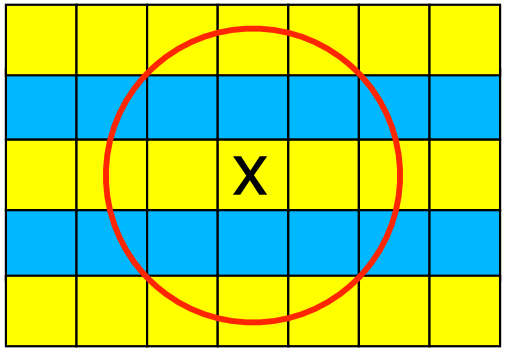
Transformers

Pascal Fua
IC-CVLab

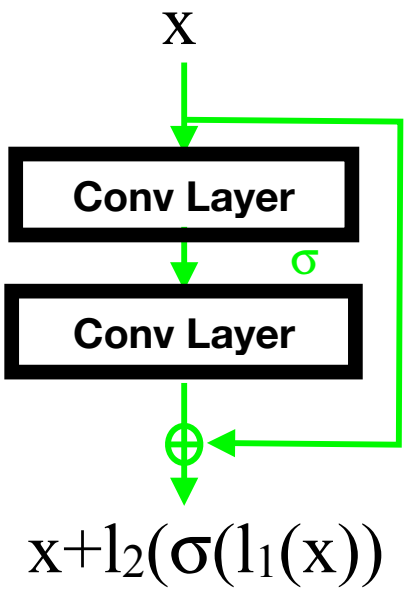
Reminder: Image Specificities



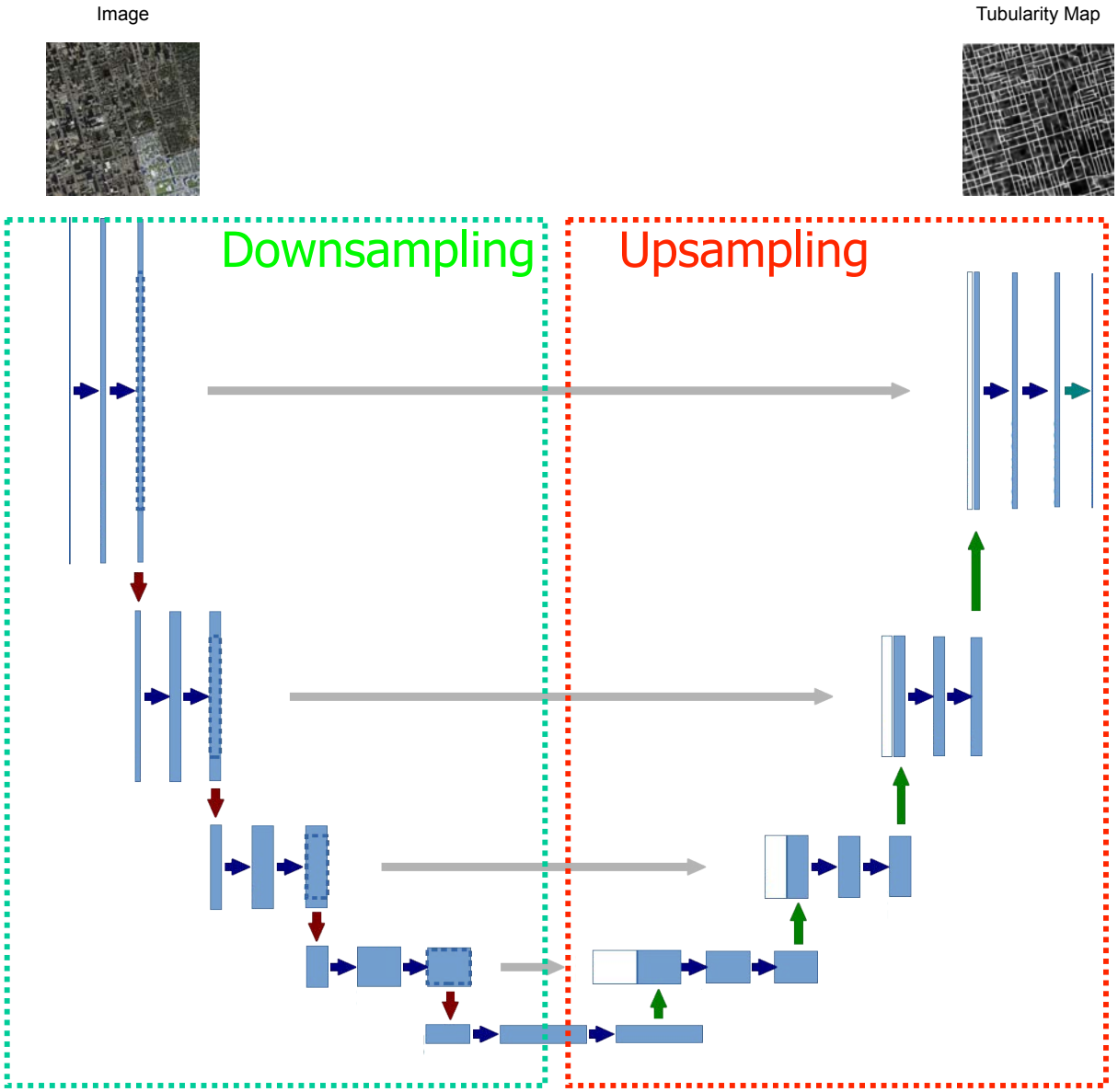
- In a typical image, the values of neighboring pixels tend to be more highly correlated than those of distant ones.
- An image filter should be translation equivariant.

