CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

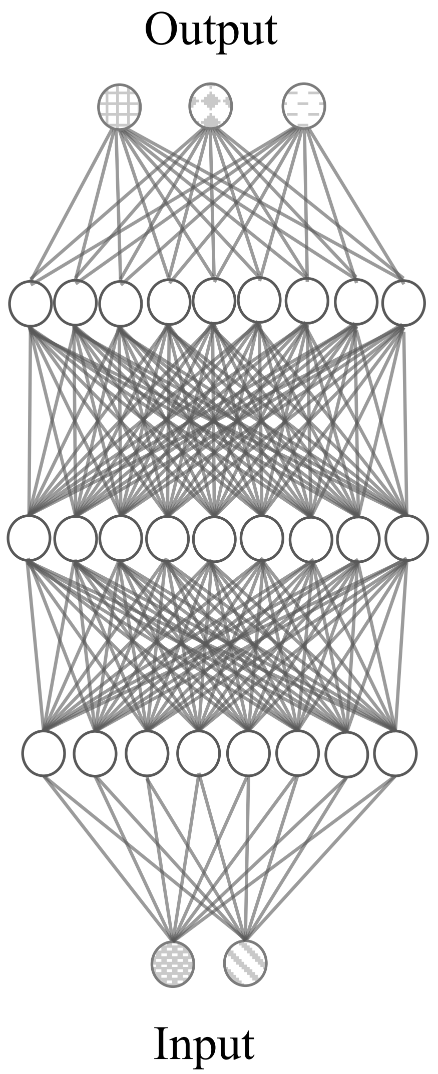
Convolutional and Residual Networks Provably Contain Lottery Tickets

Rebekka Burkholz

The quest for lottery tickets

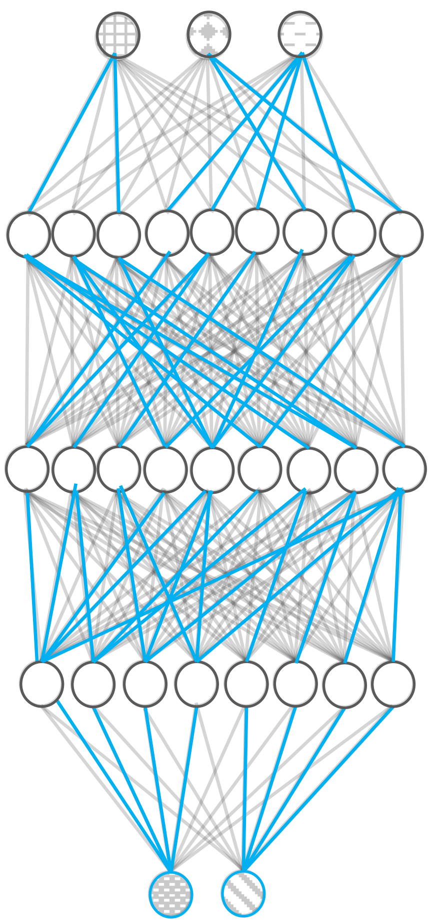
How to train sparse irregular neural nets?

The quest for lottery tickets



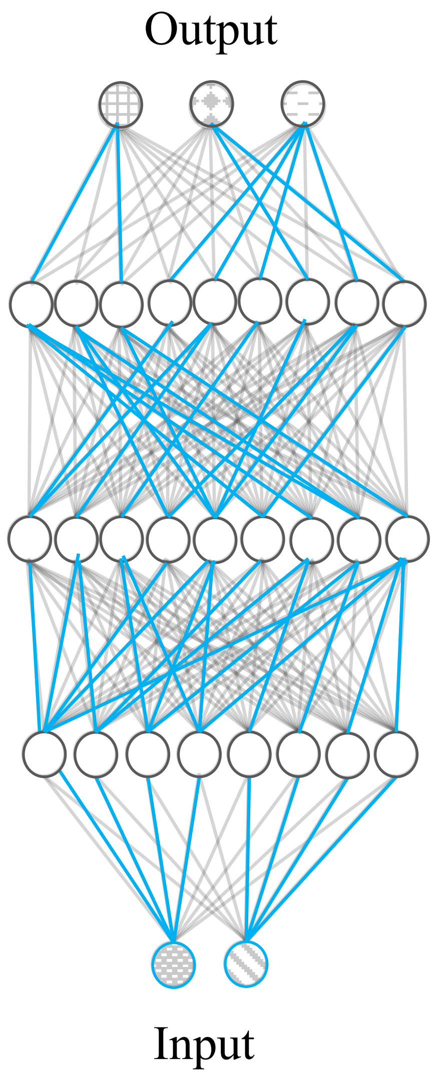
Lottery ticket hypothesis (LTH) (Frankle and Carbin, 2019)

The quest for lottery tickets



Lottery ticket hypothesis (LTH) (Frankle and Carbin, 2019)
There exist small, well *trainable* **sub-networks** of large, **randomly initialized** neural networks.

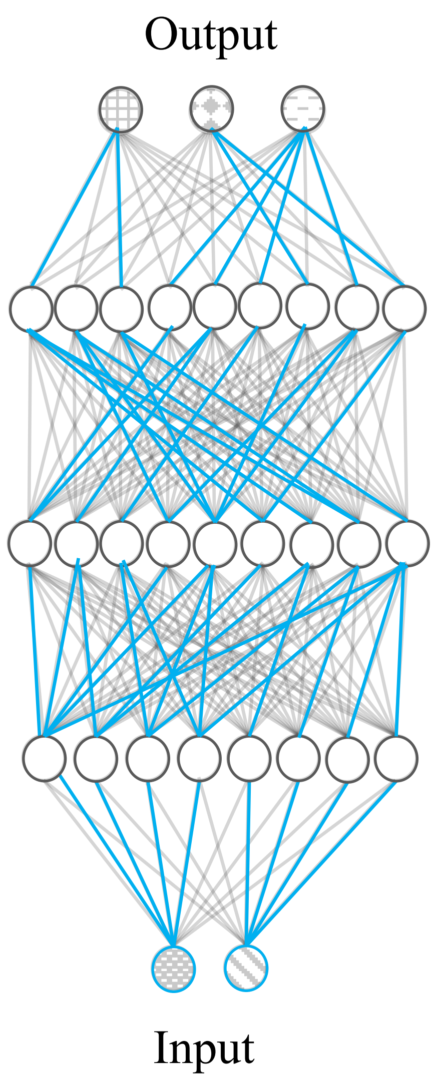
The quest for lottery tickets



Lottery ticket hypothesis (LTH) (Frankle and Carbin, 2019)
There exist small, well *trainable* **sub-networks** of large, **randomly initialized** neural networks.

