One model to learn them all

Can we create a unified deep learning model to solve tasks across multiple domains?

Лебедев Антон

Overview

To English

“A man that is sitting in front of a suitcase”

To Category

Category 127 (Male Human)

“Last week, Kigali raised the possibility of military retaliation after shells...”

To French

“La semaine dernière, Kigali a soulevé la possibilité de représailles militaires aprés avoir débarqué des coquilles...”

“Can you give our readers some details on this?”

To German

“Können Sie unseren Lesern einige Details dazu geben?”

The above represents a triumph of either apathy or civility

To Parse

“S NP DT JJS /NP VP VBZ NP NP DT NN /NP PP IN NP NP NN /NP CC NP NN /NP /NP /PP /NP /VP . /S”

Overview

(1) WSJ speech corpus [7]

(2) ImageNet dataset [23]

(3) COCO image captioning dataset [14]

(4) WSJ parsing dataset [17]

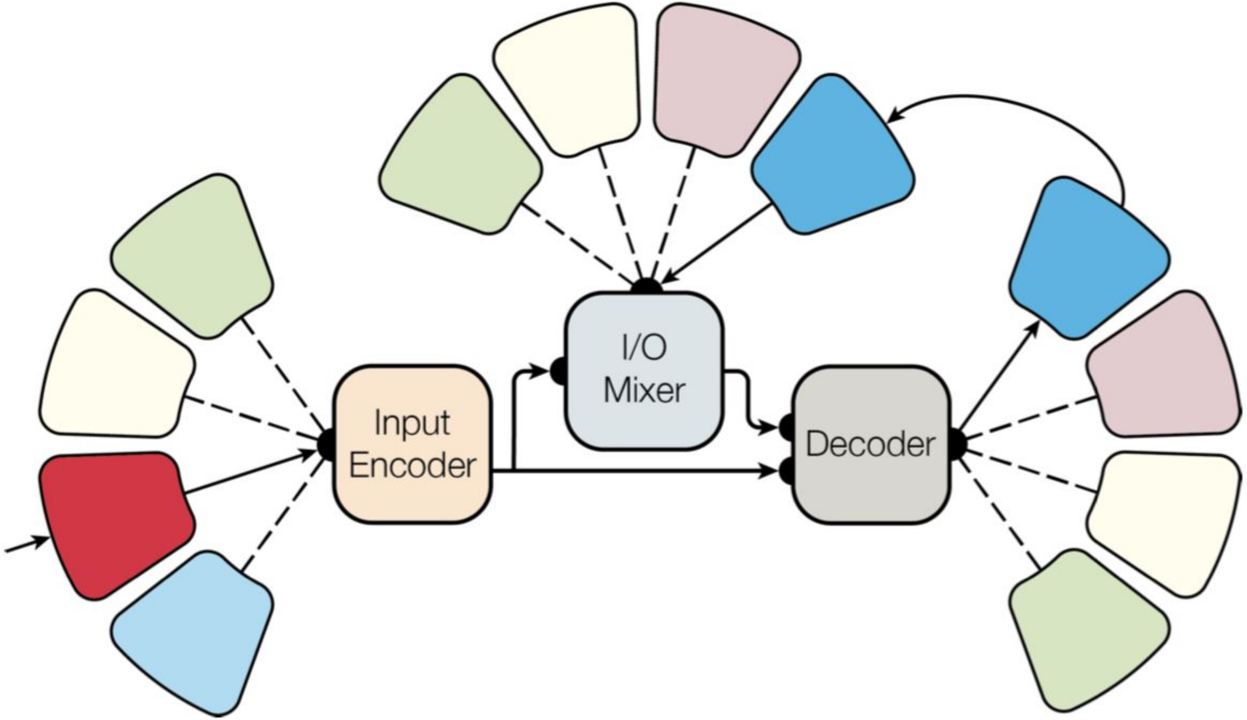
(5) WMT English-German translation corpus

(6) The reverse of the above: German-English translation.

(7) WMT English-French translation corpus

(8) The reverse of the above: German-French translation.

