Grammar Variational Autoencoder [4]

Presenter: Emanuele Rossi
Authors: Kusner, Paige, Hernandez-Lobato

University of Cambridge

March 12, 2019

Paper

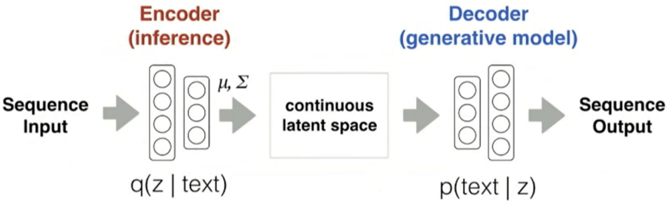
- Published in March 2017
- 91 citations until now

Motivation

- Learning a meaningful latent space for discrete inputs

Previous Approaches: Char-VAE [1] [3]

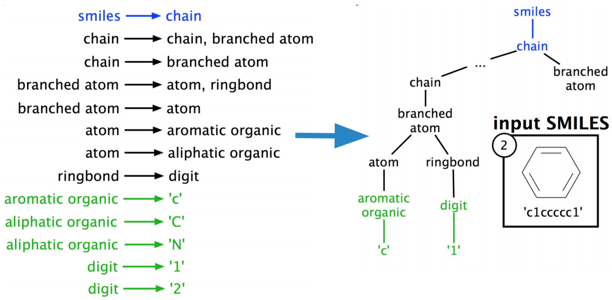
Figure 1: Model from [1]



- Problem: decoder sometimes generates invalid strings

Insight

- Many discrete objects (including molecules) can be described as a parse tree from a context free grammar

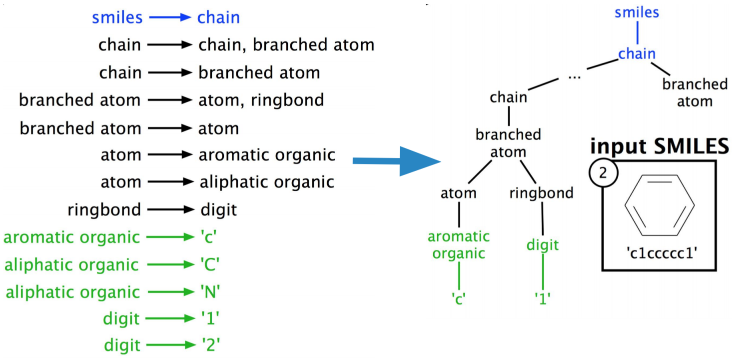


Insight(2)

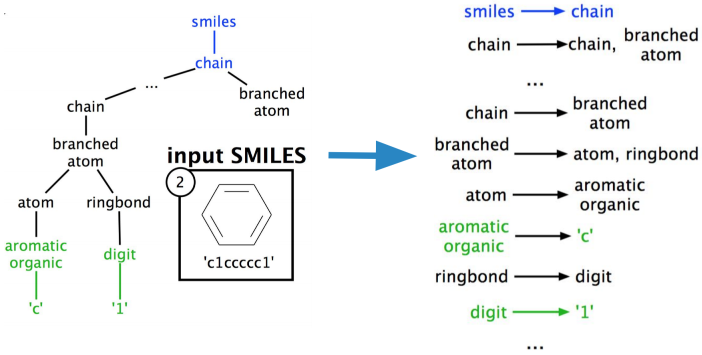
- Encoding and decoding parse trees ensures that all outputs are valid
- It also frees the model from learning ‘syntactic’ rules

