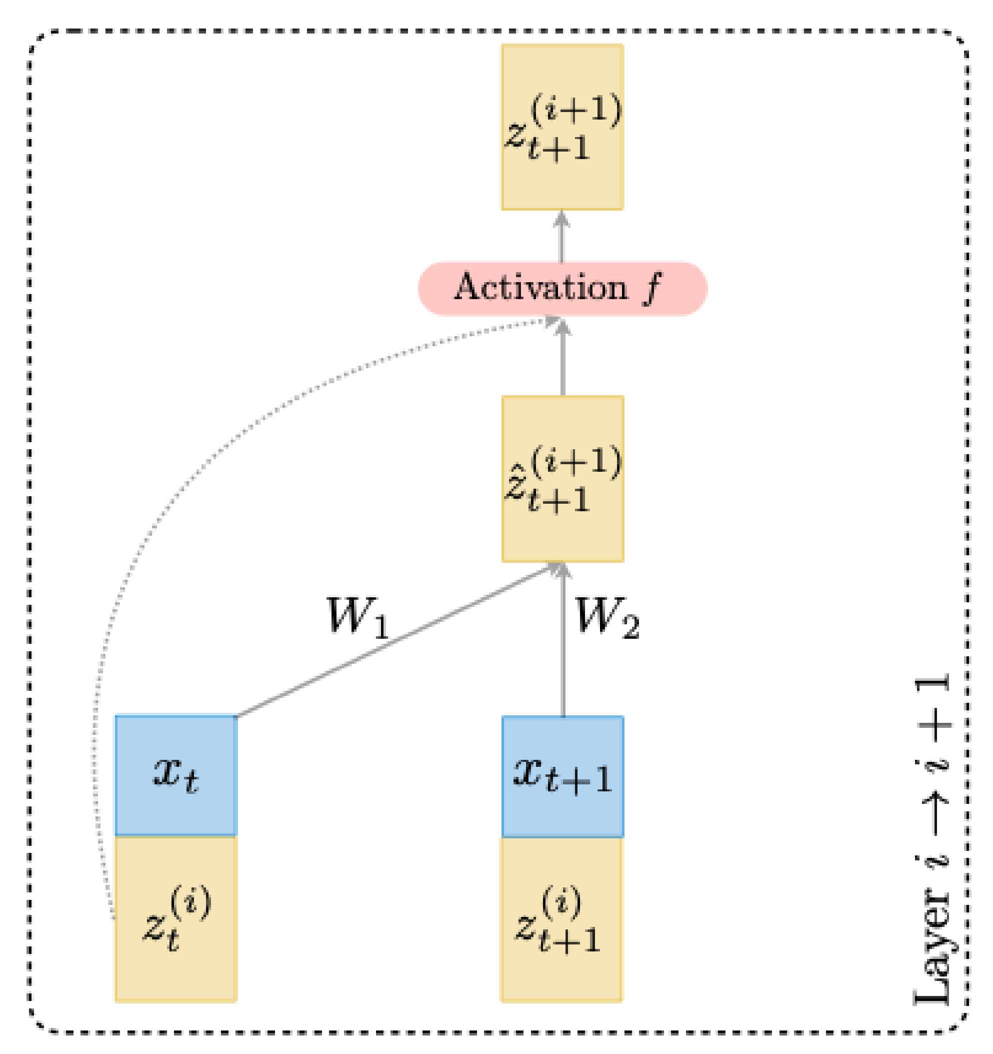
DEEP EQUILIBRIUM MODELS (DEQ)