Strong lottery ticket hypothesis (SLTH) (Ramanujan et al. 2020)
There exist small sub-networks of large, randomly initialized neural networks that perform well *at initialization*.

The quest for lottery tickets



Lottery ticket hypothesis (LTH) (Frankle and Carbin, 2019)
There exist small, well *trainable* **sub-networks** of large, **randomly initialized** neural networks.

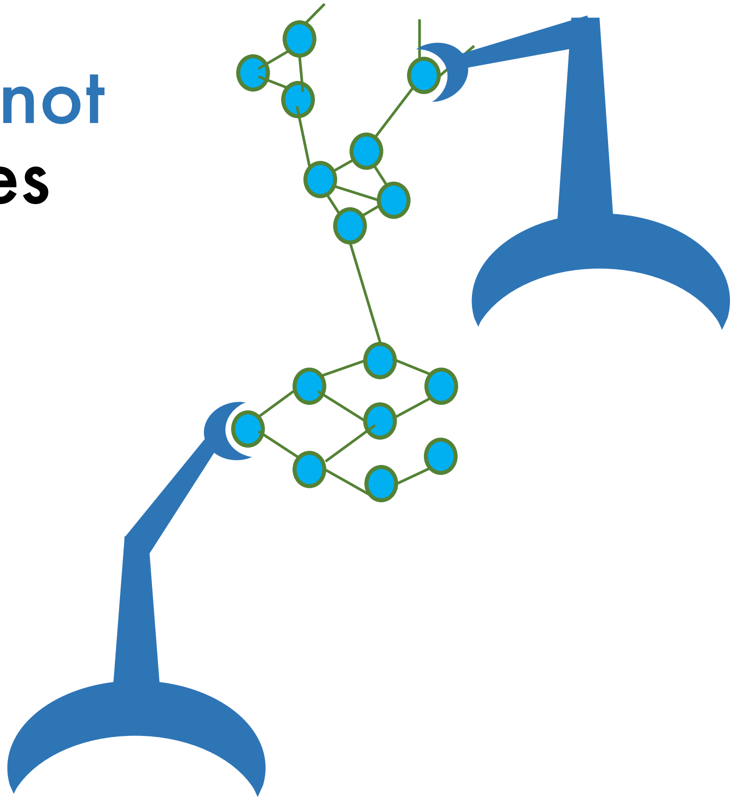
Strong lottery ticket hypothesis (SLTH) (Ramanujan et al. 2020)
There exist small sub-networks of large, randomly initialized neural networks that perform well *at initialization*.

Theory for fully connected networks

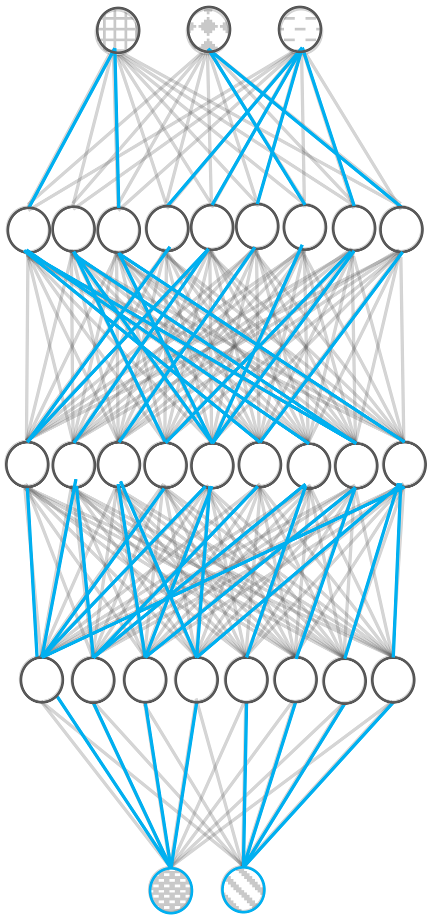
- 2L construction (Malach et al. 2020)
- Subset sum approximation (Pensia et al. 2020)
- L+1 construction, general activation functions (Burkholz 2022)

Why would we need theory?

- Contemporary pruning algorithms cannot find LTs in large, complex NN structures (unless they resort to rewinding, …).
- When do pruning algorithms have a chance to succeed?



The quest for lottery tickets



Lottery ticket hypothesis (LTH) (Frankle and Carbin, 2019)
There exist small, well *trainable* **sub-networks** of large, **randomly initialized** neural networks.

Strong lottery ticket hypothesis (SLTH) (Ramanujan et al. 2020)
There exist small sub-networks of large, randomly initialized neural networks that perform well *at initialization*.

Theory for fully connected networks

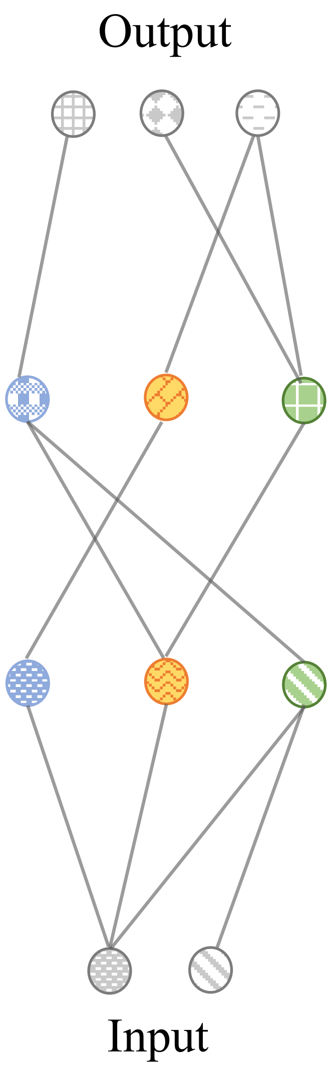
- 2L construction (Malach et al. 2020)
- Subset sum approximation (Pensia et al. 2020)
- L+1 construction, general activation functions (Burkholz 2022)

Theory for common, complex architectures

- ConvNets with **positive inputs** and **ReLUs** (da Cunha et al. 2021)
- **Here**: ConvNets and ResNets for general **inputs and act. functions**