Encoder

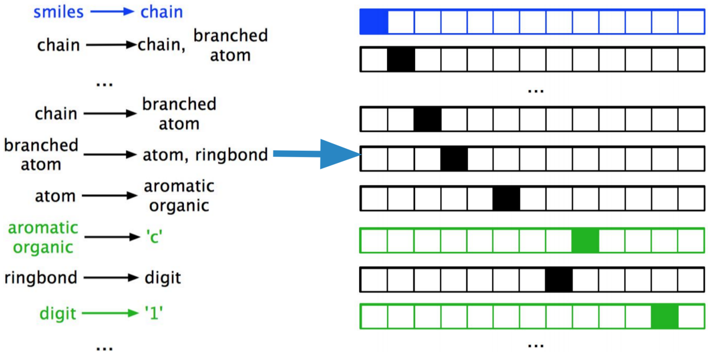
Encoder: molecule to parse tree



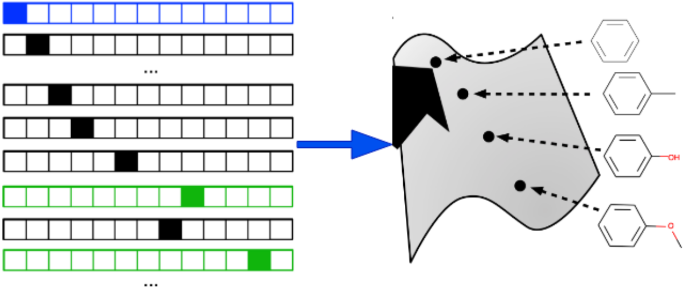
Encoder: parse tree to production rules



Encoder: production rules to one-hot embeddings

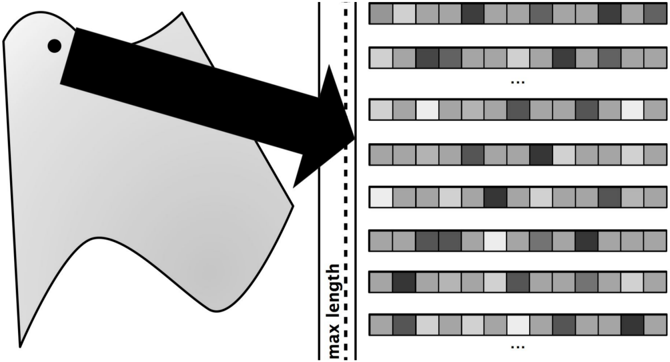


Encoder: one-hot embeddings to latent representation

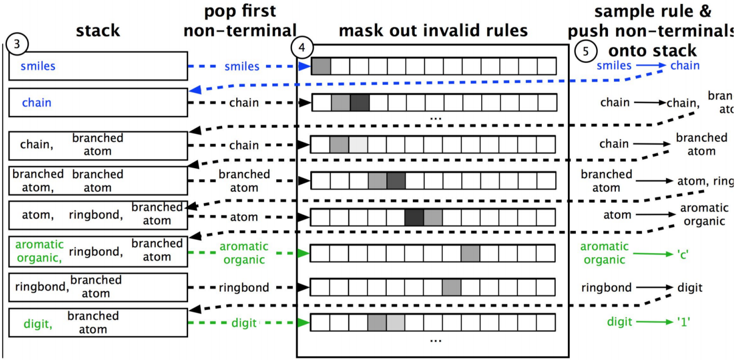


Decoder

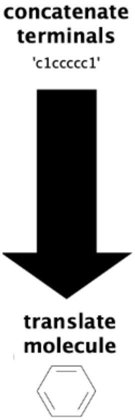
Decoder: one-hot embeddings to latent representation



Decoder: latent representation to logits sequence

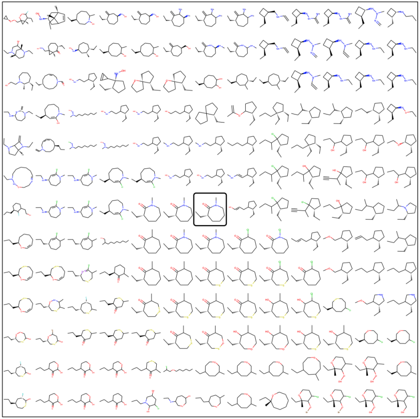


Decoder: from sequence of rules to a molecule



Evaluation

Evaluation 1: Latent Space Visualization



Evaluation 2: Predictive Performance of Latent Representation

Objective	Method	Expressions	Molecules
LL	GVAE	**-1.320±0.001**	**-1.739 ±0.004**
	CVAE	-1.397±0.003	-1.812±0.004
RMSE	GVAE	**0.884 ±0.002**	**1.404 ±0.006**
	CVAE	0.975±0.004	1.504±0.006

Evaluation 3: Optimization in Latent Space

Problem	Method	Frac. valid	Avg. score
Molecules	GVAE	**0.31±0.07**	**-9.57 ±1.77**
	CVAE	0.17±0.05	-54.66±2.66

Figure 2: Fraction of valid molecules found by method

Method	#	SMILE	Score
GVAE	1	CCCc1ccc(I)cc1C1CCC-c1	**2.94**
	2	CC(C)CCCCCc1ccc(Cl)nc1	**2.89**
	3	CCCc1ccc(Cl)cc1CCCCOC	**2.80**
CVAE	1	Cc1ccccc1CCCC1CCC1CCc1nncs1	1.98
	2	Cc1ccccc1CCCC1(COC1)CCc1nnn1	1.42
	3	CCCCCCCCC(CCCC212CCCnC1COC)c122csss1	1.19

Figure 3: Scores obtained by bayesian optimization in latent space

Discussion

Discussion 1: What about Semantic constraint?

- GVAE makes sure the ouput is *syntactically* correct, but what about *semantic* constraints?
- [2] tries to extend GVAE to include semantic constraint

Discussion 2: Evaluation

The paper presents 3 evaluation techniques

- Evaluation on supervised task is standard and not very interesting
- Latent space visualization is just a nice picture
- Optimization in latent space is insightful: what about gradient-based optimization instead of bayesian optimization?

Discussion 3: Encoder vs Decoder

Their model consists of a 1D CNN encoder and a GRU for the decoder.

- Most autoencoders seen so far have the same encoder and decoder
- Does it make a difference whether encoder and decoder are the same type of model, or are different?

Questions?

References

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio.
Generating sentences from a continuous space.
2016.

Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song.
Syntax-directed variational autoencoder for structured data.
CoRR, abs/1802.08786, 2018.

Rafael Gómez-Bombarelli, David K. Duvenaud, José Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik.
Automatic chemical design using a data-driven continuous representation of molecules.
CoRR, abs/1610.02415, 2016.

Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato.
Grammar variational autoencoder.