МАРЬИН НИКИТА

171 ГРУППА

MOTIVATION. TRELLIS NETWORKS

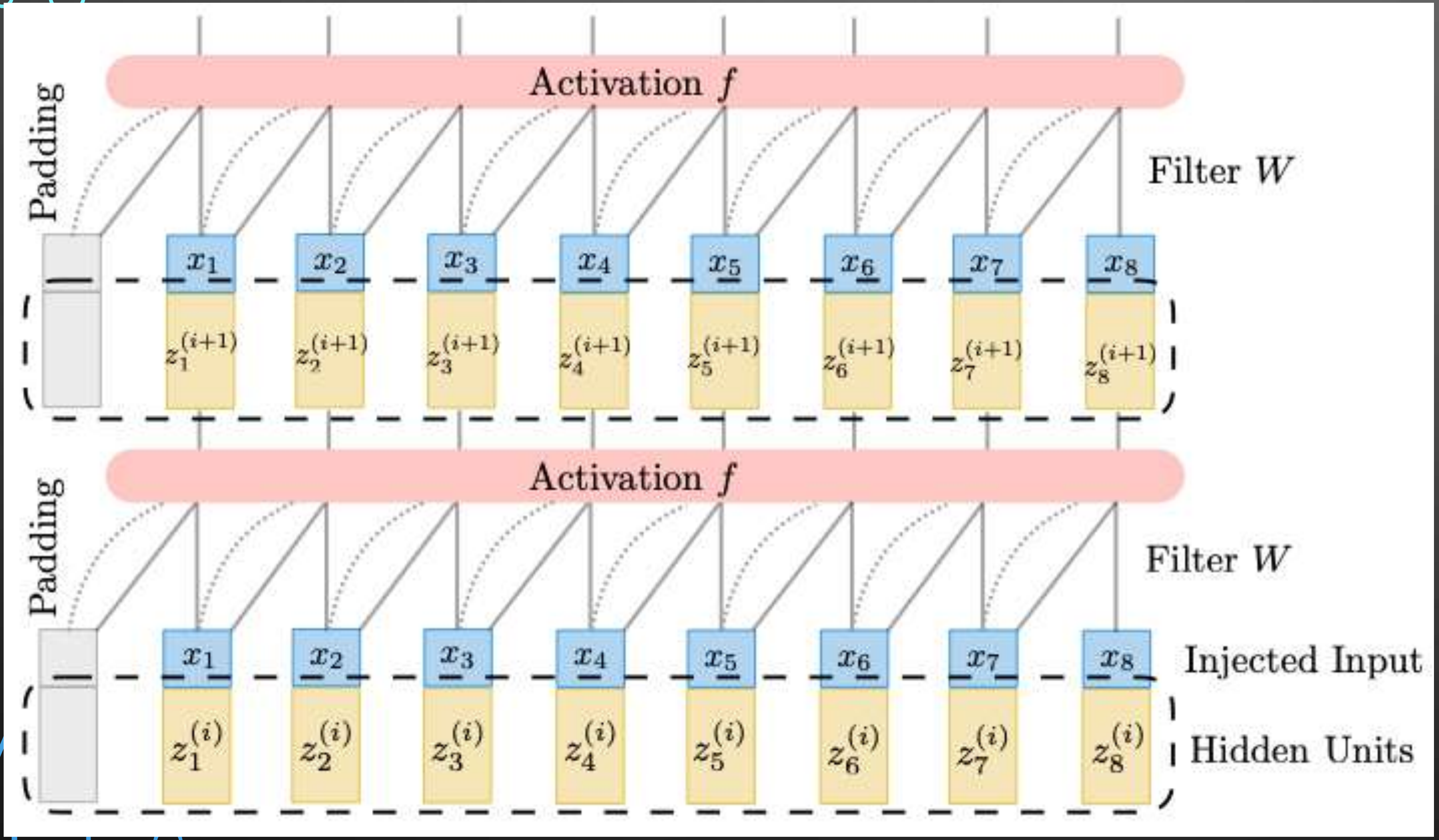


$$\hat{z}_{t+1}^{(i+1)} = W_1 \begin{bmatrix} x_t \\ z_t^{(i)} \end{bmatrix} + W_2 \begin{bmatrix} x_{t+1} \\ z_{t+1}^{(i)} \end{bmatrix}$$

$$W_1, W_2 \in \mathbb{R}^{r \times (p+q)}$$

$$z_{t+1}^{(i+1)} = f\left(\hat{z}_{t+1}^{(i+1)}, z_t^{(i)}\right)$$

MOTIVATION. TRELLIS NETWORKS

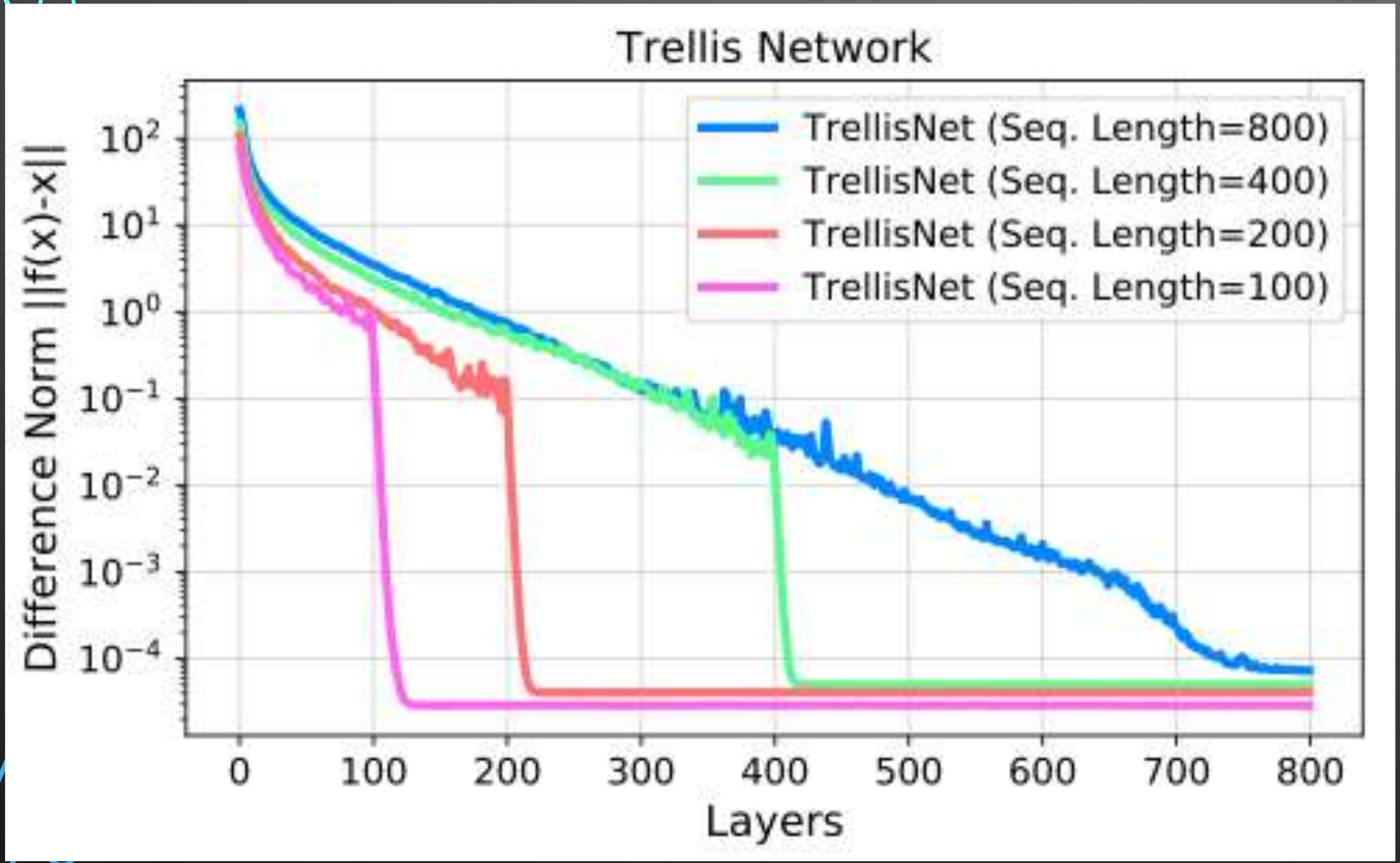


$$\tilde{x}_{t+1} = W_1^x x_t + W_2^x x_{t+1}$$

$$\hat{z}_{1:T}^{(i+1)} = \mathrm{Conv1D}\left(z_{1:T}^{(i)}; W\right) + \tilde{x}_{1:T}$$