—> These two properties can be exploited to drastically reduce the number of weights required by CNNs using so-called convolutional layers.

Reminder: ResNet to U-Net



ResNet block



U-Net

—> Long range connections are handled via downsampling.

Natural Language Example

The restaurant refused to serve me a ham sandwich, because it only cooks vegetarian food. In the end, they just gave me two slices of bread. Their ambience was just as good as the food and service.

Given your preferences should you go this restaurant?

Seriously?

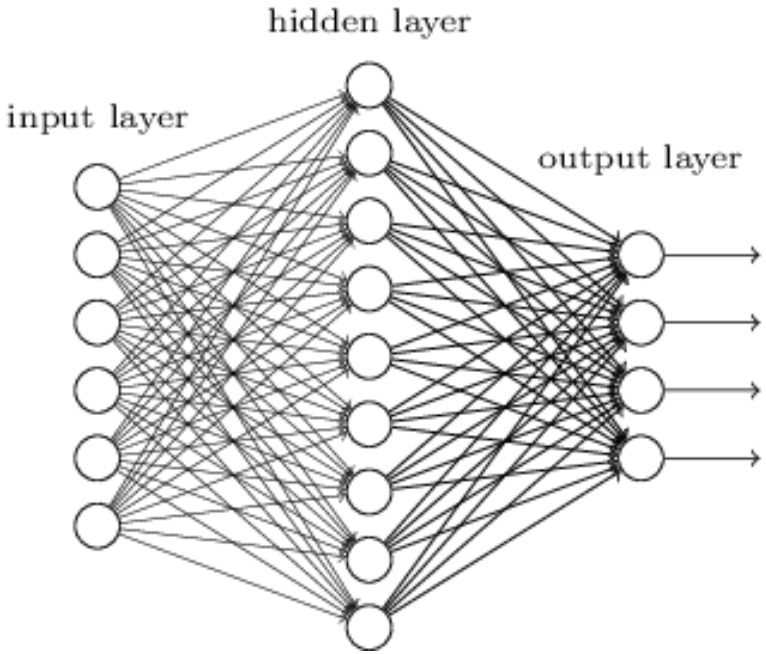
What does "it" refer to?

Context Matters

- Meaning cannot be had from the syntax alone!
- Some words are given “attention” by others.
- Such words are not necessarily close to each other.

—> Assuming that words have been converted to vectors, transformers have been invented to deal with this issue.

Dot-Product Self Attention



Standard fully connected layer

$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b})$$
$$= \sigma(\mathbf{W}\tilde{\mathbf{x}})$$

Given I words $\mathbf{x}_i$, compute the self attention

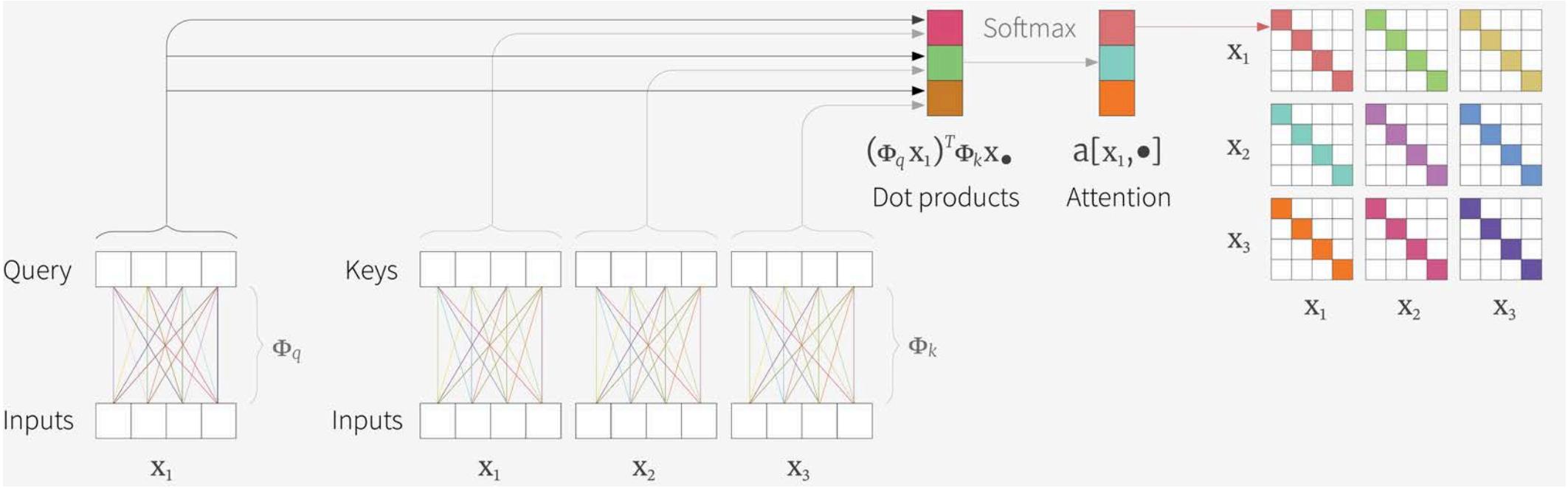
$$\forall i,\ sa[\mathbf{x}_i] = \sum_{j=1}^{I} a[\mathbf{x}_i, \mathbf{x}_j]\mathbf{W}_v\mathbf{x}_j$$