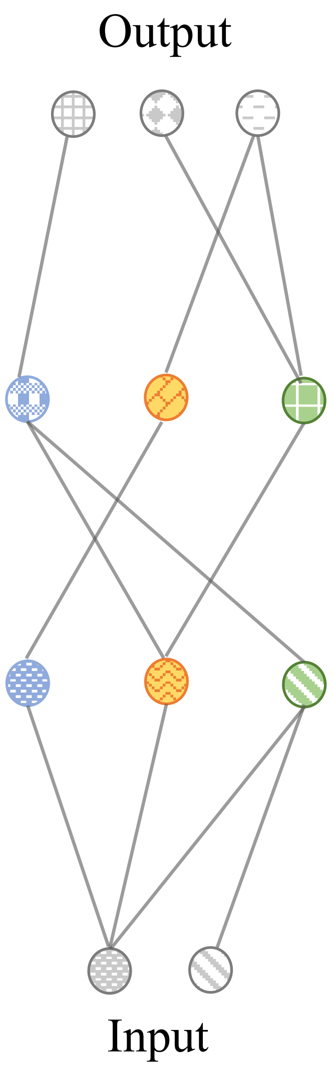
Existence proof set-up

Sparse target network f_t



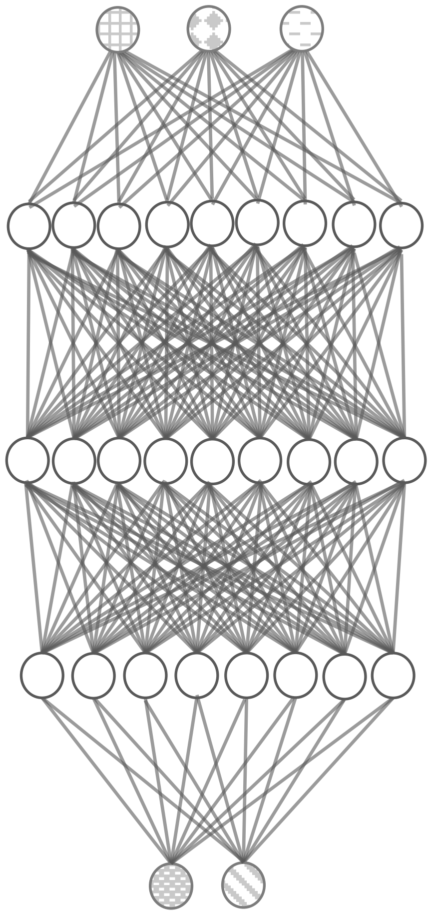
Existence proof set-up

Sparse target **network** f_t



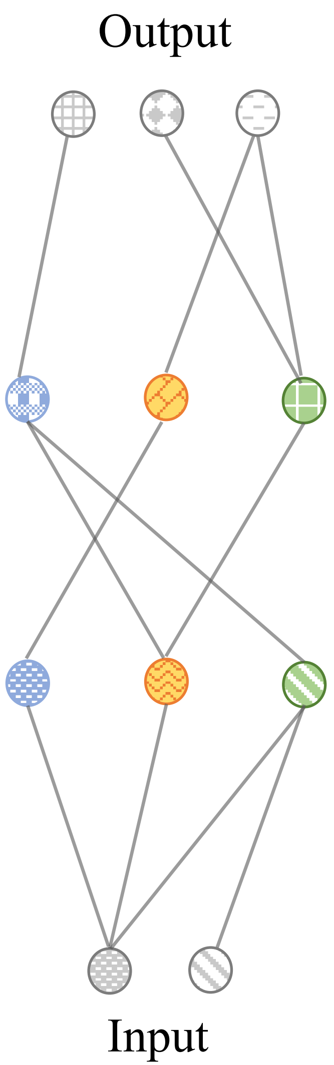
$\cong$

Dense source **network** f_0



Existence proof set-up

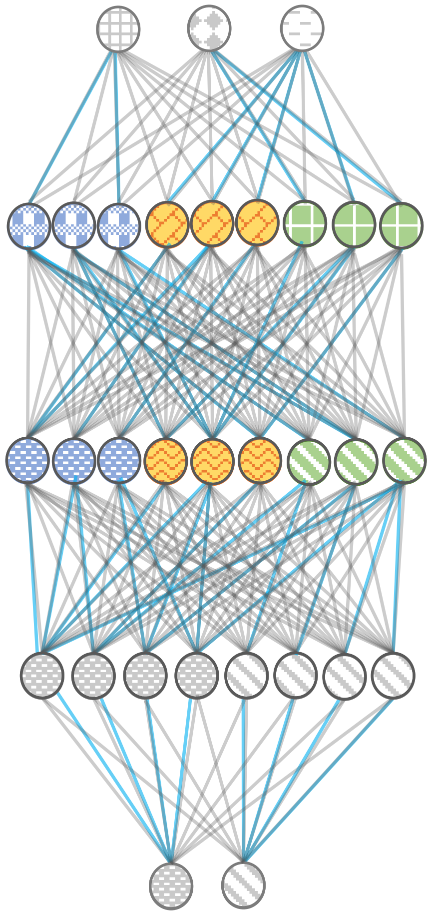
Sparse target network f_t



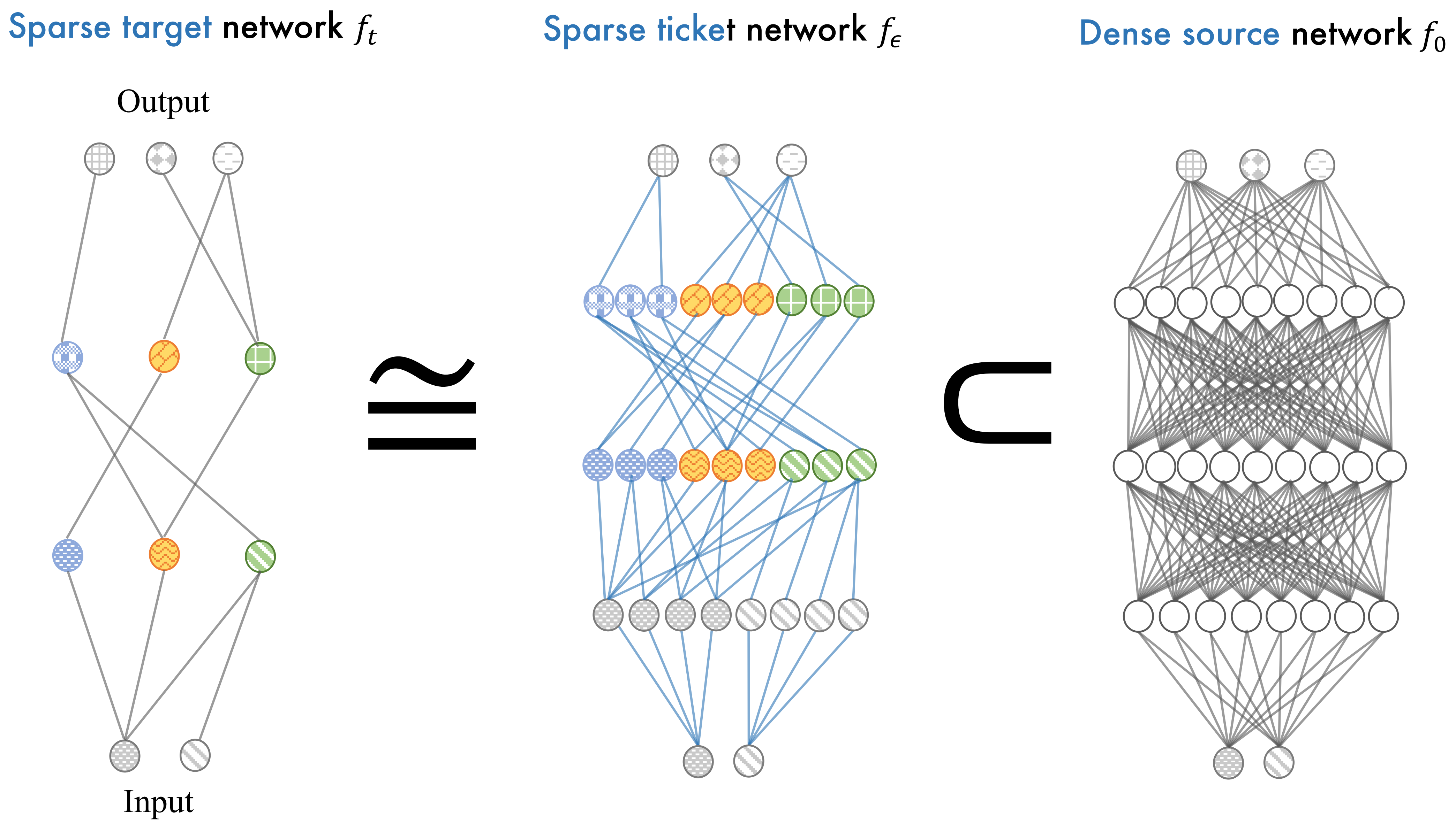
$\cong$

$\subset$

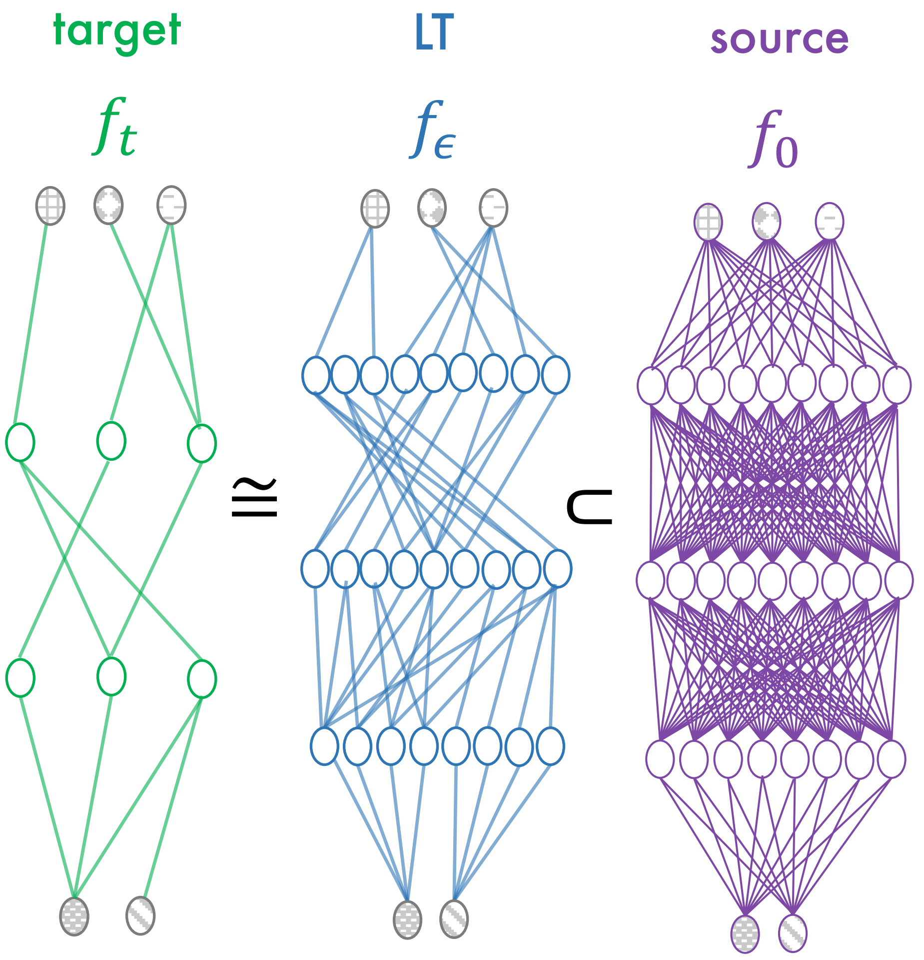
Dense source network f_0



Existence proof set-up



LT existence



Theorem