$$z_{1:T}^{(i+1)} = f\left(\hat{z}_{1:T}^{(i+1)}, z_{1:T-1}^{(i)}\right)$$

MOTIVATION. TRELLIS NETWORKS



Видно, что сеть сходится к какой – то точке равновесия. Вопрос : можно ли явно найти эту точку?

WEIGHT-TIED DEEP SEQUENCE MODELS

$$\mathbf{z}_{1:T}^{[i+1]} = f_\theta(\mathbf{z}_{1:T}^{[i]}; \mathbf{x}_{1:T}), \quad i = 0, \dots, L-1, \quad \mathbf{z}_{1:T}^{[0]} = \mathbf{0}, \quad G(\mathbf{x}_{1:T}) \equiv \mathbf{z}_{1:T}^{[L]}$$

Свойства weight-tied:

1) Такая модель уменьшает риск переобучиться.
2) Значительно уменьшает размер модели.
3) Можно показать, что любая сеть может быть представлена как weight-tied такой же глубины, но с увеличением ширины.
4) Сеть может быть развернута на любую глубину.

$$\lim_{i\to\infty} \mathbf{z}_{1:T}^{[i]} = \lim_{i\to\infty} f_\theta\left(\mathbf{z}_{1:T}^{[i]}; \mathbf{x}_{1:T}\right) \equiv f_\theta\left(\mathbf{z}_{1:T}^{\star}; \mathbf{x}_{1:T}\right) = \mathbf{z}_{1:T}^{\star}$$

DEQ APPROACH. FORWARD PASS

$$\mathbf{z}_{1:T}^{[i+1]} = f_\theta\left(\mathbf{z}_{1:T}^{[i]}; \mathbf{x}_{1:T}\right) \quad \text{for } i = 0, 1, 2, \ldots$$

По сути это можно рассматривать как уравнение. Тогда корнем этого уравнения будет точка эквилибриума.

$$g_\theta(\mathbf{z}_{1:T}^\star; \mathbf{x}_{1:T}) = f_\theta\left(\mathbf{z}_{1:T}^\star; \mathbf{x}_{1:T}\right) - \mathbf{z}_{1:T}^\star \to 0$$

Перепишем уравнение по другому и будем оптимизировать до определенной точности.

$$\mathbf{z}_{1:T}^{[i+1]} = \mathbf{z}_{1:T}^{[i]} - \alpha B g_\theta(\mathbf{z}_{1:T}^{[i]}; \mathbf{x}_{1:T}) \quad \text{for } i = 0, 1, 2, \ldots$$

Метод оптимизации Бройдена.

$$\mathbf{z}_{1:T}^\star = \mathrm{RootFind}(g_\theta; \mathbf{x}_{1:T})$$

По сути, forward pass – это алгоритм оптимизации.

DEQ APPROACH. BACKWARD PASS

Theorem 1. ***(Gradient of the Equilibrium Model)*** *Let $\mathbf{z}^\star_{1:T} \in \mathbb{R}^{T\times d}$ be an equilibrium hidden sequence with length T and dimensionality d, and $\mathbf{y}_{1:T} \in \mathbb{R}^{T\times q}$ the ground-truth (target) sequence. Let $h : \mathbb{R}^d \to \mathbb{R}^q$ be any differentiable function and let $\mathcal{L} : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ be a loss function (where $h, \mathcal{L}$ are applied in a vectorized manner) that computes*

$$\ell = \mathcal{L}(h(\mathbf{z}^\star_{1:T}), \mathbf{y}_{1:T}) = \mathcal{L}(h(\mathsf{RootFind}(g_\theta; \mathbf{x}_{1:T})), \mathbf{y}_{1:T}). \tag{7}$$

Then the loss gradient w.r.t. $(\cdot)$ (for instance, θ or $\mathbf{x}_{1:T}$) is

$$\frac{\partial \ell}{\partial(\cdot)} = -\frac{\partial \ell}{\partial \mathbf{z}^\star_{1:T}} \left(J^{-1}_{g_\theta}\big|_{\mathbf{z}^\star_{1:T}}\right) \frac{\partial f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\partial(\cdot)} = -\frac{\partial \ell}{\partial h} \frac{\partial h}{\partial \mathbf{z}^\star_{1:T}} \left(J^{-1}_{g_\theta}\big|_{\mathbf{z}^\star_{1:T}}\right) \frac{\partial f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\partial(\cdot)}, \tag{8}$$

where $J^{-1}_{g_\theta}\big|_{\mathbf{x}}$ is the inverse Jacobian of g_θ evaluated at $\mathbf{x}$.