Attention given to $\mathbf{x}_i$

Attention given by $\mathbf{x}_j$ of $\mathbf{x}_i$

Value of $\mathbf{x}_j$

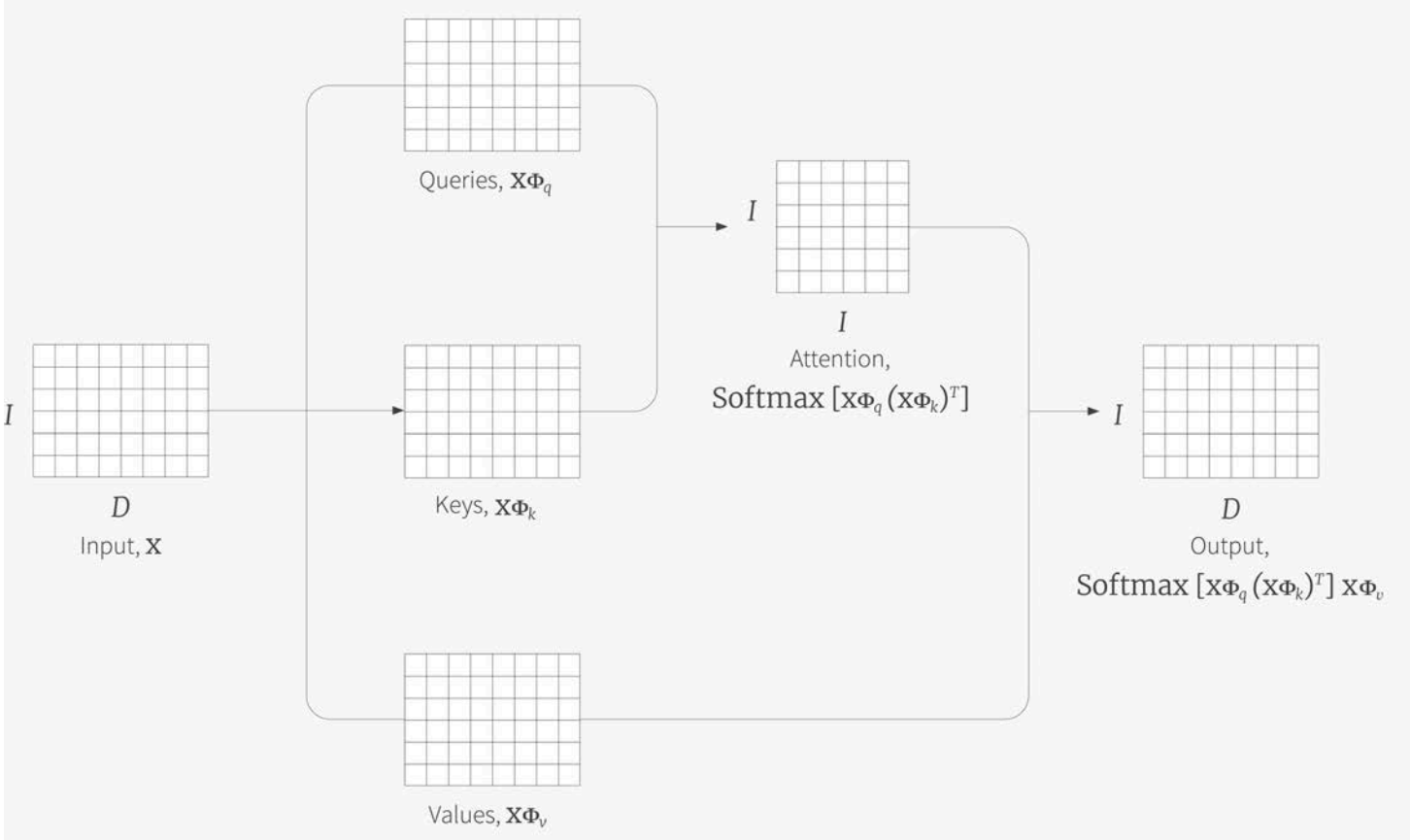
Attention Weights



$$a[\mathbf{x}_i, \mathbf{x}_j] = softmax_j[(\mathbf{W}_q\mathbf{x}_i)^T\mathbf{W}_k\mathbf{x}_j]$$