For every target f_t with width c_t, N_t nonzero params., and depth L_t exists with prob. $1-\delta$ a LT $f_\epsilon \subset f_0$ that approx. f_t up to error $\varepsilon > 0$ if the random source f_0 has depth $L_0 \geq L_t + 1$ and width

$$c_0 \geq c_t \, C \log\left(\frac{c_t N_t}{\min\{\epsilon, \delta\}}\right)$$

Subset sum approx.

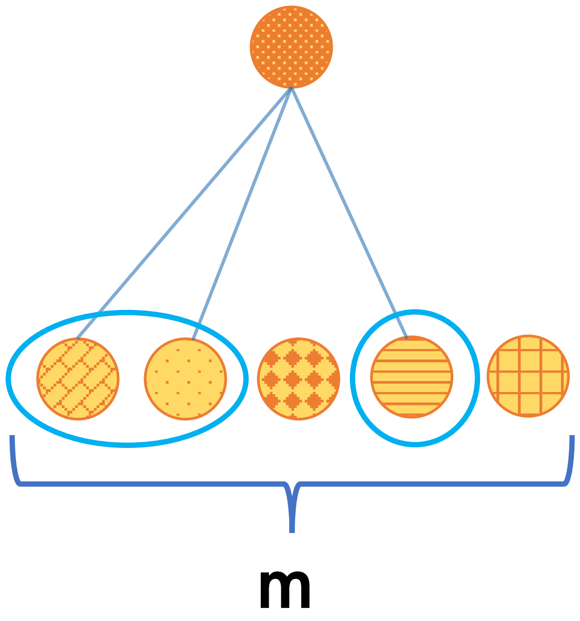
- **Lueker 1998, Pensia et al. 2020.**

Given:
- Target parameter $\theta_t \in [-1, 1]$
- $X_1, \ldots, X_m$ iid, $X_k \sim U[-1, 1]$
- $\varepsilon, \delta \in (0, 1)$