DEQ APPROACH. BACKWARD PASS

Proof of Theorem 1. We first write out the equilibrium sequence condition: $f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T}) = \mathbf{z}^\star_{1:T}$. By implicitly differentiating two sides of this condition with respect to $(\cdot)$:

$$\frac{\mathrm{d}\mathbf{z}^\star_{1:T}}{\mathrm{d}(\cdot)} = \frac{\mathrm{d}f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\mathrm{d}(\cdot)} = \frac{\partial f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\partial(\cdot)} + \frac{\partial f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^\star_{1:T}} \frac{\mathrm{d}\mathbf{z}^\star_{1:T}}{\mathrm{d}(\cdot)}$$

$$\Longrightarrow \left(I - \frac{\partial f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^\star_{1:T}}\right) \frac{\mathrm{d}\mathbf{z}^\star_{1:T}}{\mathrm{d}(\cdot)} = \frac{\partial f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\partial(\cdot)}$$

Since $g_\theta(\mathbf{z}^\star_{1:T}) = f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T}) - \mathbf{z}^\star_{1:T}$, we have

$$J_{g_\theta}\big|_{\mathbf{z}^\star_{1:T}} = -\left(I - \frac{\partial f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\partial \mathbf{z}^\star_{1:T}}\right),$$

which implies

$$\frac{\partial \ell}{\partial(\cdot)} = \frac{\partial \ell}{\partial \mathbf{z}^\star_{1:T}} \frac{\mathrm{d}\mathbf{z}^\star_{1:T}}{\mathrm{d}(\cdot)} = -\frac{\partial \ell}{\partial \mathbf{z}^\star_{1:T}} \left(J^{-1}_{g_\theta}\big|_{\mathbf{z}^\star_{1:T}}\right) \frac{\partial f_\theta(\mathbf{z}^\star_{1:T}; \mathbf{x}_{1:T})}{\partial(\cdot)}.$$

$\square$

OPTIMIZATION

Аппроксимация якобиана через формулу Шермана – Моррисона:

$$J_{g_\theta}^{-1}\Big|_{\mathbf{z}_{1:T}^{[i+1]}} \approx B_{g_\theta}^{[i+1]} = B_{g_\theta}^{[i]} + \frac{\Delta\mathbf{z}^{[i+1]} - B_{g_\theta}^{[i]}\Delta g_\theta^{[i+1]}}{{\Delta\mathbf{z}^{[i+1]}}^\top B_{g_\theta}^{[i]}\Delta g_\theta^{[i+1]}}{\Delta\mathbf{z}^{[i+1]}}^\top B_{g_\theta}^{[i]}$$

$$B_{g_\theta}^{[0]} = -I$$

Нахождение выражения через решение системы линейных уравнений:

$$-\frac{\partial \ell}{\partial \mathbf{z}_{1:T}^\star}\left(J_{g_\theta}^{-1}\Big|_{\mathbf{z}_{1:T}^\star}\right) \longrightarrow \left(J_{g_\theta}^\top\big|_{\mathbf{z}_{1:T}^\star}\right)\mathbf{x}^\top + \left(\frac{\partial \ell}{\partial \mathbf{z}_{1:T}^\star}\right)^\top = \mathbf{0}$$

PROPERTIES OF DEEP EQUILIBRIUM MODELS

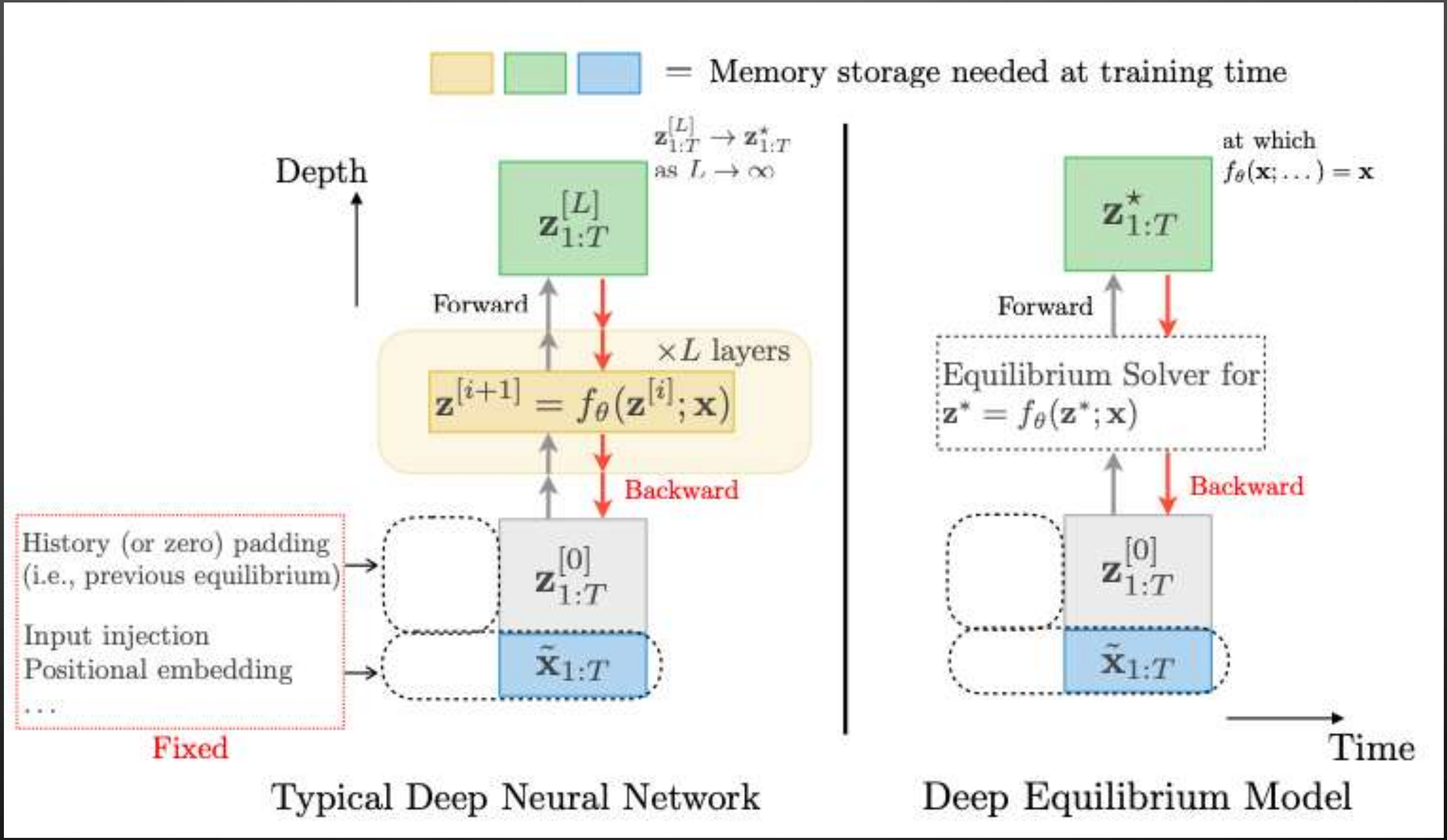
1) Стоимость памяти : храним x, точку эквилибриума z, функцию f. Так как сам якобиан нам не нужен, а лишь умножение его на вектор, то явно мы его не храним.

