Ivan Titov · Иван Титов

Reader (≈ US Associate Prof)
Insitute for Language, Cognition and Computation, School of Informatics
University of Edinburgh, UK

UHD (≈ US Associate Prof), part-time
Institute for Logic, Language and Computation, FNWI
University of Amsterdam, Netherlands

Office: IF 3.28
E-mail: ititov (at) inf.ed.ac.uk
Phone: +44 131 6513092

I am an associate professor (Reader) in the Institute for Language, Cognition and Computation (ILCC) at the School of Informatics of the University of Edinburgh. I am also a part-time faculty at the Institute of Logic, Language and Computation of the University of Amsterdam. My research interests are in natural language processing (incl. semantics and syntax) and machine learning.

My research is supported by personal grants (ERC Starting grant and NWO VIDI), as well as industrial funding / collaborations (incl. Google, SAP and Yandex).

I am an action editor for the journal of machine learning research (JMLR), Transactions of ACL (TACL), a member of editorial board of JAIR, an advisory board member for European Chapter of ACL. My other professional services include being a PC co-chair for *SEM 2016 and CoNLL 2018, a senior area chair for ACL 2019, an area chair for at ACL 2016, EMNLP 2014, EACL 2012, ICLR 2017 and 2019 and NIPS 2017, a senior PC member for IJCAI 2011

Ivan Titov · Иван Титов

Reader (≈ US Associate Prof)
Insitute for Language, Cognition and Computation, School of Informatics
University of Edinburgh, UK

UHD (≈ US Associate Prof), part-time
Institute for Logic, Language and Computation, FNWI
University of Amsterdam, Netherlands

Office:	IF 3.28
E-mail:	ititov (at) inf.ed.ac.uk
Phone:	+44 131 6513092

I am an associate professor (Reader) in the Institute for Language, Cognition and Computation (ILCC) at the School of Informatics of the University of Edinburgh. I am also a part-time faculty at the Institute of Logic, Language and Computation of the University of Amsterdam. My research interests are in natural language processing (incl. semantics and syntax) and machine learning.

My research is supported by personal grants (ERC Starting grant and NWO VIDI), as well as industrial funding / collaborations (incl. Google, SAP and Yandex).

I am an action editor for the journal of machine learning research (JMLR), Transactions of ACL (TACL), a member of editorial board of JAIR, an advisory board member for European Chapter of ACL. My other professional services include being a PC co-chair for *SEM 2016 and CoNLL 2018, a senior area chair for ACL 2019, an area chair for at ACL 2016, EMNLP 2014, EACL 2012, ICLR 2017 and 2019 and NIPS 2017, a senior PC member for IJCAI 2011

Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned

Elena Voita
Yandex, Russia
University of Amsterdam, Netherlands
lena-voita@yandex-team.ru

David Talbot
Yandex, Russia
talbot@yandex-team.ru

Fedor Moiseev
Yandex, Russia
Moscow Institute of Physics and Technology, Russia
femoiseev@yandex-team.ru

Rico Sennrich
University of Edinburgh, Scotland
University of Zurich, Switzerland
rico.sennrich@ed.ac.uk

Ivan Titov
University of Edinburgh, Scotland
University of Amsterdam, Netherlands
ititov@inf.ed.ac.uk