Overview

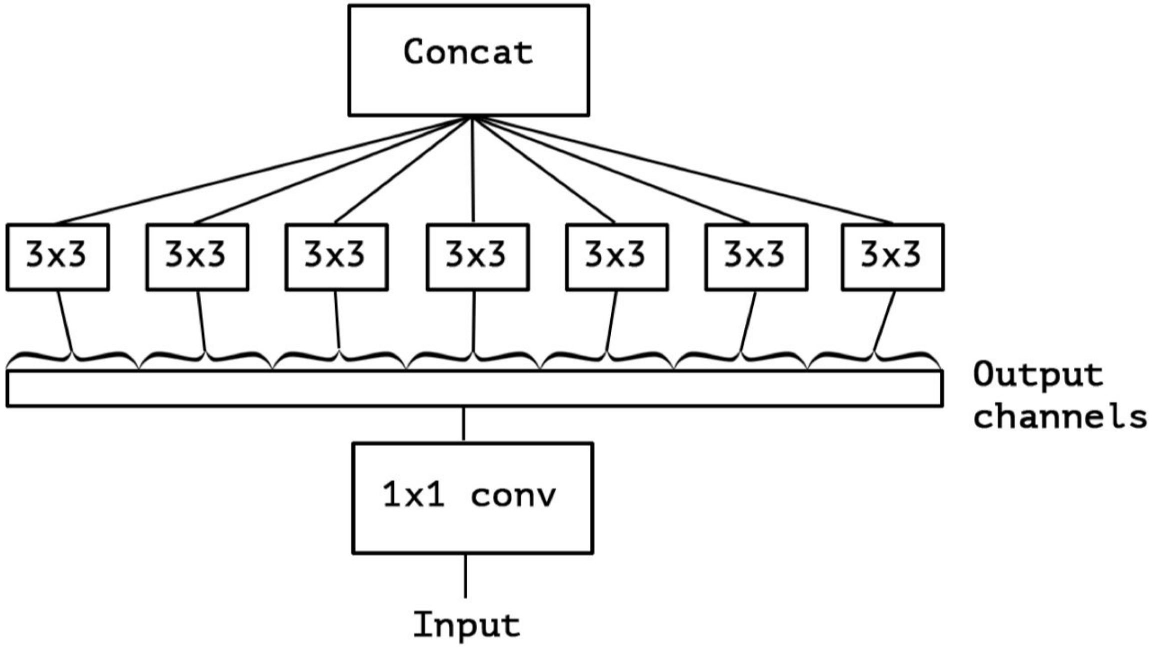


Convolution Blocks

Depthwise separable convolutions

Each channel handles separately

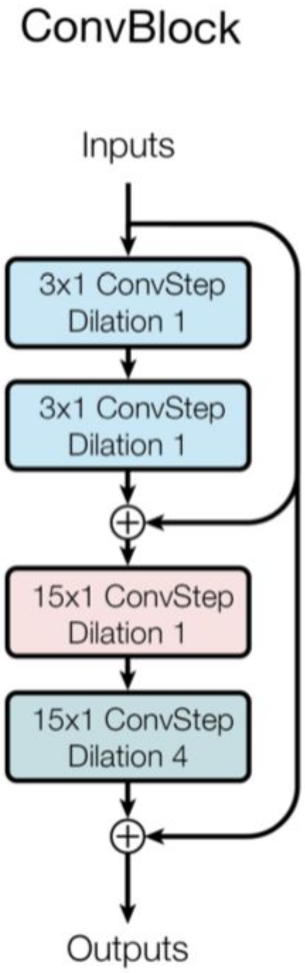
parameter- and computationally-efficient variant of the traditional convolution