2) Шаги не зависят от выбора f, однако для уверенности в сходимости, f должна быть устойчивой и ограниченной.

Theorem 2. ***(Universality of "single-layer" DEQs.)*** *Let* $\mathbf{x}_{1:T} \in \mathbb{R}^{T\times p}$ *be the input sequence, and* $\theta^{[1]}, \theta^{[2]}$ *the sets of parameters for stable transformations* $f_{\theta^{[1]}} : \mathbb{R}^r \times \mathbb{R}^p \to \mathbb{R}^r$ *and* $v_{\theta^{[2]}} : \mathbb{R}^d \times \mathbb{R}^r \to \mathbb{R}^d$*, respectively. Then there exists* $\Gamma_\Theta : \mathbb{R}^{d+r} \times \mathbb{R}^p \to \mathbb{R}^{d+r}$*, where* $\Theta = \theta^{[1]} \cup \theta^{[2]}$*, s.t.*

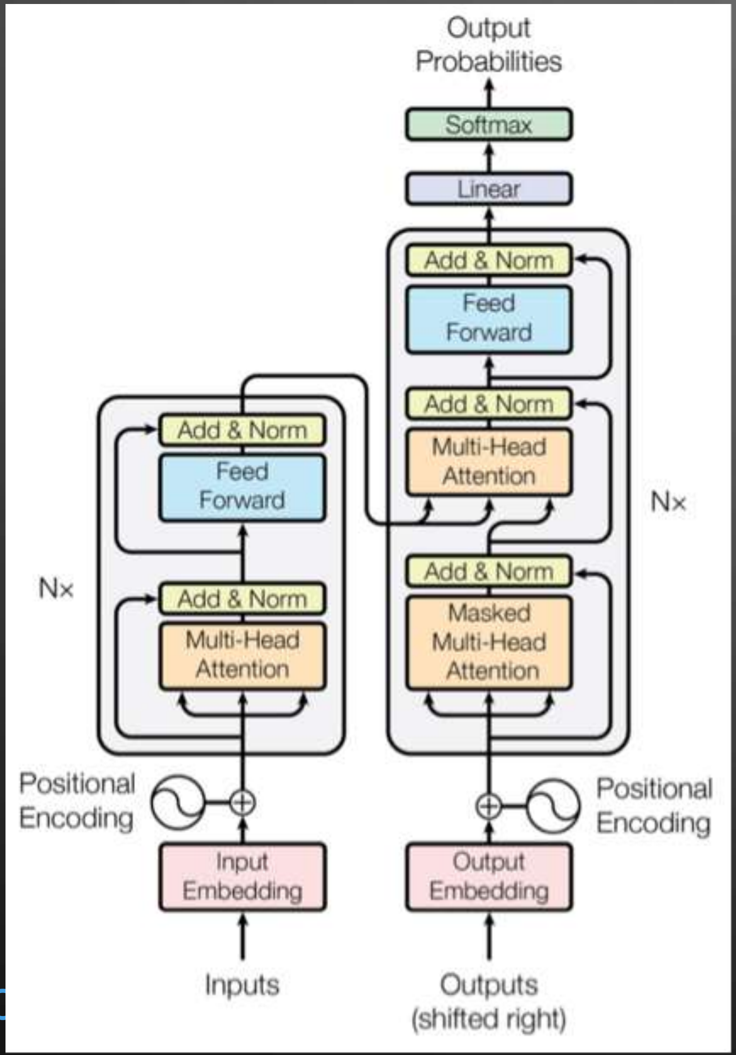
$$\mathbf{z}^\star_{1:T} = \mathsf{RootFind}\big(g^f_{\theta^{[2]}}; \mathsf{RootFind}\big(g^v_{\theta^{[1]}}; \mathbf{x}_{1:T}\big)\big) = \mathsf{RootFind}\big(g^\Gamma_\Theta; \mathbf{x}_{1:T}\big)_{[:,-d:]}, \qquad (12)$$