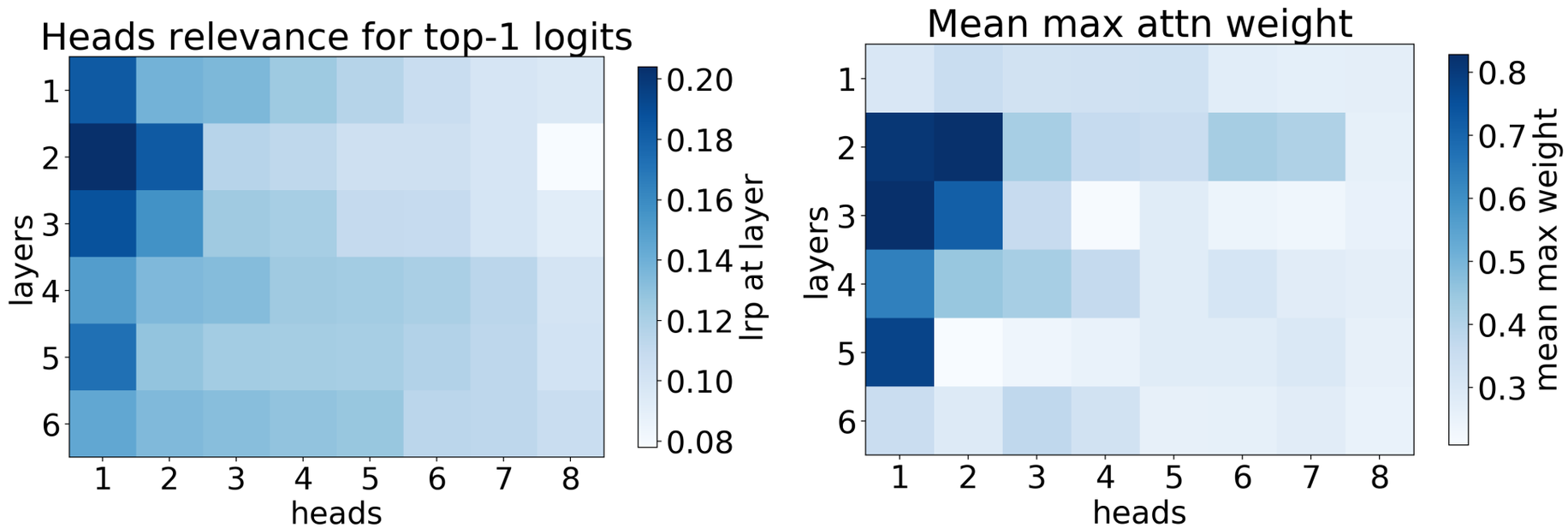
Research Questions

- To what extent does translation quality depend on individual encoder heads?
- Do individual encoder heads play consistent and interpretable roles? If so, which are the most important ones for translation quality?
- Which types of model attention (encoder self-attention, decoder self-attention or decoder-encoder attention) are most sensitive to the number of attention heads and on which layers?
- Can we significantly reduce the number of attention heads while preserving translation quality?

Identify Important Heads

- Confident heads
  - Usually assign a high proportion of its at- tention to a single token
- Layer-wise relevance propagation (LRP)
  - Contribute most to the top-1 logit predicted by the model