If: $$m \geq C \log\left(\frac{1}{\min(\delta, \varepsilon)}\right)$$

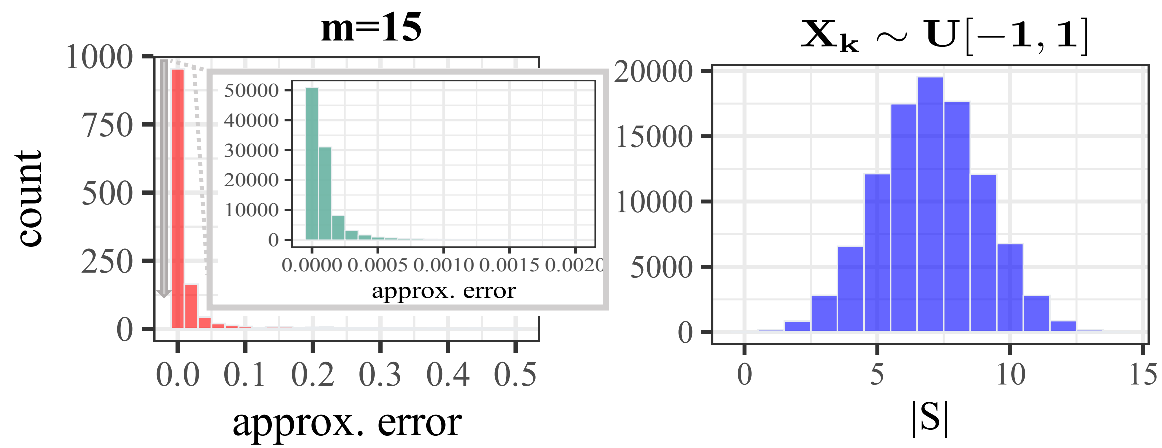
Then: With prob. $1 - \delta$ exists $S \subset [m]$ s.t.

$$|\theta_t - \sum_{k \in S} X_k| \leq \epsilon$$

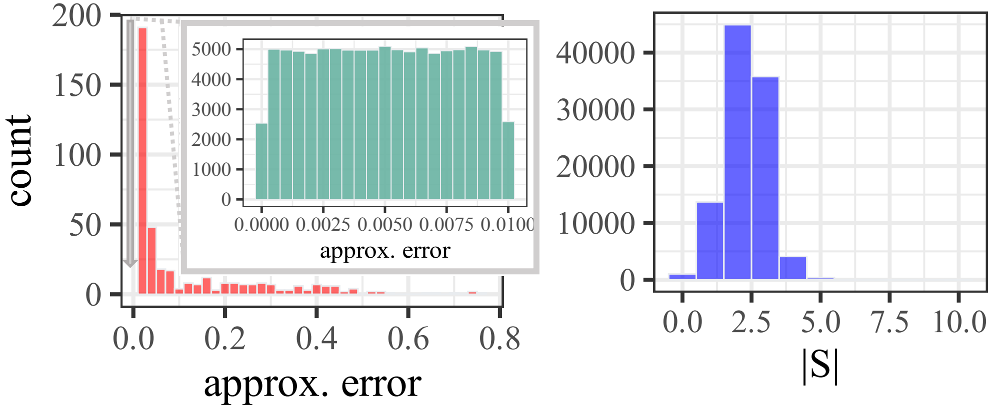


Is that useful?

Best approximation



Smallest subset

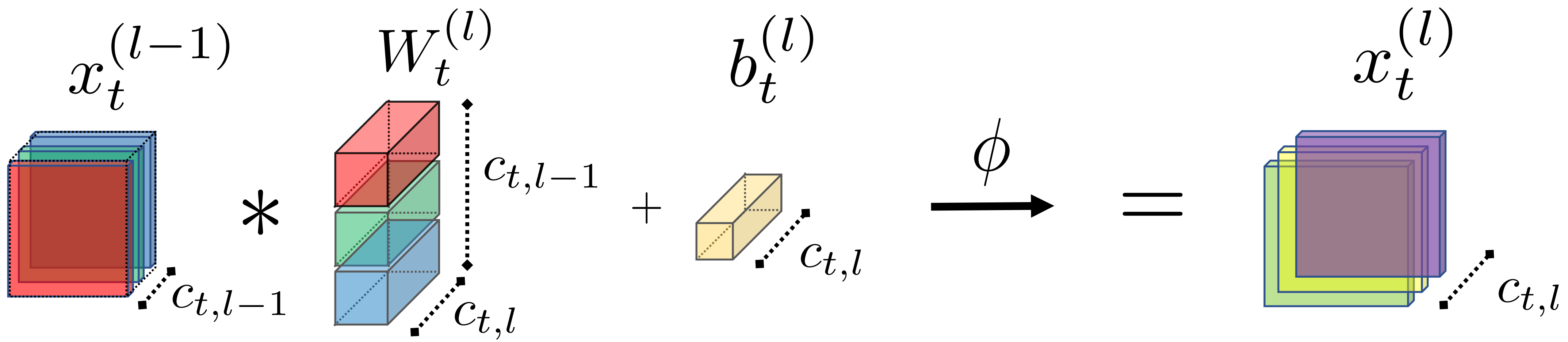


Yes!

- $C \approx 3$.
- Set sizes of m = 15 is more than sufficient.
- On average: $|S| \approx 2 - 3$