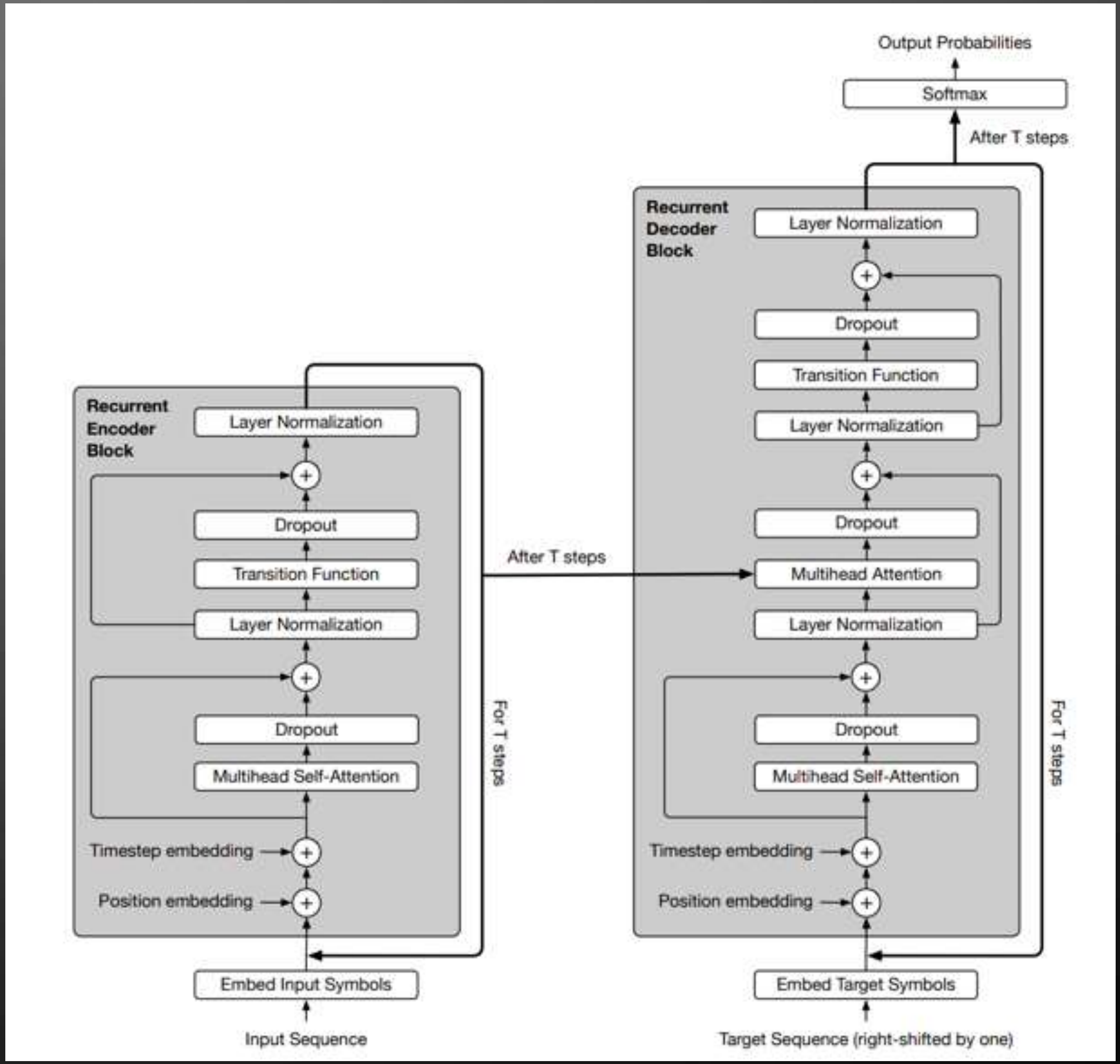
where $[\cdot]_{[:,-d:]}$ *denotes the last d feature dimensions of* $[\cdot]$.



= Memory storage needed at training time
Depth
$\mathbf{z}_{1:T}^{[L]} \to \mathbf{z}_{1:T}^{\star}$ as $L \to \infty$
$\mathbf{z}_{1:T}^{[L]}$
Forward
$\times L$ layers
$\mathbf{z}^{[i+1]} = f_\theta(\mathbf{z}^{[i]}; \mathbf{x})$
Backward
History (or zero) padding
(i.e., previous equilibrium)
Input injection
Positional embedding
...
Fixed
$\mathbf{z}_{1:T}^{[0]}$
$\tilde{\mathbf{x}}_{1:T}$
Typical Deep Neural Network
at which $f_\theta(\mathbf{x}; \dots) = \mathbf{x}$
$\mathbf{z}_{1:T}^{\star}$
Forward
Equilibrium Solver for
$\mathbf{z}^* = f_\theta(\mathbf{z}^*; \mathbf{x})$
Backward
$\mathbf{z}_{1:T}^{[0]}$
$\tilde{\mathbf{x}}_{1:T}$
Time
Deep Equilibrium Model

UNIVERSAL TRANSFORMER





TRELLISNET AND WEIGHT-TIED TRANSFORMERS AS DEQ

TrellisNet:

$$\begin{aligned} \tilde{\mathbf{x}}_{1:T} &= \text{Input injection (i.e., linearly transformed inputs by } \mathsf{Conv1D}(\mathbf{x}_{1:T}; W_x)) \\ f_\theta(\mathbf{z}_{1:T}; \mathbf{x}_{1:T}) &= \psi(\mathsf{Conv1D}([\mathbf{u}_{-(k-1)s:}, \mathbf{z}_{1:T}]; W_z) + \tilde{\mathbf{x}}_{1:T}) \end{aligned}$$

Universal Transformer:

$$\begin{aligned} \tilde{\mathbf{x}}_{1:T} &= \text{Input injection (i.e., linearly transformed inputs by } \mathbf{x}_{1:T} W_x) \\ f_\theta(\mathbf{z}_{1:T}; \mathbf{x}_{1:T}) &= \mathsf{LN}(\phi(\mathsf{LN}(\mathsf{SelfAttention}(\mathbf{z}_{1:T} W_{QKV} + \tilde{\mathbf{x}}_{1:T}; \mathsf{PE}_{1:T})))) \end{aligned}$$

ЭКСПЕРИМЕНТЫ

COPY MEMORY TASK

Задача проверить способность модели долгое время точно запоминать последовательность.