Convolution Blocks

$$SepConv_{d,s,f}(W, x)$$

$$ConvStep_{d,s,f}(W, x) = LN(SepConv_{d,s,f}(W, ReLU(x)))$$

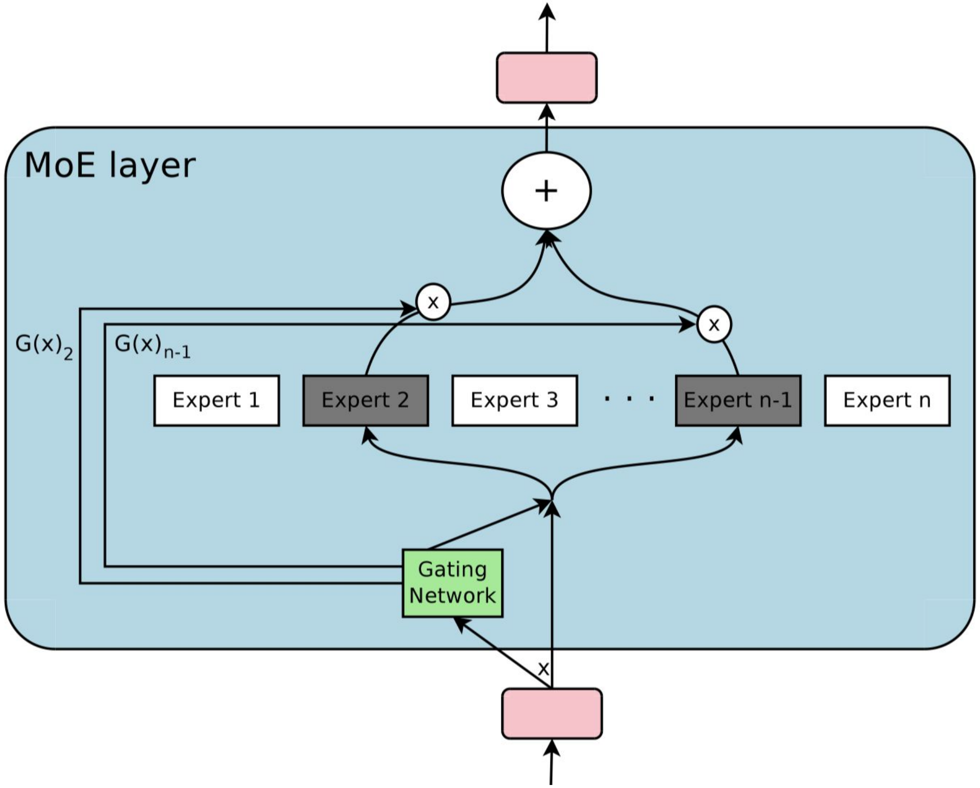


Mixture-of-Experts Blocks

240 experts when training on 8 problems jointly

60 experts when training on each problem separately

4 experts out of the whole expert pool



Mixture-of-Experts Blocks

$$G(x) = Softmax(KeepTopK(H(x), k))$$

$$H(x)_i = (x \cdot W_g)_i + StandardNormal() \cdot Softplus((x \cdot W_{noise})_i)$$

$$KeepTopK(v, k)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ elements of } v. \\ -\infty & \text{otherwise.} \end{cases}$$

Mixture-of-Experts Blocks

$$P(x,i) = Pr\Big((x \cdot W_g)_i + StandardNormal() \cdot Softplus((x \cdot W_{noise})_i)$$
$$> kth\_excluding(H(x), k, i)\Big)$$