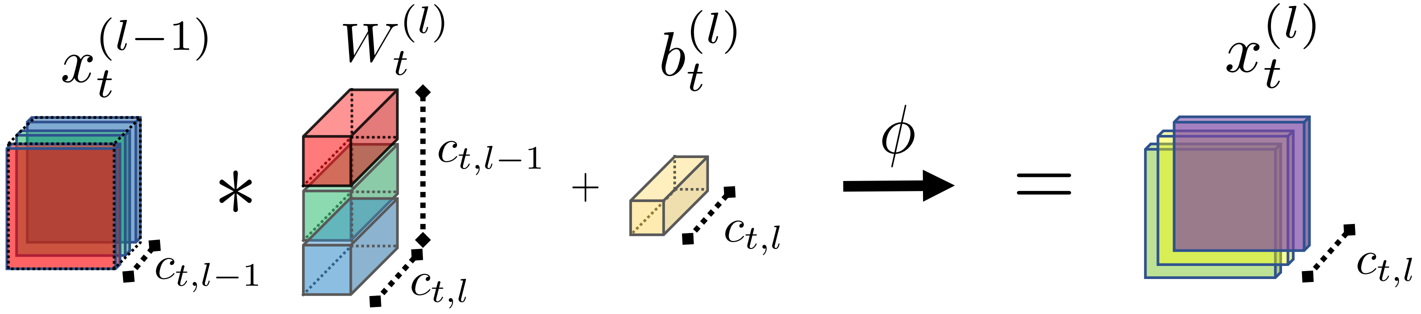
A convolutional layer

Target



Approximating a CNN layer

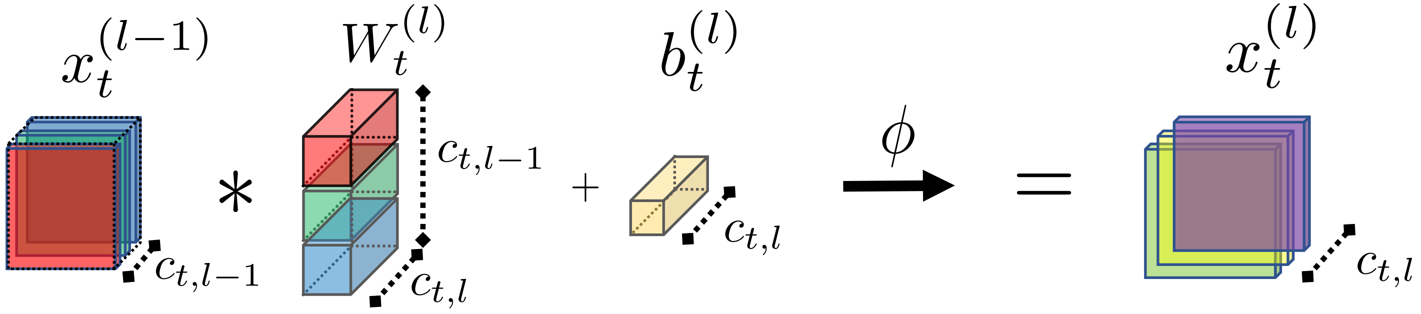
Target



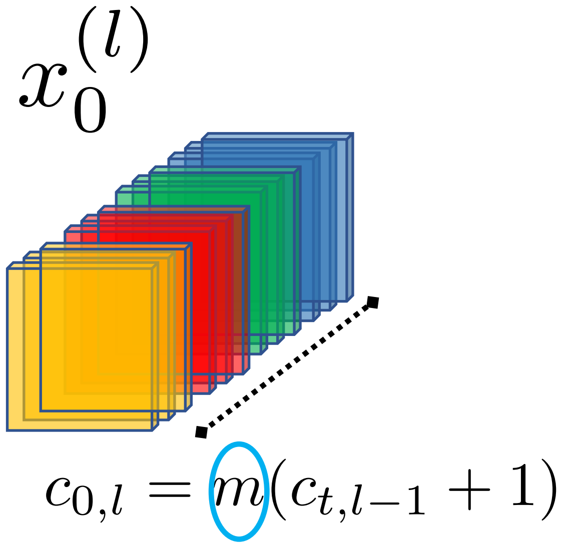
Source

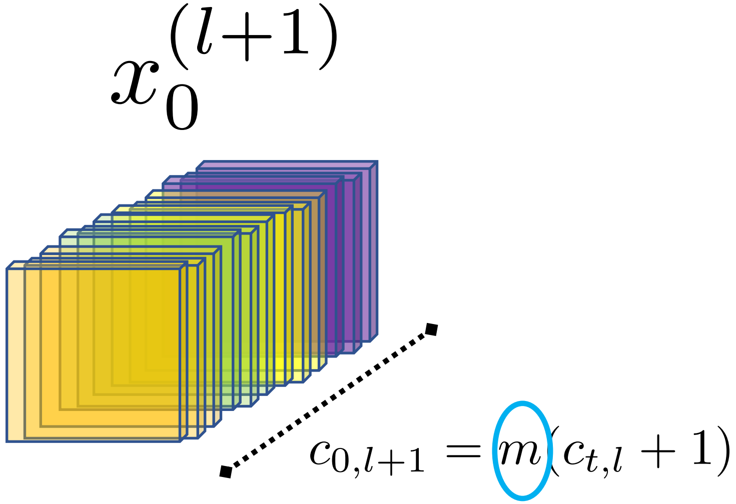
Approximating a CNN layer

Target

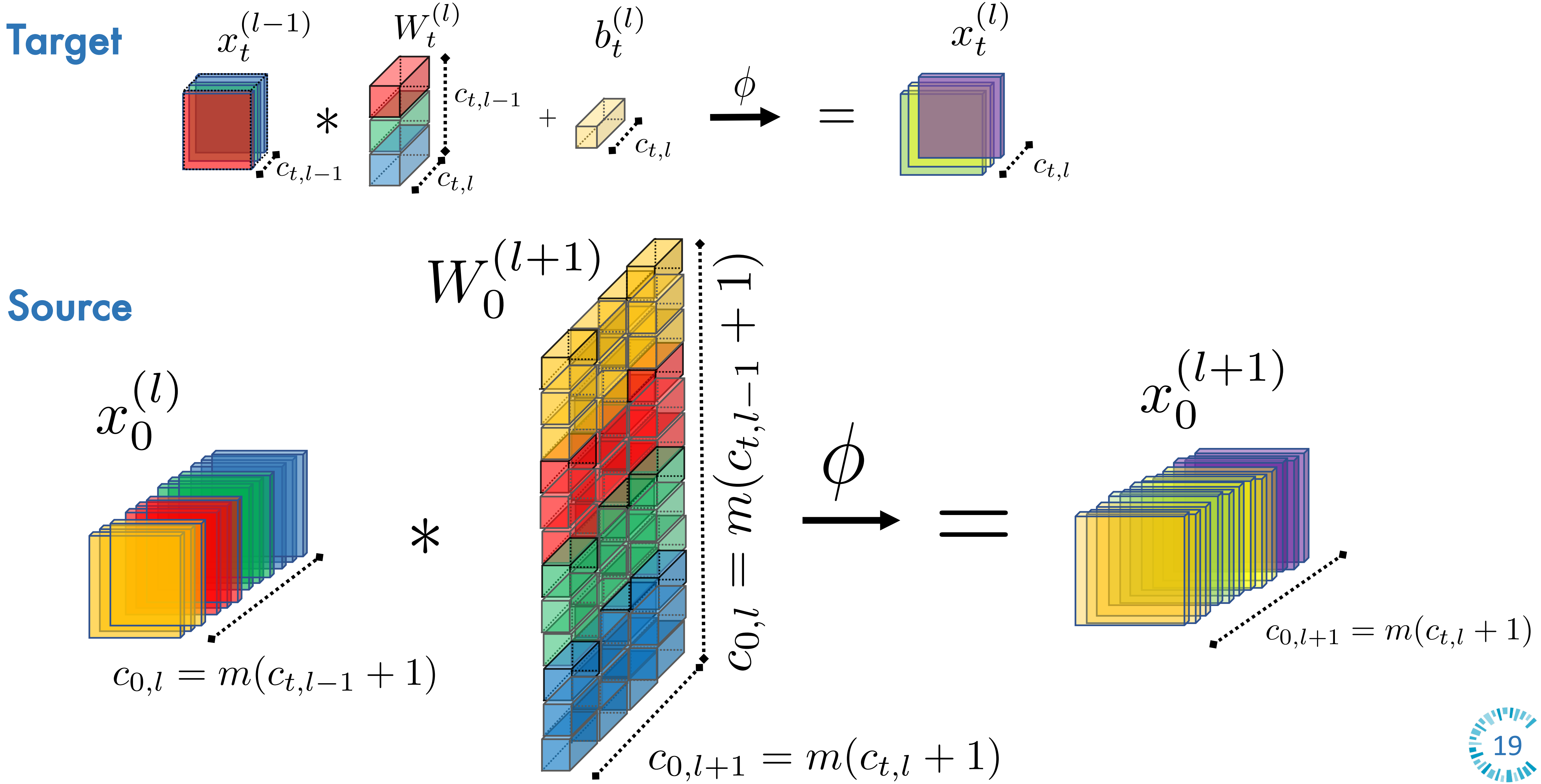


Source

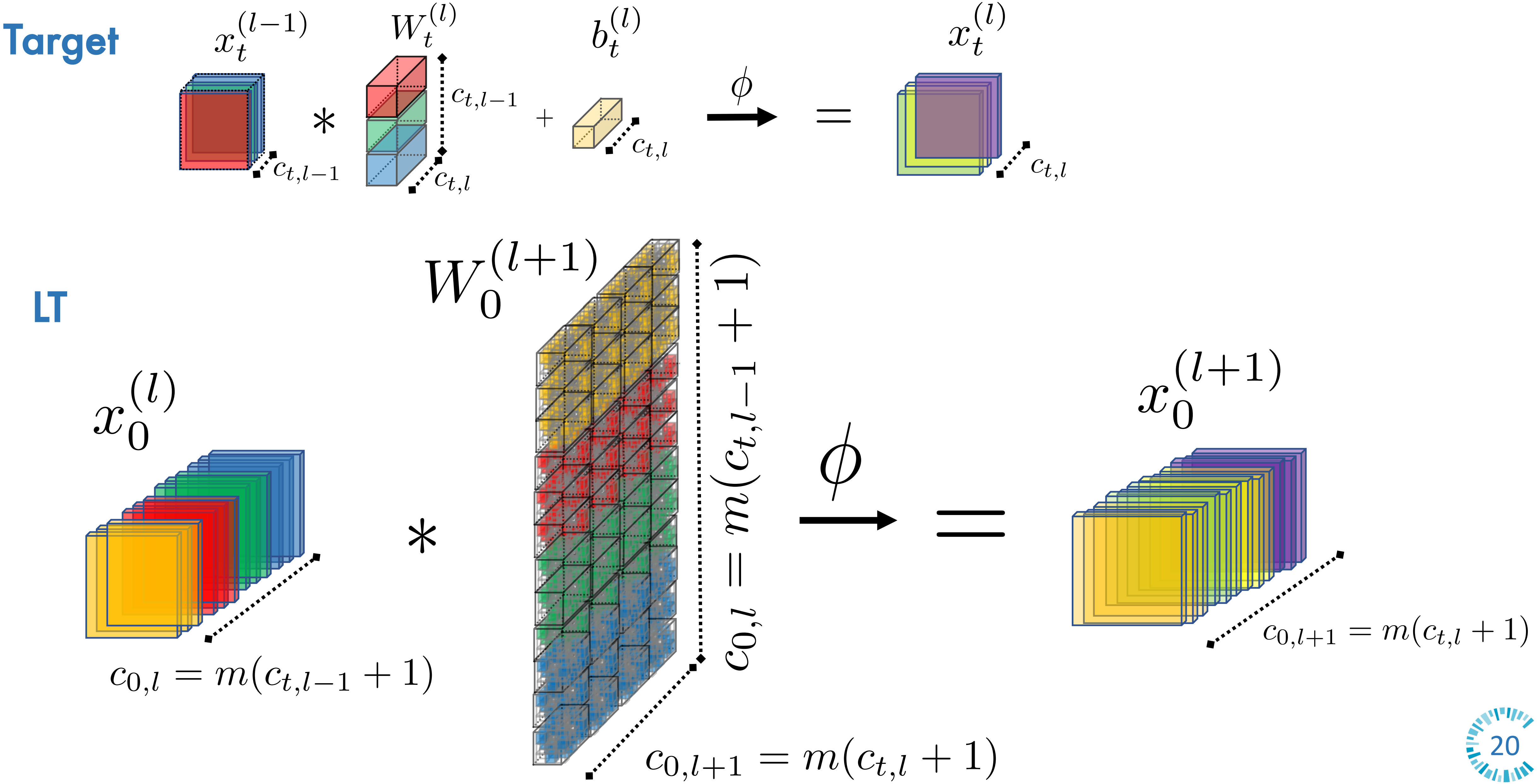