Table 1: DEQ achieves strong performance on the long-range copy-memory task.

	Models (Size)			
	DEQ-Transformer (ours) (14K)	TCN [7] (16K)	LSTM [26] (14K)	GRU [14] (14K)
Copy Memory T=400 Loss	**3.5e-6**	**2.7e-5**	0.0501	0.0491

LANGUAGE MODELING. PERFORMANCE ON PENN TREEBANK

Table 2: DEQ achieves competitive performance on word-level Penn Treebank language modeling (on par with SOTA results, without fine-tuning steps [34]). †The memory footprints are benchmarked (for fairness) on input sequence length 150 and batch size 15, which does not reflect the actual hyperparameters used; the values also do *not* include the memory for word embeddings.

Word-level Language Modeling w/ Penn Treebank (PTB)				
Model	# Params	Non-embedding model size	Test perplexity	Memory†
Variational LSTM [22]	66M	-	73.4	-
NAS Cell [55]	54M	-	62.4	-
NAS (w/ black-box hyperparameter tuner) [32]	24M	20M	59.7	-
AWD-LSTM [34]	24M	20M	58.8	-
DARTS architecture search (second order) [29]	23M	20M	**55.7**	-
60-layer TrellisNet (w/ auxiliary loss, w/o MoS) [8]	24M	20M	57.0	8.5GB
DEQ-TrellisNet (ours)	24M	20M	57.1	**1.2GB**