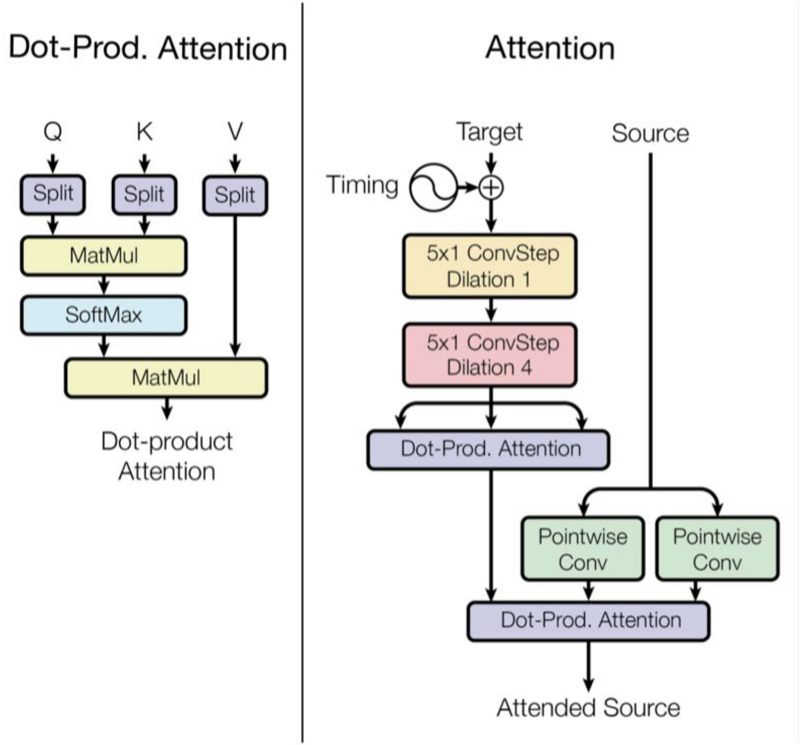
$$Load(X)_i = \sum_{x \in X} P(x,i)$$

$$L_{load}(X) = w_{load} \cdot CV(Load(X))^2$$

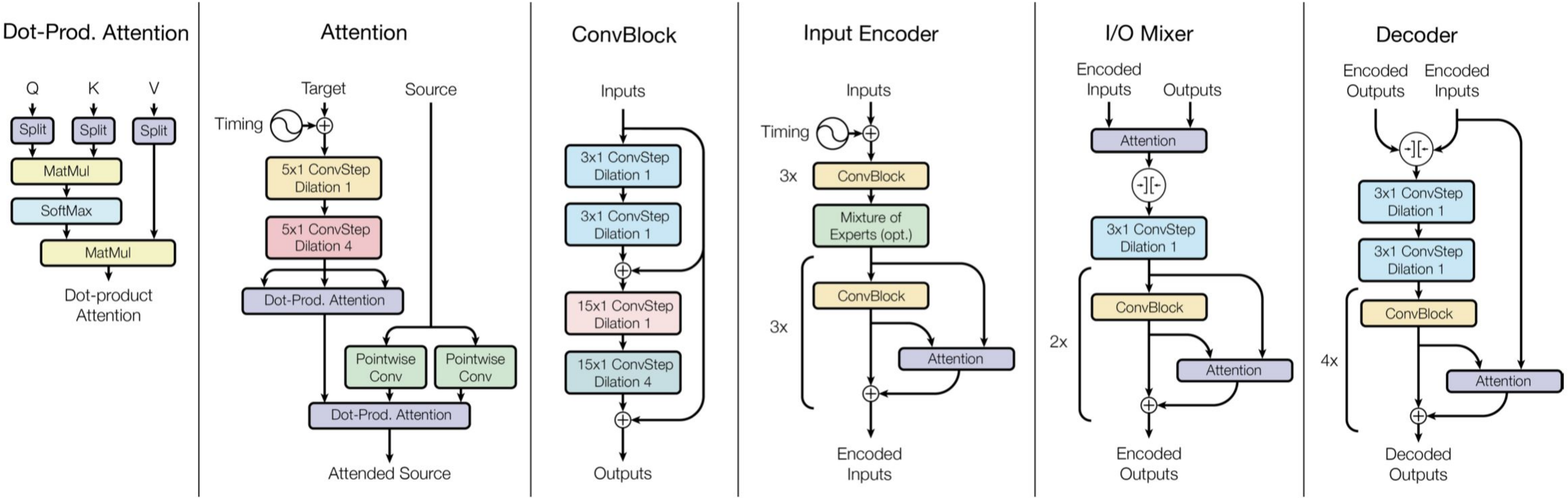
Attention blocks

$$\Delta(2d) = 1e4^{-\frac{2d}{depth}}$$

$$timing(t, [2d, 2d+1]) = [\sin(t\Delta(2d)) \,\|_2\, \cos(t\Delta(2d))]$$



Architecture of the MultiModel



Language modality nets

$$LanguageModality_{\text{in}}(x, W_E) = W_E \cdot x$$
$$LanguageModality_{\text{out}}(x, W_S) = Softmax(W_S \cdot x)$$