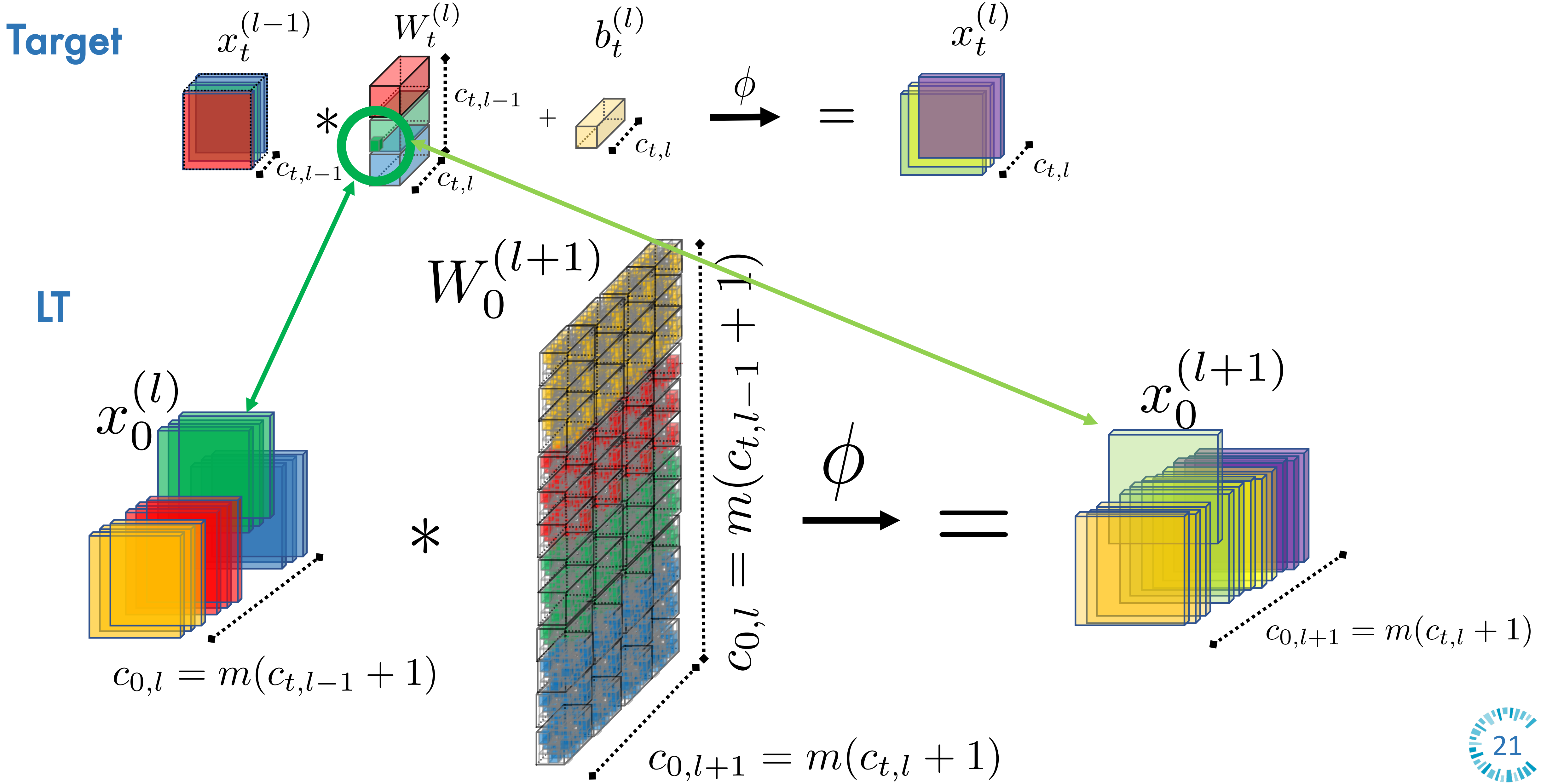
Approximating a CNN layer



Approximating a CNN layer



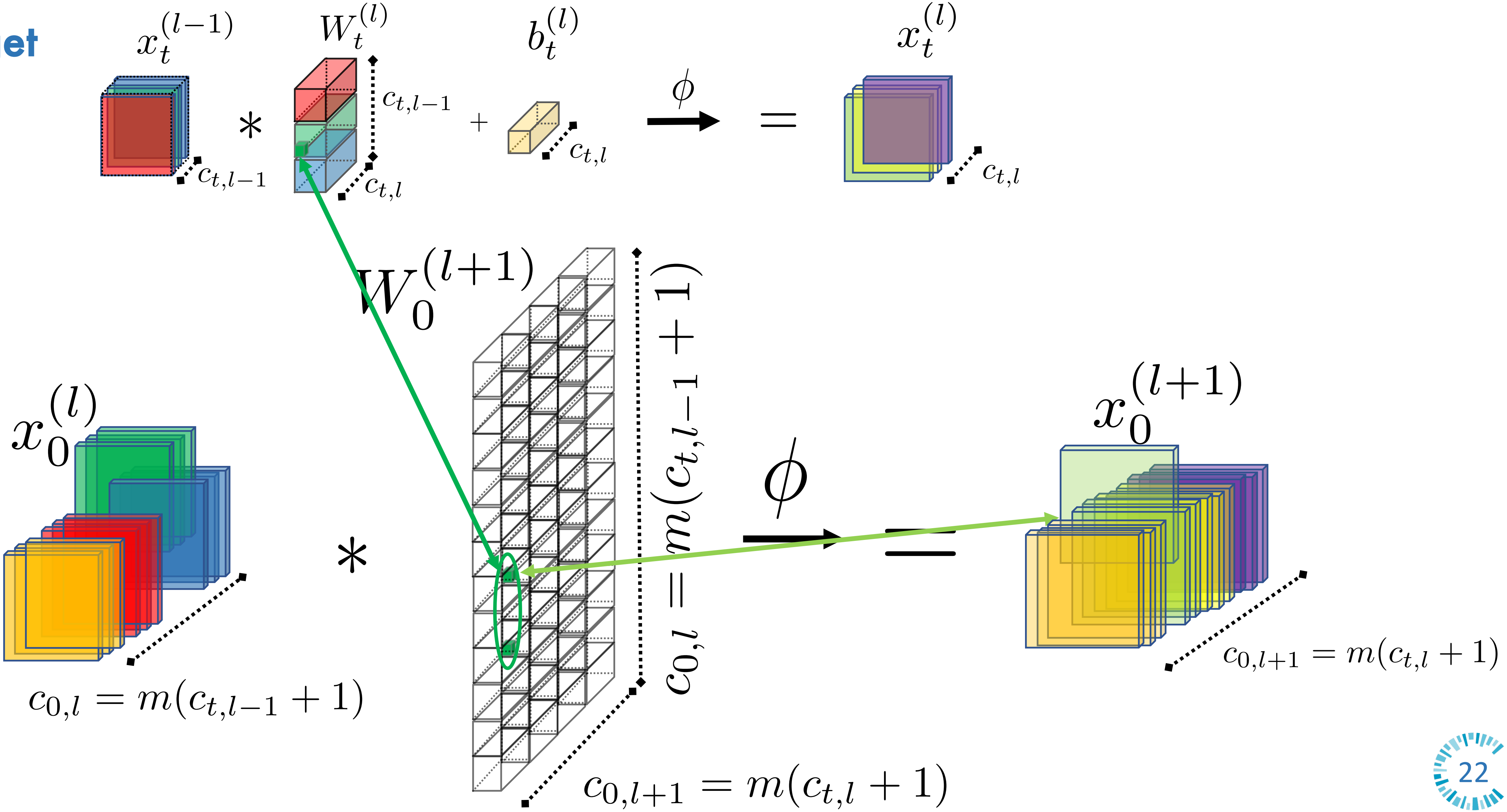
Approximating a CNN layer



Approximating a CNN layer

Target

LT



ResNet-22 target on CIFAR10

- Proofs constructive.
- Defines algorithm to construct LTs that approx. given target.

	Target	LT	LT (best 50%)
ReLU	80.68	80.32 ± 0.13	80.69 ± 0.09
Leaky ReLU	80.68	80.11 ± 0.18	80.6 ± 0.1
Tanh	80.86	80.48 ± 0.13	80.8 ± 0.1
Sigmoid	72.66	69 ± 2	73.2 ± 0.9

Contributions

- Proofs of LT existence for:
  - CNNs with and without skip connections
  - general inputs
  - general activation functions
  - L+1 (instead of 2L) source depth.

- Experimental validation of theory.

- Analysis of subset sum approx.:
  - Width requirement by factor 9-15.