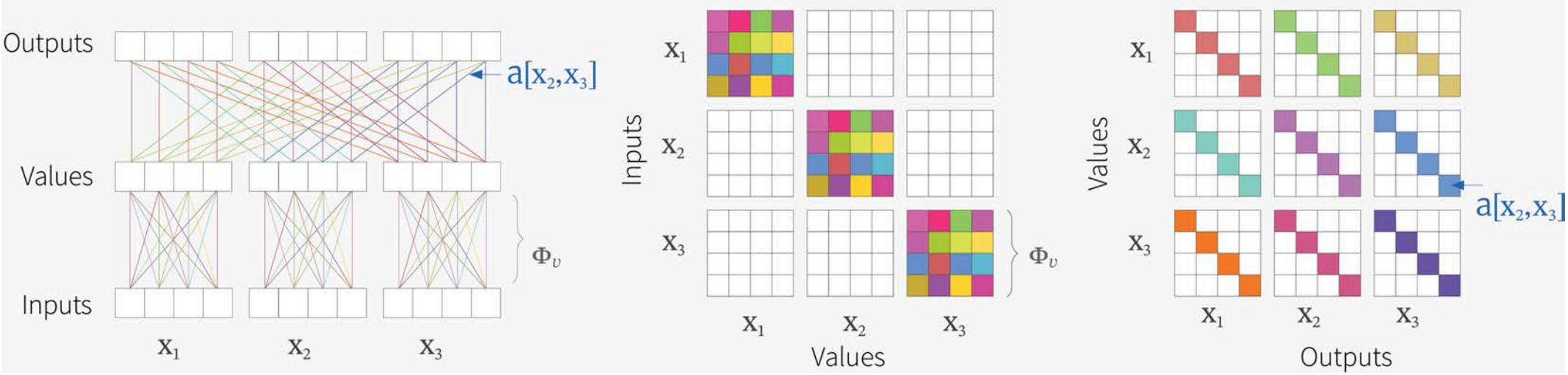
$$= \frac{\exp[(\mathbf{W}_q\mathbf{x}_i)^T\mathbf{W}_k\mathbf{x}_j]}{\sum_{j=1}^{I} exp[(\mathbf{W}_q\mathbf{x}_i)^T\mathbf{W}_k\mathbf{x}_j]}$$

In Matrix Form



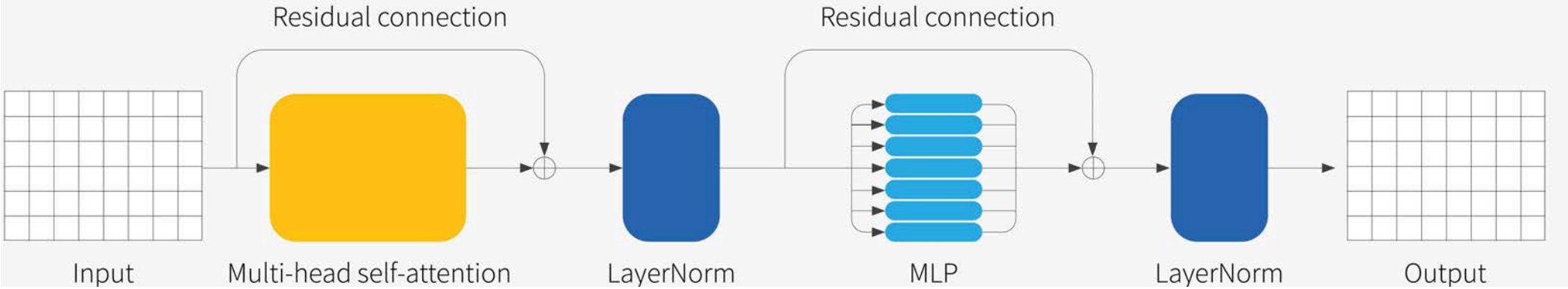
$$Sa(\mathbf{X}) = Softmax[\mathbf{X}\mathbf{W}_q\mathbf{W}_k^T\mathbf{X}^T]\mathbf{X}\mathbf{W}_w$$

Dot-Product Self Attention



Given $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_I}]$, we compute $Sa(\mathbf{X})$ using far fewer weights than if we used a fully connected network.