Identifying Important Heads

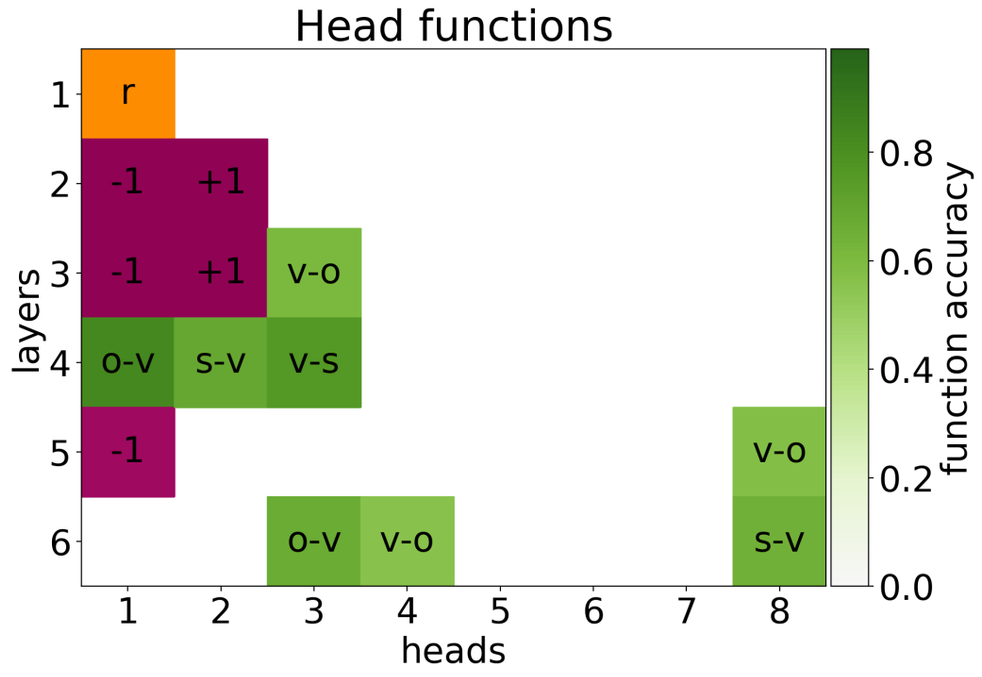


(a) LRP (b) confidence

- The relevance of a head as computed by LRP agrees to a reasonable extent with its confidence.

Characterizing heads

- Positional heads
- Syntactic heads
- Rare word heads



Pruning Attention Heads

- We have identified certain functions of the most relevant heads at each layer and showed that to a large extent they are interpretable

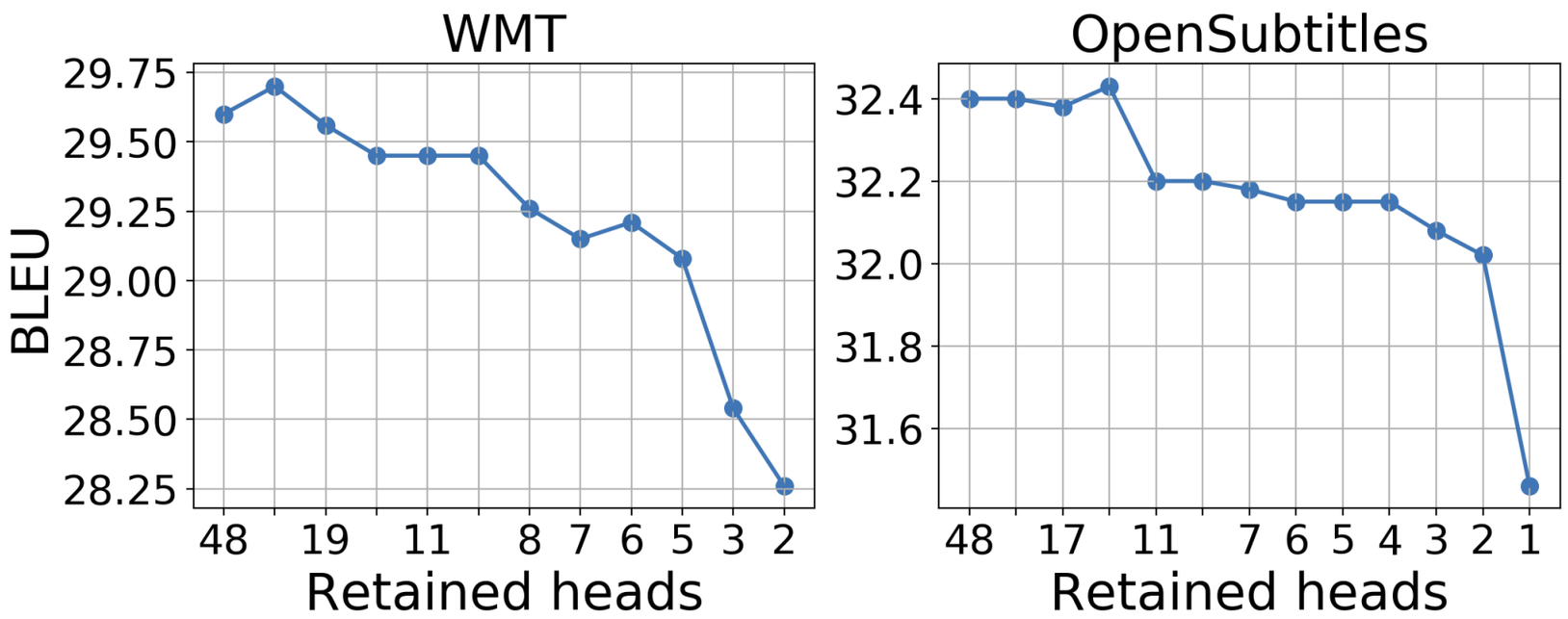
- What of the remaining heads?