Categorical modality nets

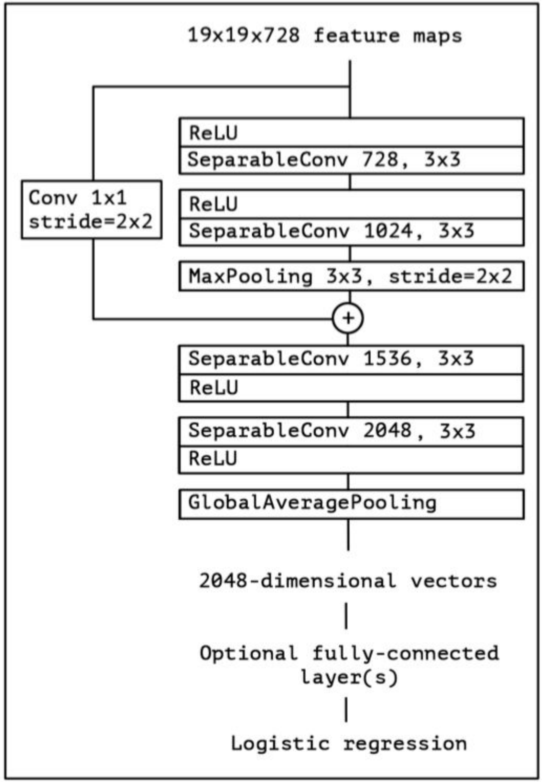
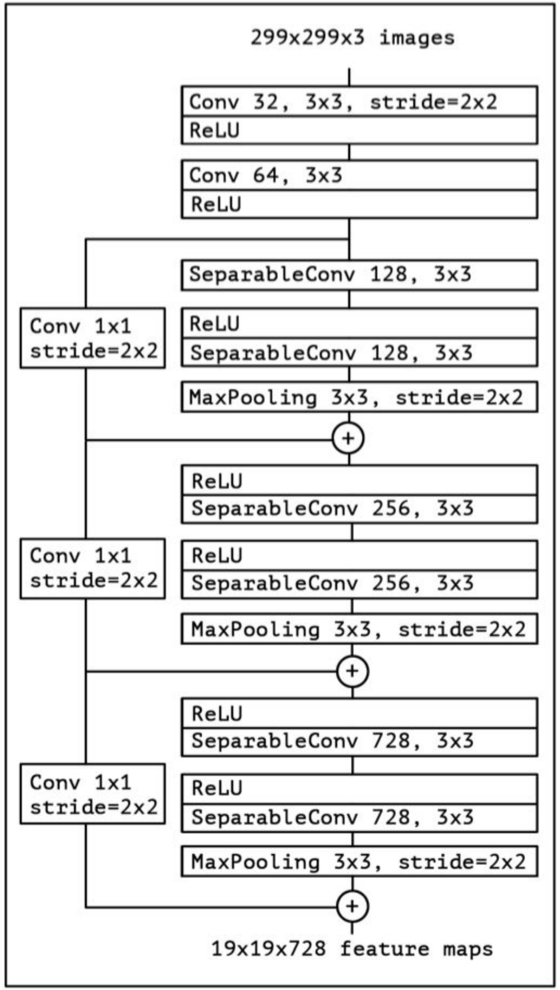
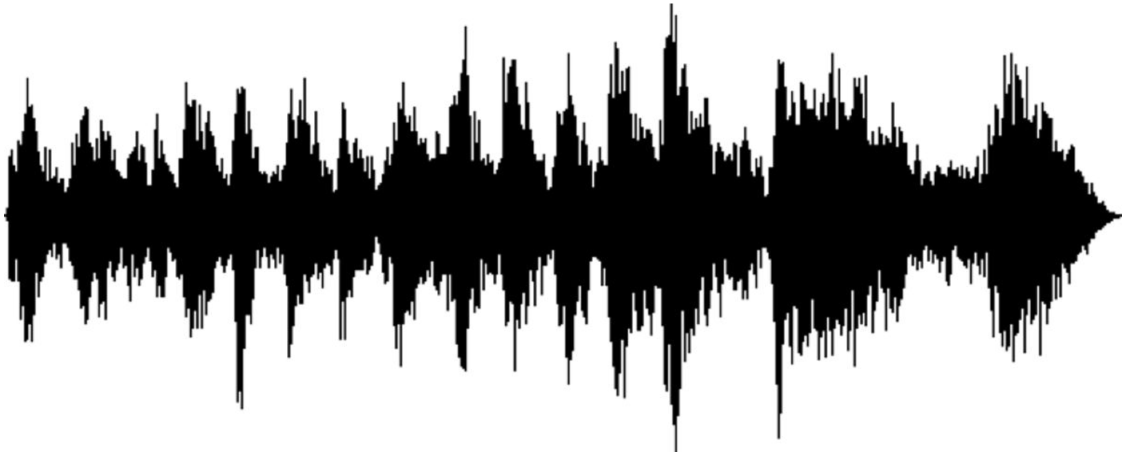