Transformer Layer



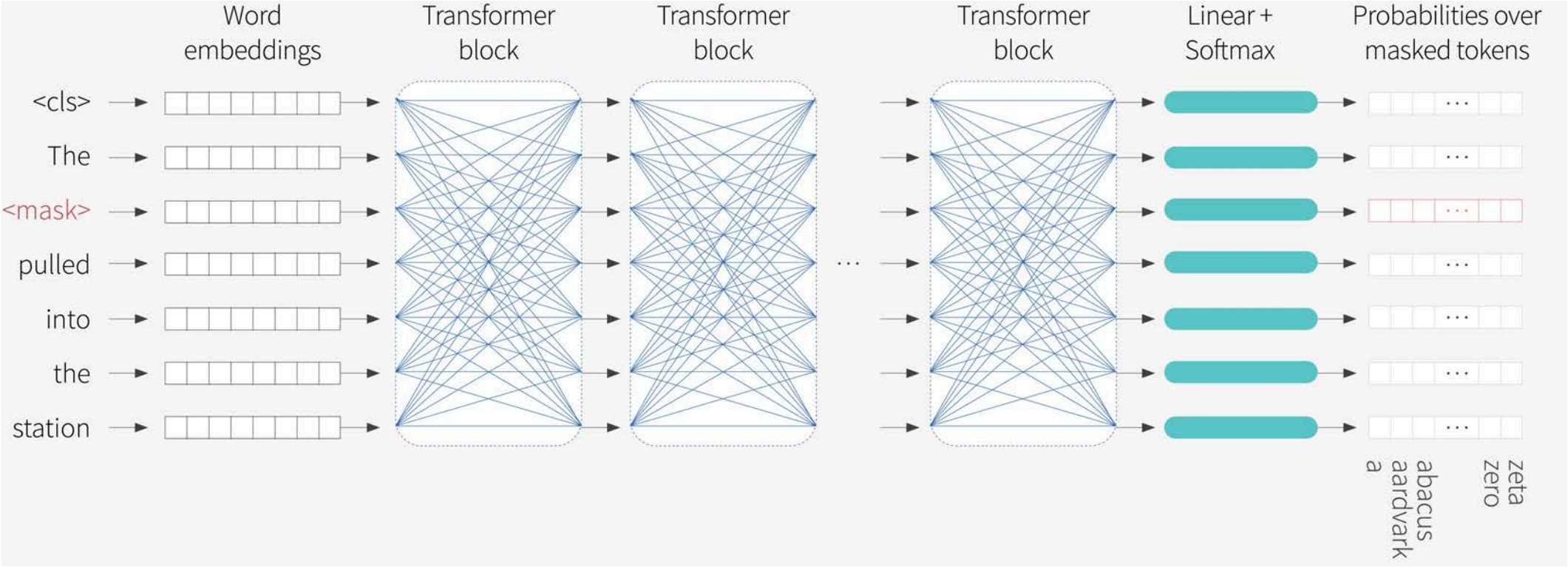
$$\mathbf{X} \leftarrow \mathbf{X} + Sa(\mathbf{X})$$

$$\mathbf{X} \leftarrow LayerNorm(\mathbf{X})$$

$$\mathbf{x}_i \leftarrow \mathbf{x}_i + mlp[\mathbf{x}_i] \quad \forall i$$

$$\mathbf{X} \leftarrow LayerNorm(\mathbf{X})$$

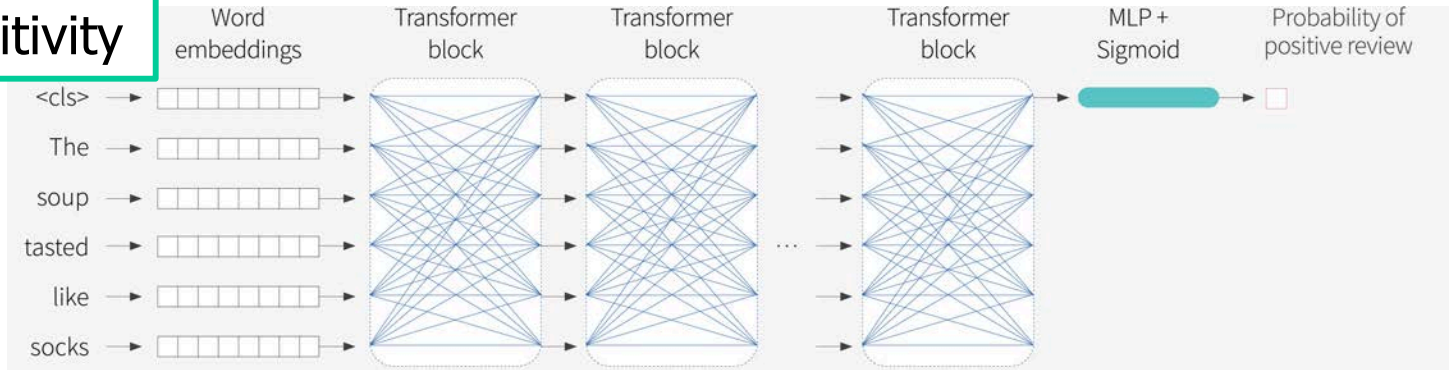
Optional: Bidirectional Encoder Representations from Transformers (BERT)



