMT14 English-German and WMT14 English-French translation at an order of magnitude faster speed, both on GPU and CPU.

Our unified model (with inconsistency loss):

The dominant approach to date encodes the input sequence with a series of bi-directional recurrent neural networks (RNN) and generates a variable length output.

The effective context size of the network can easily be made larger by stacking several layers on top of each other.

This allows to precisely control the maximum length of dependencies to be modeled.

Convolutional networks do not depend on the computations of the previous time step and allow parallelization over every element in a sequence.

Figure 7: Our model attends at the first two paragraphs (blue bold font). Though the first two sentences that our model outputs are not very coherent, in the last two sentences, it successfully mentions the benefits of parallelization in convolutional neural networks which is the main contribution of this paper.

Original introduction (Bahdanau et al., 2014):

Neural machine translation is a newly emerging approach to machine translation, recently proposed by Kalchbrenner and Blunsom (2013), Sutskever et al. (2014) and Cho et al. (2014b). **Unlike the traditional phrase-based translation system (see, e.g.,** Koehn et al., 2003) which consists of many small sub-components that are tuned separately, **neural machine translation attempts to build and train a single, large neural network that reads a sentence and outputs a correct translation.**

Most of the proposed neural machine translation models belong to a family of encoder-decoders (Sutskever et al., 2014; Cho et al., 2014a), with an encoder and a decoder for each language, or involve a language-specific encoder applied to each sentence whose outputs are then compared (Hermann and Blunsom, 2014). An encoder neural network reads and encodes a source sentence into a fixed-length vector. A decoder then outputs a translation from the encoded vector. The whole encoder-decoder system, which consists of the encoder and the decoder for a language pair, is jointly trained to maximize the probability of a correct translation given a source sentence.

A potential issue with this encoder-decoder approach is that a neural network needs to be able to compress all the necessary information of a source sentence into a fixed-length vector. This may make it difficult for the neural network to cope with long sentences, especially those that are longer than the sentences in the training corpus. Cho et al. (2014b) showed that indeed the performance of a basic encoder-decoder deteriorates rapidly as the length of an input sentence increases.

In order to address this issue, we introduce an extension to the encoder-decoder model which learns to align and translate jointly. Each time the proposed model generates a word in a translation, it (soft-)searches for a set of positions in a source sentence where the most relevant information is concentrated. The model then predicts a target word based on the context vectors associated with these source positions and all the previous generated target words.

The most important distinguishing feature of this approach from the basic encoder-decoder is that it does not attempt to encode a whole input sentence into a single fixed-length vector. Instead, it encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively while decoding the translation. This frees a neural translation model from having to squash all the information of a source sentence, regardless of its length, into a fixed-length vector. We show this allows a model to cope better with long sentences.

In this paper, we show that the proposed approach of jointly learning to align and translate achieves significantly improved translation performance over the basic encoder-decoder approach. The improvement is more apparent with longer sentences, but can be observed with sentences of any length. On the task of English-to-French translation, the proposed approach achieves, with a single model, a translation performance comparable, or close, to the conventional phrase-based system. Furthermore, qualitative analysis reveals that the proposed model finds a linguistically plausible (soft-)alignment between a source sentence and the corresponding target sentence.

Original abstract (Bahdanau et al., 2014):

Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder-decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder-decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.

Our unified model (with inconsistency loss):

Neural machine translation is a newly emerging approach to machine translation.

Unlike the traditional phrase-based translation system (see, e.g., neural machine translation attempts to build and train a single, large neural network that reads a correct translation.

Most of the proposed neural machine translation models belong to a family of encoder decoders (sutskever et al., 2014; cho et al., 2014a, with an encoder and a decoder for each language.

Figure 8: Our model attends at the first two paragraphs (blue bold font) and generates the summary matching will with the first three sentences in the abstract.