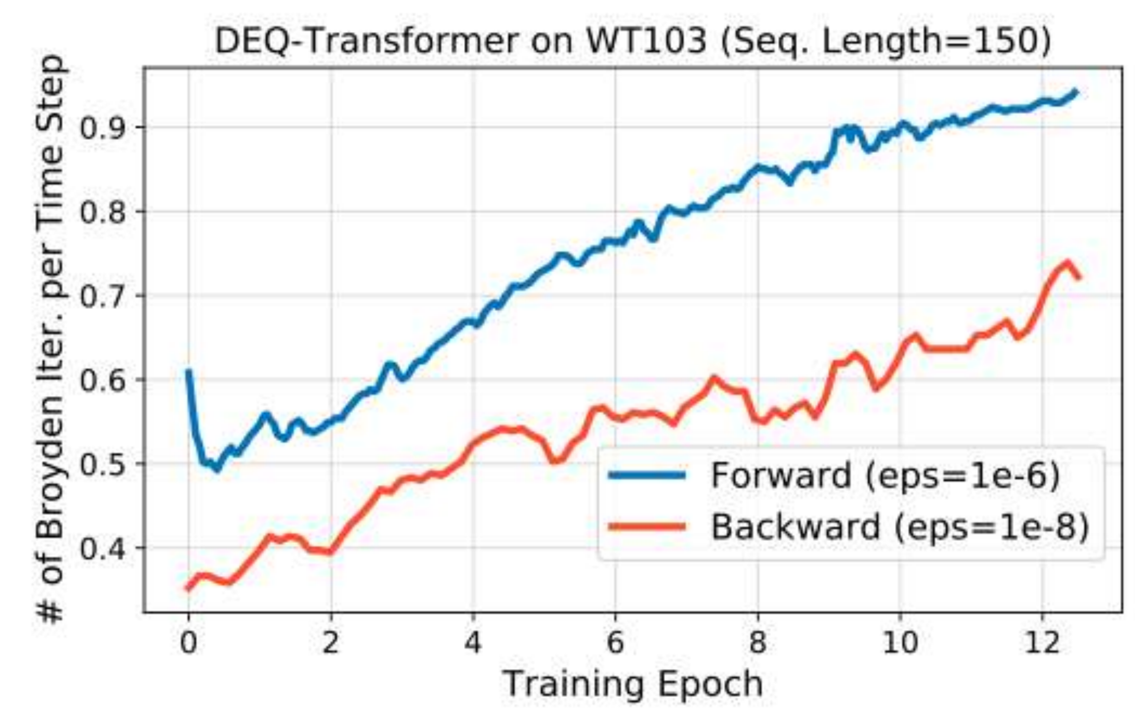
LANGUAGE MODELING. PERFORMANCE ON WIKITEXT-103

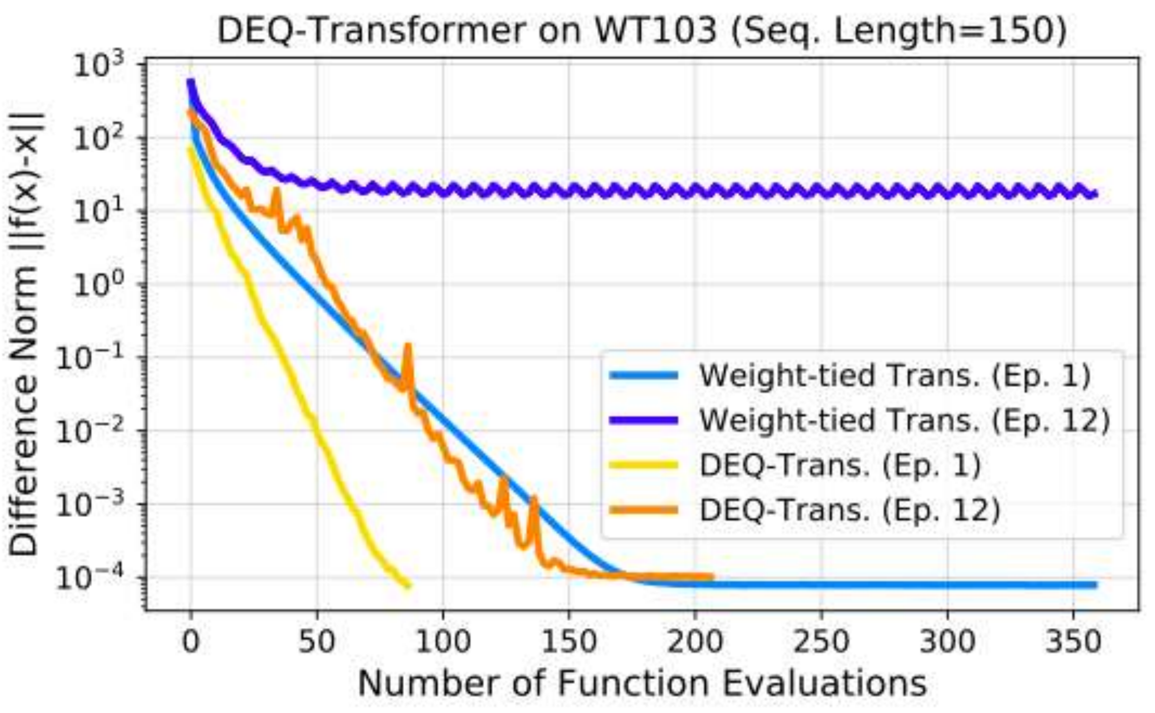
Table 3: DEQ-based models are competitive with SOTA deep networks of the same model size on the WikiText-103 corpus, with significantly less memory. †See Table 2 for more details on the memory benchmarking. Transformer-XL models are not weight-tied, unless specified otherwise.

Word-level Language Modeling w/ WikiText-103 (WT103)				
Model	# Params	Non-Embedding Model Size	Test perplexity	Memory†
Generic TCN [7]	150M	34M	45.2	-
Gated Linear ConvNet [17]	230M	-	37.2	-
AWD-QRNN [33]	159M	51M	33.0	7.1GB
Relational Memory Core [40]	195M	60M	31.6	-
Transformer-XL (X-large, adaptive embed., on TPU) [16]	257M	224M	**18.7**	12.0GB
70-layer TrellisNet (+ auxiliary loss, etc.) [8]	180M	45M	29.2	24.7GB
70-layer TrellisNet with *gradient checkpointing*	180M	45M	29.2	5.2GB
DEQ-TrellisNet (ours)	180M	45M	**29.0**	**3.3GB**
Transformer-XL (medium, 16 layers)	165M	44M	24.3	8.5GB
DEQ-Transformer (medium, ours).	172M	43M	24.2	**2.7GB**
Transformer-XL (medium, 18 layers, adaptive embed.)	110M	72M	23.6	9.0GB
DEQ-Transformer (medium, adaptive embed., ours)	110M	70M	**23.2**	3.7GB
Transformer-XL (small, 4 layers)	139M	4.9M	35.8	4.8GB
Transformer-XL (small, weight-tied 16 layers)	138M	4.5M	34.9	6.8GB
DEQ-Transformer (small, ours).	138M	4.5M	**32.4**	**1.1GB**

CONVERGENCE TO EQUILIBRIUM

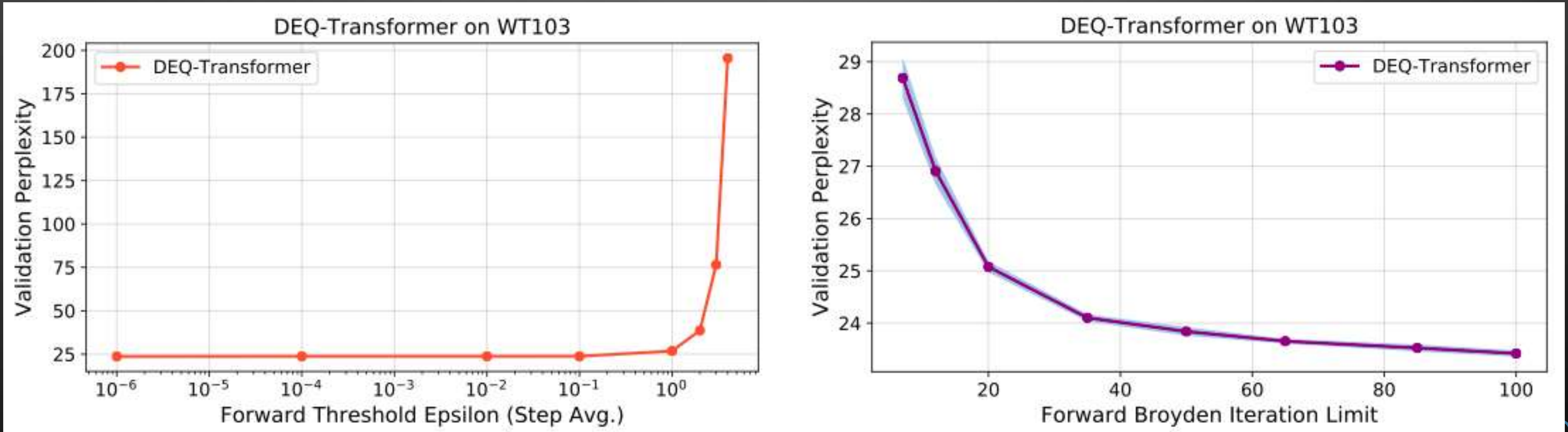
$$\frac{\text{Total Broyden Iterations}}{\text{Sequence Length}}$$





BROYDEN ITERATIONS AND THE RUNTIME OF DEQ

Время работы преимущественно зависит от количества итераций алгоритма Бройдена.



CONCLUSION

- Экономия памяти.
- Конкурентоспособные результаты на реальных задачах.
- Новый подход к обучению через неявные слои.