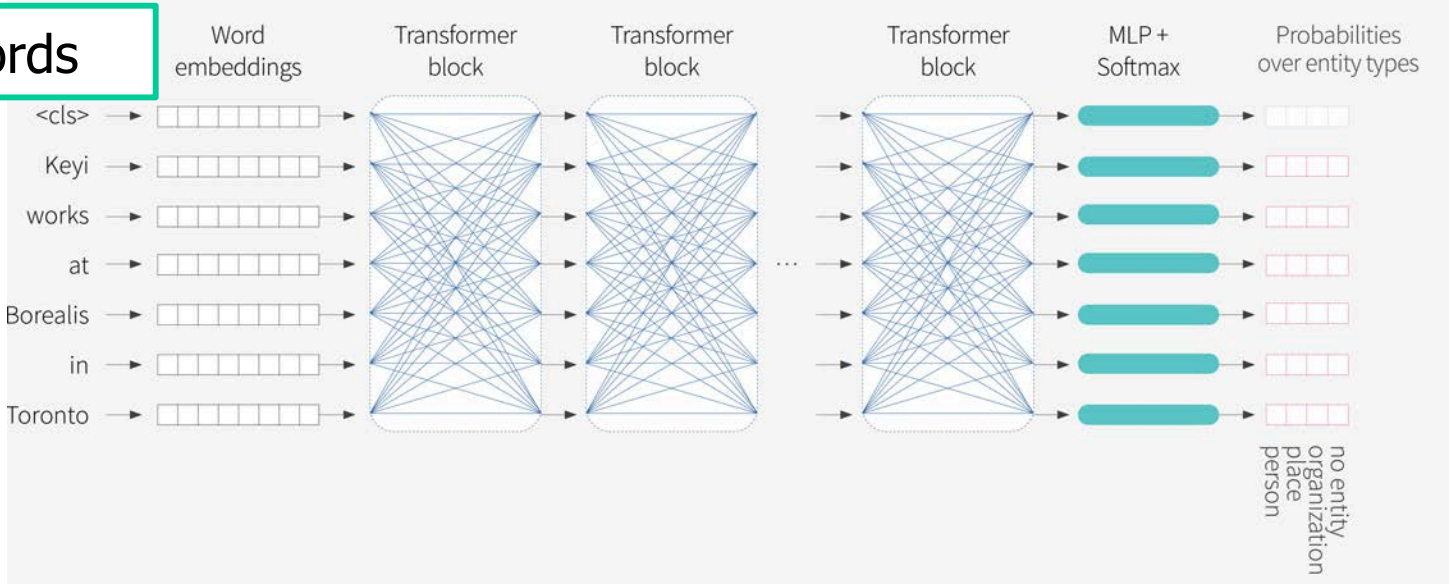
- A sequence of transformer blocks is trained to predict a missing word in a sentence.
- This forces the transformer to learn something about the syntax of the language.

Optional: Bidirectional Encoder Representations from Transformers (BERT)

Rate the positivity



Classify words

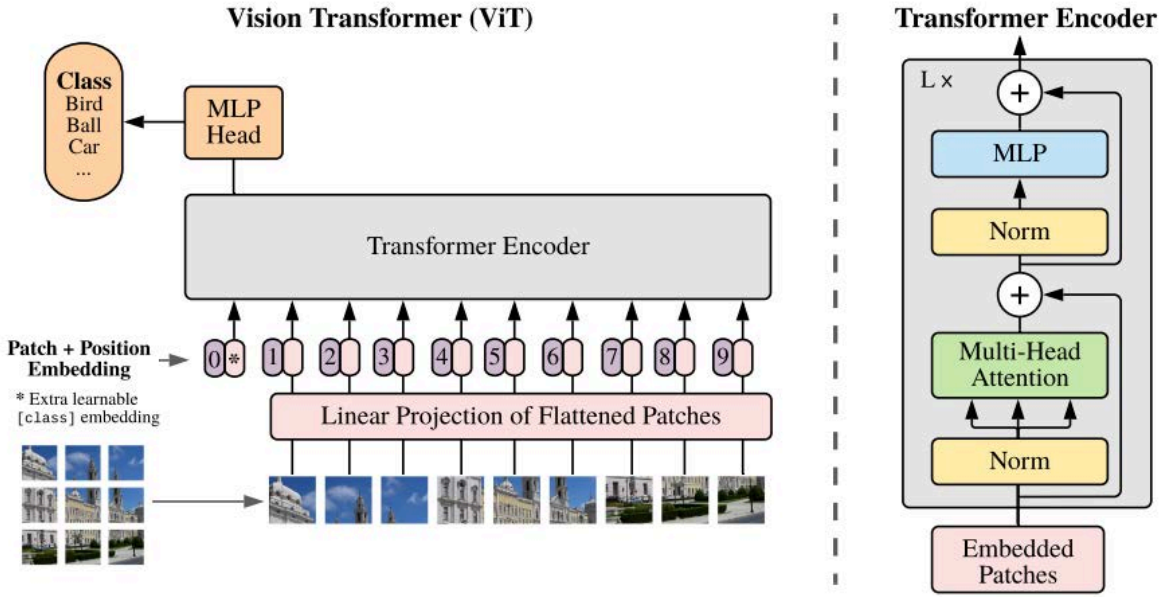


Given additional training data, the representation learned is this way can be used to perform various tasks.