L0-norm

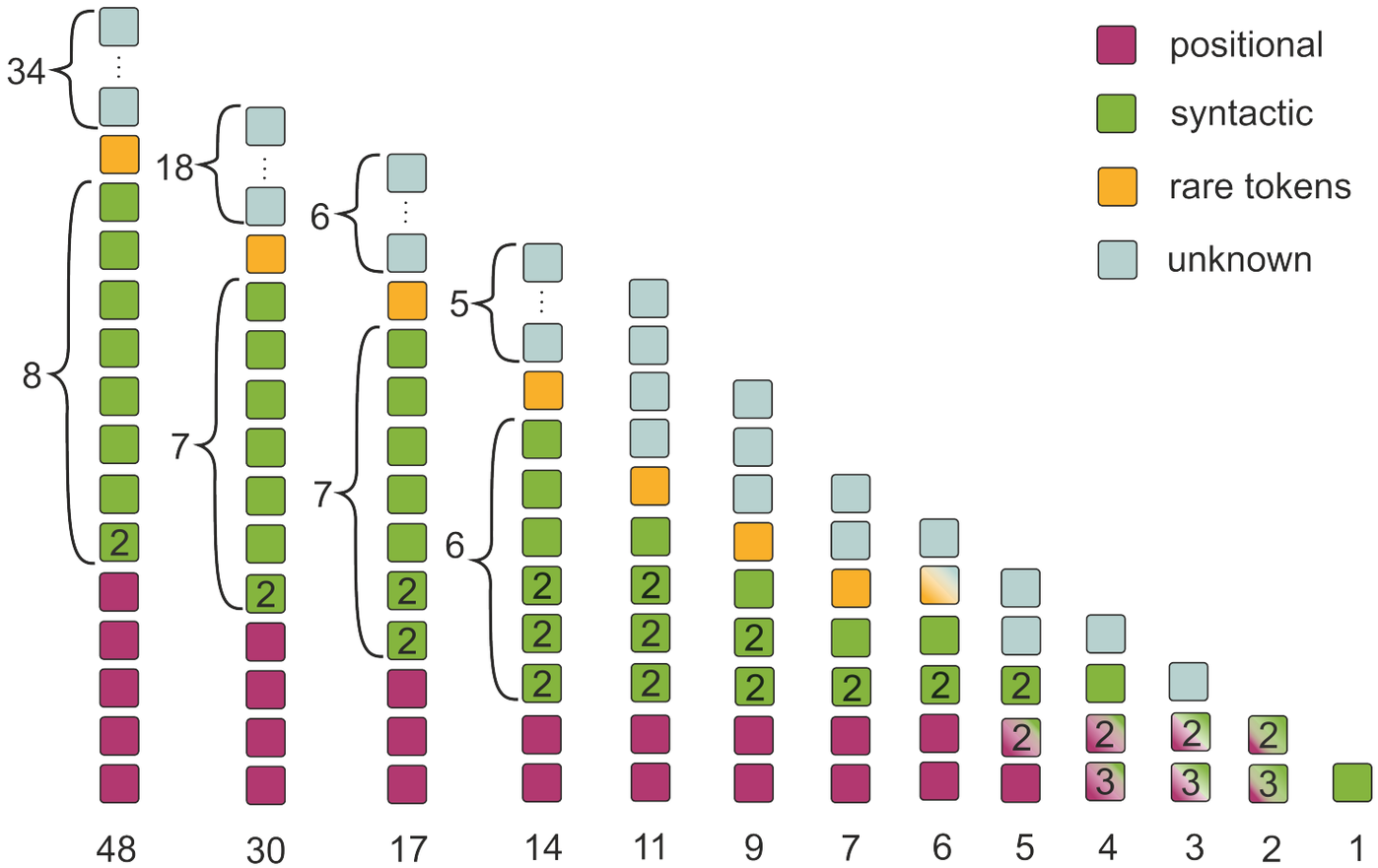
$$\text{MultiHead}(Q, K, V) = \text{Concat}_i(g_i \cdot \text{head}_i)W^O$$

