arXiv:1701.02854v3 [cs.CL] 9 Feb 2017

Decoding as Continuous Optimization in Neural Machine Translation

Cong Duy Vu Hoang
University of Melbourne
Melbourne, VIC, Australia
vhoang2@student.unimelb.edu.au

Gholamreza Haffari
Monash University
Clayton, VIC, Australia
gholamreza.haffari@monash.edu

Trevor Cohn
University of Melbourne
Melbourne, VIC, Australia
t.cohn@unimelb.edu.au

Abstract

We propose a novel decoding approach for neural machine translation (NMT) based on continuous optimisation. The resulting optimisation problem is then tackled using constrained gradient optimisation. Our powerful decoding framework, enables decoding intractable models such as the intersection of left-to-right and right-to-left (bidirectional) as well as source-to-target and target-to-source (bilingual) NMT models. Our empirical results show that our decoding framework is effective, and leads to substantial improvements in translations generated from the intersected models where the typical greedy or beam search is infeasible.

1 Introduction

Sequence to sequence learning with neural networks (Graves, 2013; Sutskever et al., 2014; Lipton et al., 2015) is typically associated with two phases: training and decoding (*a.k.a.* inference). Model parameters are learned by optimising the training objective, so that the model generalises well when the unknown test data is decoded. The majority of literature have been focusing on developing better training paradigms or network architectures, but the decoding problem is arguably under-investigated. Conventional heuristic-based approaches for approximate inference include greedy, beam, and stochastic search. Greedy and beam search have been empirically proved to be adequate for many sequence to sequence tasks, and are the standard methods for decoding in NMT.

However, these approximate inference approaches have several drawbacks. Firstly, due to sequential decoding of symbols of the target sequence, the inter-dependencies among the target symbols are not fully exploited. For example, when decoding the words of the target sentence in a left-to-right manner, the right context is not exploited leading potentially to inferior performance (see Watanabe and Sumita (2002a) who apply this idea in traditional statistical MT). Secondly, it is not trivial to apply greedy or beam search to decode in NMT models involving global features or constraints, e.g., intersecting left-to-right and right-to-left models which do not follow the same generation order. These global constraints capture different aspects and can be highly useful in producing better and more diverse translations.

We introduce a novel decoding framework (§ 3) that effectively relaxes this *discrete* optimisation problem into a *continuous* optimisation problem. This is akin to linear programming relaxation approach for approximate inference in graphical models with discrete random variables where the exact inference is NP-hard (Sontag, 2010). Our continuous optimisation problems are challenging due to the non-linearirty and non-convexity of the relaxed decoding objective. We make use of stochastic gradient descent (SGD) and exponentiated gradient (EG) algorithms, which are mainly used for training in the literature, for decoding based on our relaxation approach. Our decoding framework is powerful and flexible, as it enables us to decode with global constraints involving intersection of multiple NMT models (§4). We present experimental results on Chinese-English and German-English translation tasks, confirming the effectiveness of our relaxed optimisation method for decoding (§5).[1]

[1] The source code will be released upon publication.