Optional: AI Poetry

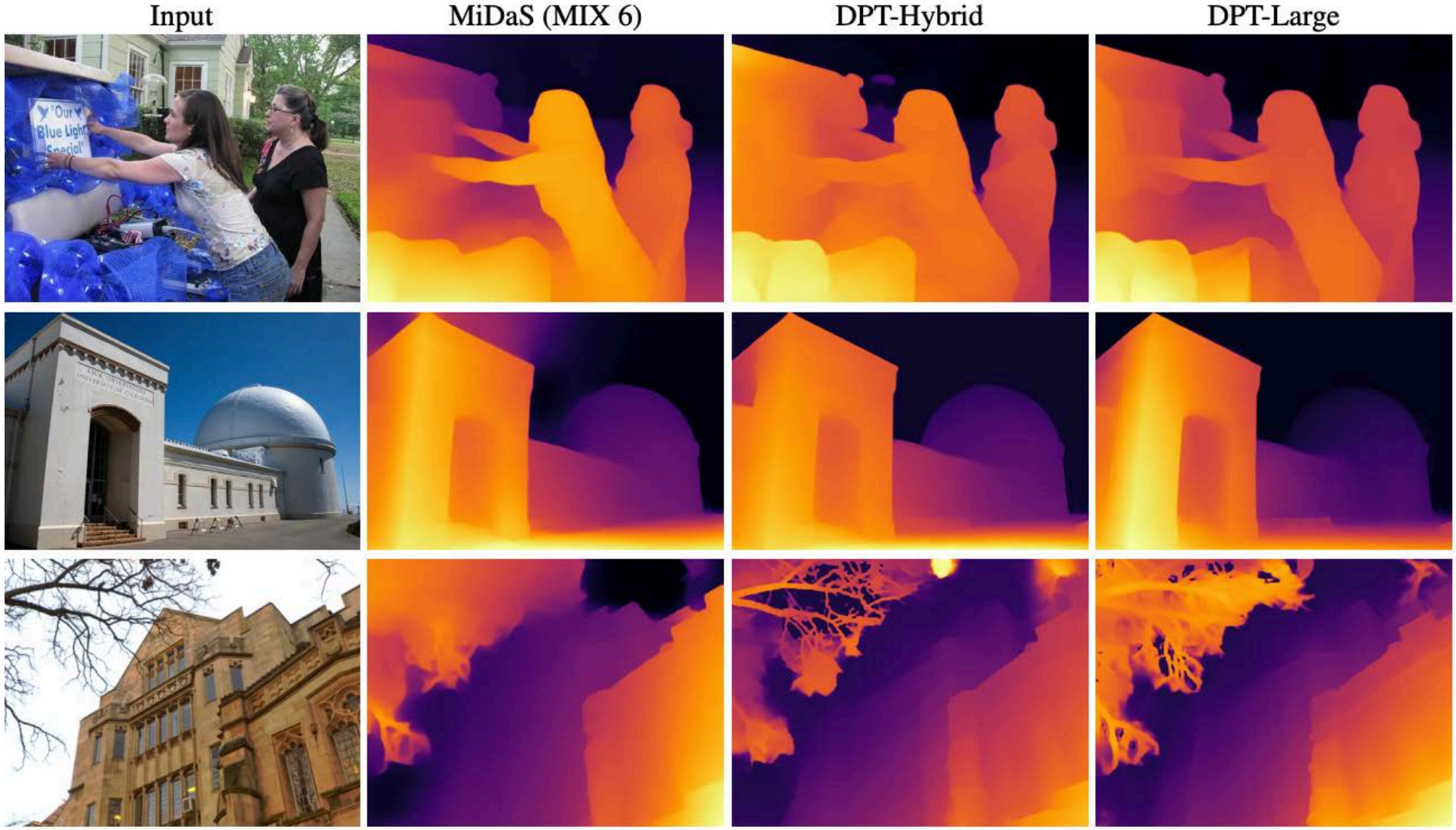
Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date;
A winter's day, when beams of sun are few,
And skies are grey with clouds intemperate,
Is to thy beauty somehow sordid grown.
Her lovely-lovely eye is quite shut up,
Like windows pale before a sunny shower.
The earth that's cold and moist, like to my mistress,
When once that she unbrac'd, her lovely jewel,
Grows as both ugly and uncourtly.
Poor soul, the center of my motley host,
See, now they request thy sovereign presence;
The din of merry hearts hath brought me thus
To greet thee, thou unsoil'd and saint-like beauty.
A Translation by an Artificial Neural Network that draws inspiration from François Connor's translation.
Soften thou, for showers often stain the daylight,
And longer they torment my Mary's day,
But the midday sunshine of thy radiance
Burns in my heart reasonable days.
But rain in an April evening is worse,
No pleasure does disturb the tending rain
That cheeks soon fade and pleasures fade in haste,
When cloudy skies and humid days we wait.
Neither flesh of love nor love's herself my love,
The sun to love is now unfelt, and rare,
My hole sun showing your religion here,
Hastens to go, a blush on your forehead.