$$L_0(g_1, \ldots, g_h) = \sum_{i=1}^{h}(1 - [[g_i = 0]])$$

Result


WMT
OpenSubtitles
BLEU
29.75
29.50
29.25
29.00
28.75
28.50
28.25
32.4
32.2
32.0
31.8
31.6
48 19 11 8 7 6 5 3 2
48 17 11 7 6 5 4 3 2 1
Retained heads
Retained heads

Results


positional
syntactic
rare tokens
unknown
34
8
18
7
6
7
5
6
2
3
48
30
17
14
11
9
7
6
5
4
3
2
1

Interpretable Neural Predictions with Differentiable Binary Variables

Joost Bastings
ILLC
University of Amsterdam
bastings@uva.nl

Wilker Aziz
ILLC
University of Amsterdam
w.aziz@uva.nl

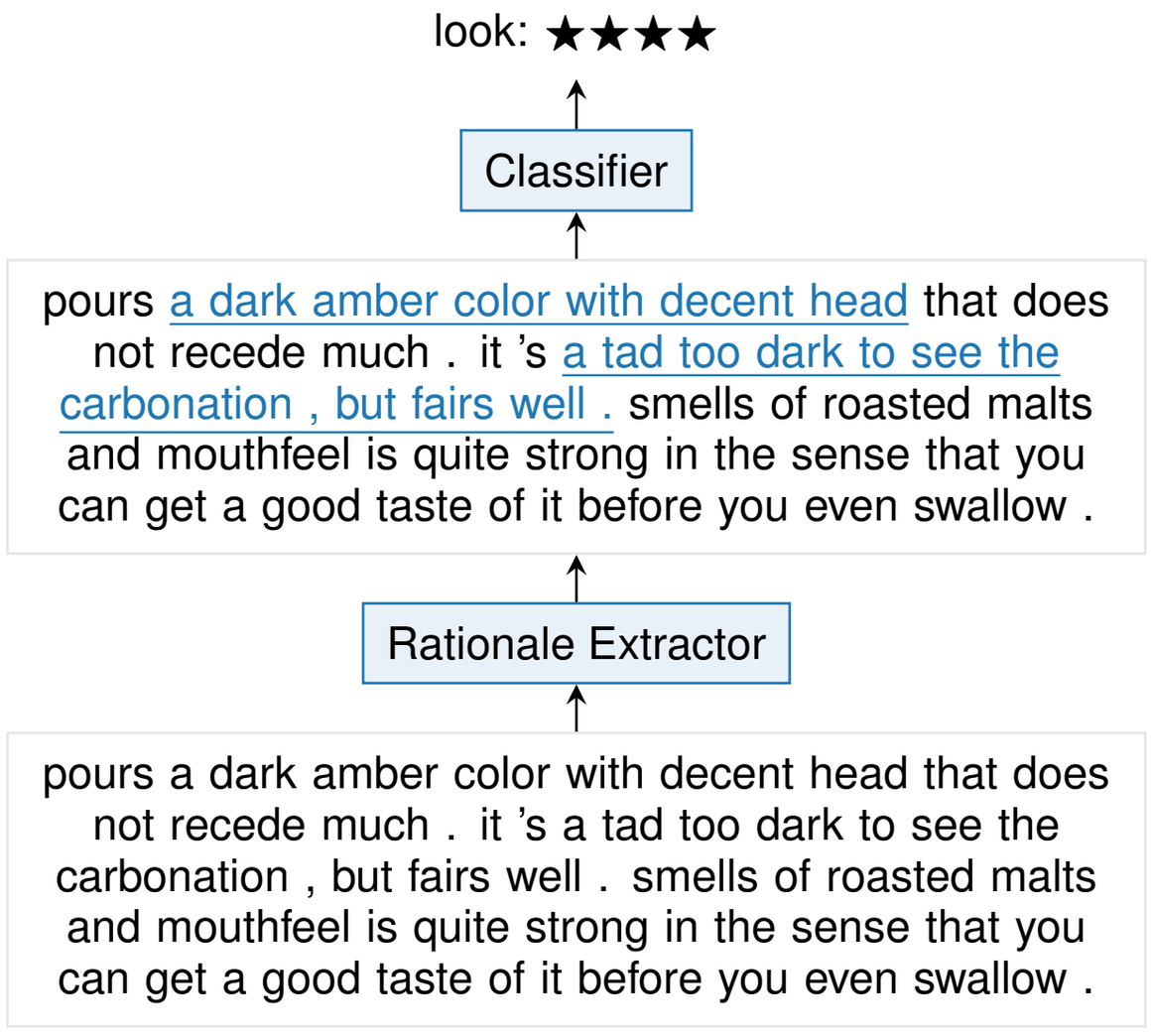
Ivan Titov
ILLC, University of Amsterdam
ILCC, University of Edinburgh
ititov@inf.ed.ac.uk