A.3 Human Evaluation User Interface

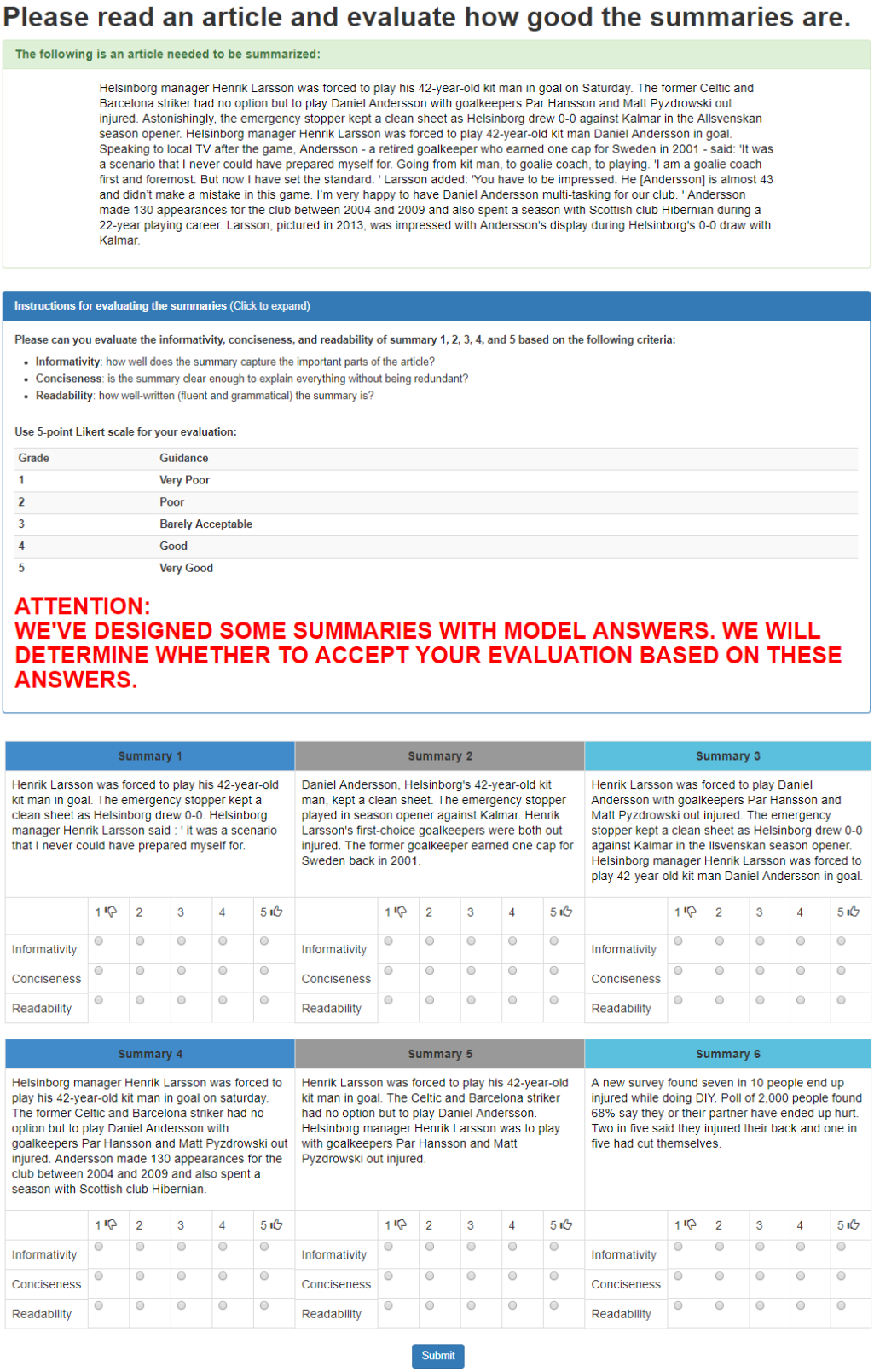

Figure 9: Our human evaluation user interface. The human evaluators are asked to read one article along with six summaries. There is a trap summary (i.e., a random summary which belongs to another article) among the six summaries. We will remove the evaluation which gives the trap summary a high informativity score.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 2015 International Conference on Learning Representations (ICLR)*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning (ICML)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.