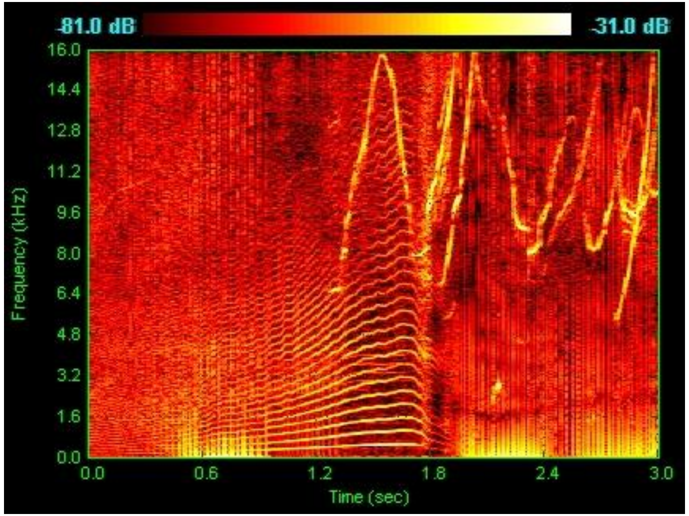
Image modality net



Audio modality net

We accept audio input in the form of a 1-dimensional waveform over time or as a 2-dimensional spectrogram. Both the waveform and spectral input modalities use a stack of 8 $ConvRes$ blocks from the $ImageInputModality$ (Section 2.5.2). The i^{th} block has the form: $l_i = ConvRes(l_{i-1}, 2^i)$. The spectral modality does not perform any striding along the frequency bin dimension, preserving full resolution in the spectral domain.





Experiments

Problem	MultiModel (joint 8-problem)	State of the art
ImageNet (top-5 accuracy)	86%	95%
WMT EN $\rightarrow$ DE (BLEU)	21.2	26.0
WMT EN $\rightarrow$ FR (BLEU)	30.5	40.5

Table 1: Comparing MultiModel to state-of-the-art from [28] and [21].

Problem	Joint 8-problem		Single problem	
	log(perpexity)	accuracy	log(perplexity)	accuracy
ImageNet	1.7	66%	1.6	67%
WMT EN$\rightarrow$DE	1.4	72%	1.4	71%
WSJ speech	4.4	41%	5.7	23%
Parsing	0.15	98%	0.2	97%

Table 2: Comparison of the MultiModel trained jointly on 8 tasks and separately on each task.

Experiments

Problem	Alone			W/ ImageNet			W/ 8 Problems		
	log(ppl)	acc.	full	log(ppl)	acc.	full	log(ppl)	acc.	full
Parsing	0.20	97.1%	11.7%	0.16	97.5%	12.7%	0.15	97.9%	14.5%

Table 3: Results on training parsing alone, with ImageNet, and with 8 other tasks. We report log-perplexity, per-token accuracy, and the percentage of fully correct parse trees.

Problem	All Blocks		Without MoE		Without Attention	
	log(perpexity)	accuracy	log(perplexity)	accuracy	log(perplexity)	accuracy
ImageNet	1.6	67%	1.6	66%	1.6	67%
WMT EN$\rightarrow$FR	1.2	76%	1.3	74%	1.4	72%

Table 4: Ablating mixture-of-experts and attention from MultiModel training.

https://arxiv.org/abs/1706.05137