Rational

- Can we trust neural models?

- What if the model could provide us the most important parts of the document, as a justification for its prediction?

Rationale


look: ★★★★
Classifier
pours a dark amber color with decent head that does not recede much . it 's a tad too dark to see the carbonation , but fairs well . smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow .
Rationale Extractor
pours a dark amber color with decent head that does not recede much . it 's a tad too dark to see the carbonation , but fairs well . smells of roasted malts and mouthfeel is quite strong in the sense that you can get a good taste of it before you even swallow .

L0-norm

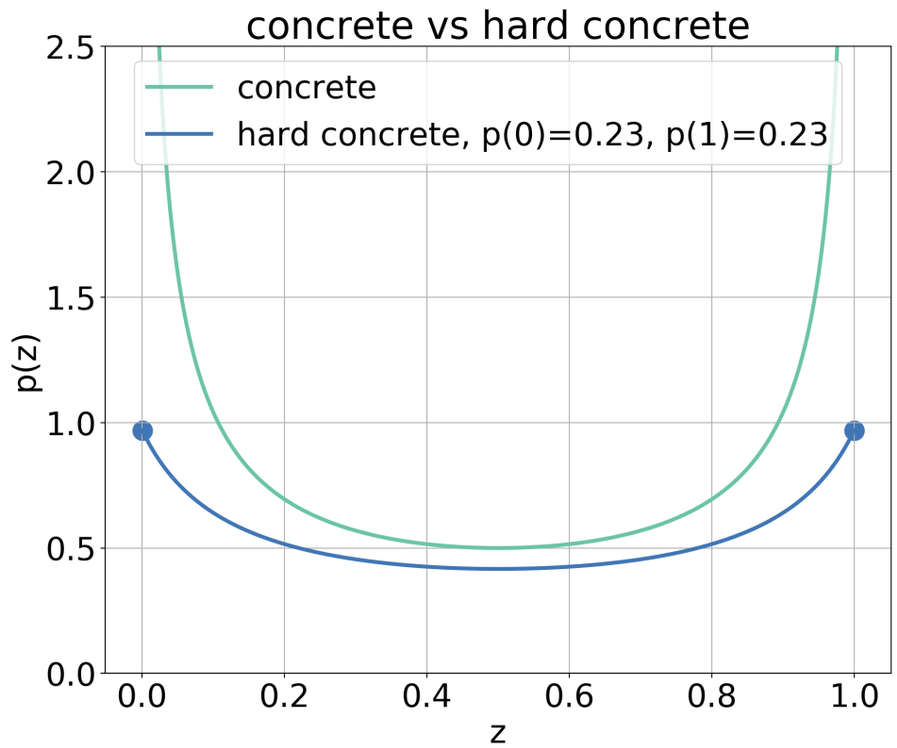
$$Z_i|x \sim \text{Bern}(g_i(x;\phi))$$

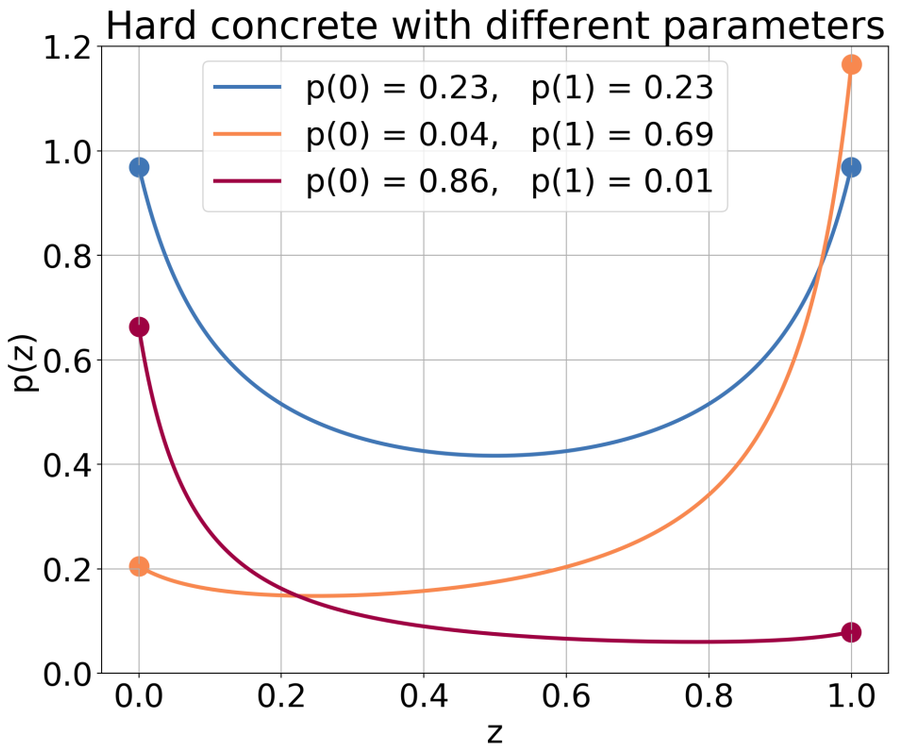
$$Y|x,z \sim \text{Cat}(f(x \odot z;\theta))$$

The Trick

- We start from a distribution over the open interval (0, 1)
  - Closed form solution for P(not zero)
  - Most probability lies on the two ends
- We then *stretch* its support from $l < 0$ to $r > 1$ in order to include {0} and {1}
- We collapse the probability mass over the interval (l, 0] to {0}, and similarly, the probability mass over the interval [1, r) to {1}

Concrete


concrete vs hard concrete
concrete
hard concrete, p(0)=0.23, p(1)=0.23
p(z)
z
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0


Hard concrete with different parameters
p(0) = 0.23, p(1) = 0.23
p(0) = 0.04, p(1) = 0.69
p(0) = 0.86, p(1) = 0.01
p(z)
z
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.2
0.4
0.6
0.8
1.0

Kumaraswamy