Vision Transformers



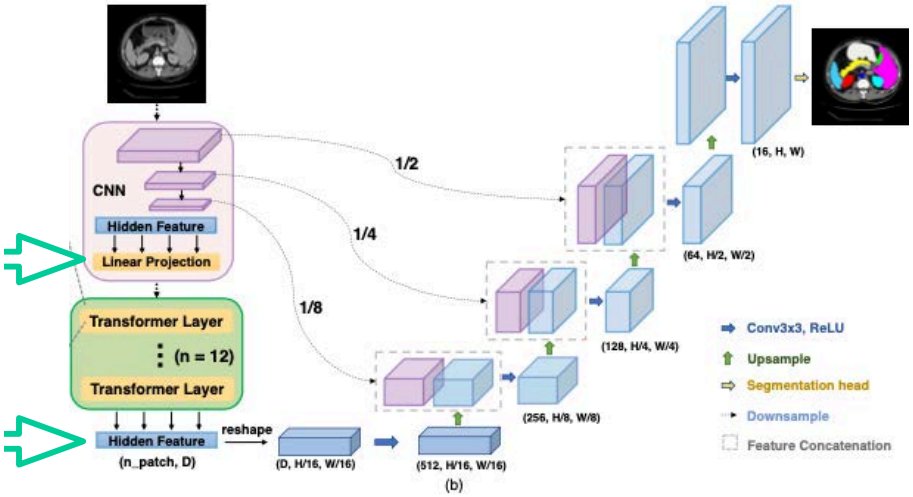
- Break up the images into square patches.
- Transform each path into a feature vector.
- Feed to a transformer architecture.

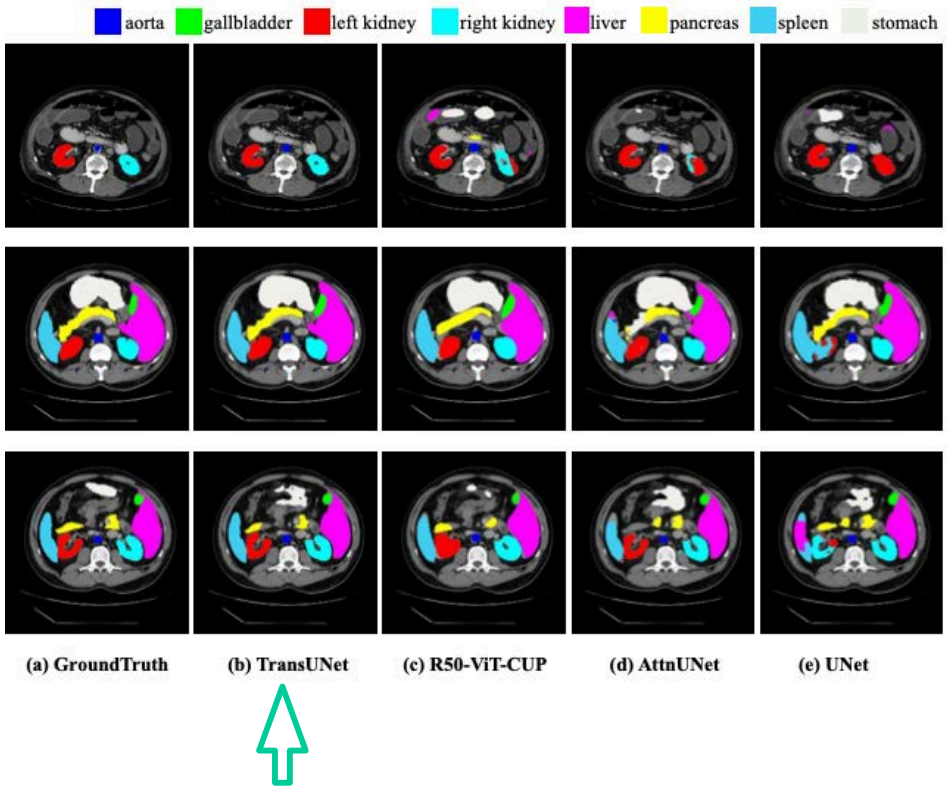
Depth from Single Images



- Pros: Good at modeling long range relationships.
- Cons: Flattening the patches looses some amount of information.

U-NET + Transformers





- A CNN operates at low-resolution and produces a feature vector.
- A transform operates on that feature vector.
- The upsampling is similar to that of U-Net

—> Best of both worlds?