Text Classification

Lena Voita

Lecture-blog and lots of additional materials are here:
https://lena-voita.github.io/nlp_course/text_classification.html

What is going to happen:

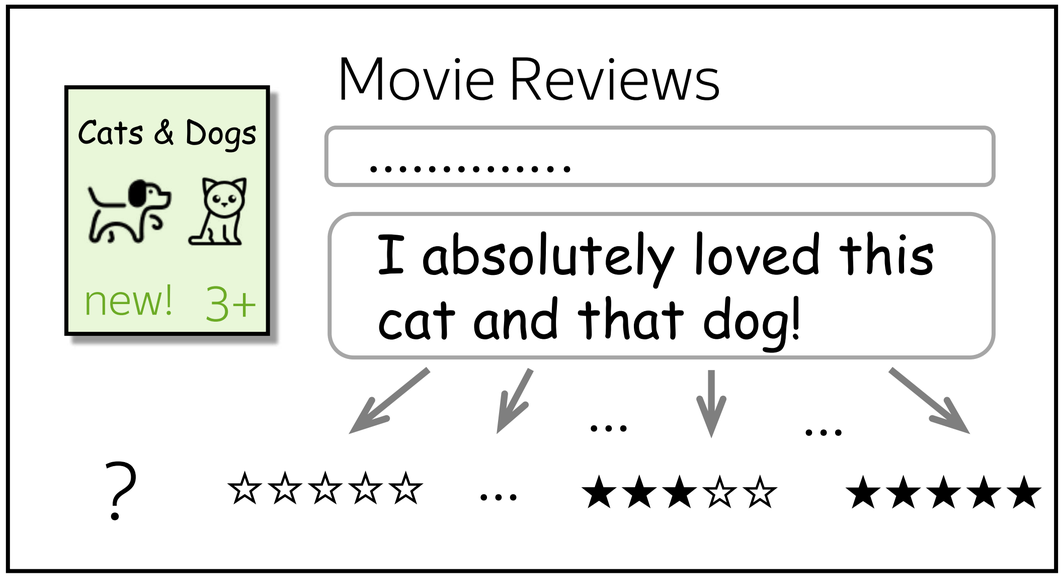
- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

What is going to happen:

- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

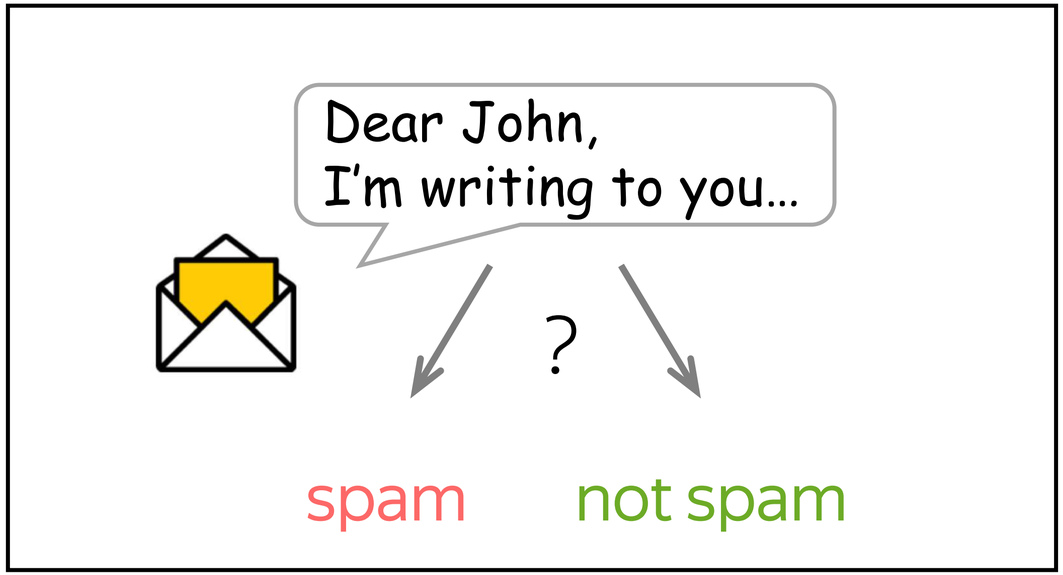
Text Classification

- Multi-class: many labels
- Single label: only one label is correct



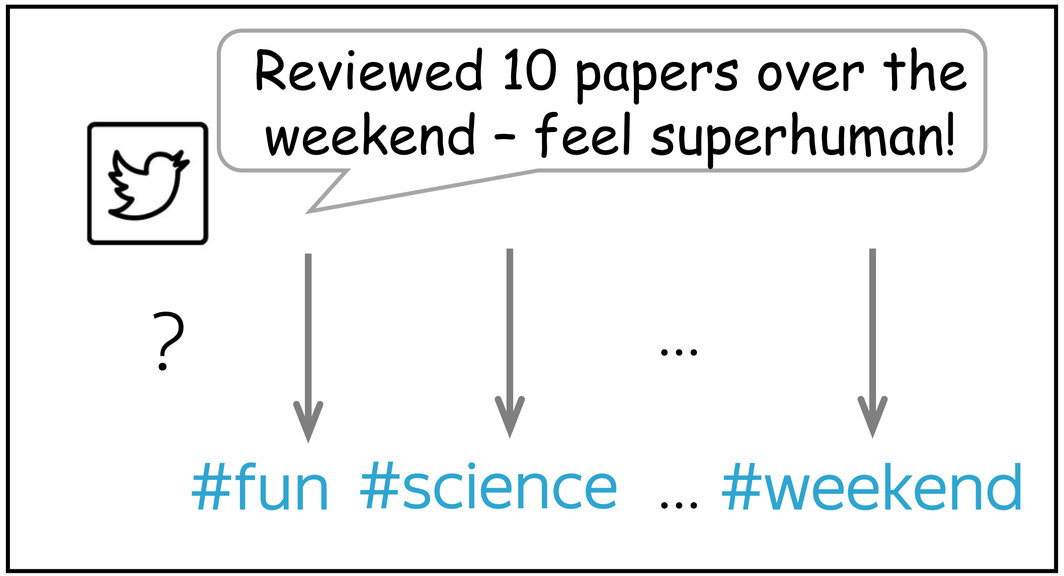
Text Classification

- Binary: two labels
- Single label: only one label is correct



Text Classification

- Multi-class: many labels
- Multi-label: several labels can be correct



Datasets for Text Classification

Datasets vary a lot in:

Dataset	Type	Number of labels	Size (train/test)	Avg. length (tokens)
SST	sentiment	5 or 2	8.5k / 1.1k	19
IMDb Review	sentiment	2	25k / 25k	271
Yelp Review	sentiment	5 or 2	650k / 50k	179
Amazon Review	sentiment	5 or 2	3m / 650k	79
TREC	question	6	5.5k / 0.5k	10
Yahoo! Answers	question	10	1.4m / 60k	131
AG's News	topic	4	120k / 7.6k	44
Sogou News	topic	6	54k / 6k	737
DBPedia	topic	14	560k / 70k	67

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification

Datasets vary a lot in:

- Type

Dataset	Type	Number of labels	Size (train/test)	Avg. length (tokens)
SST	sentiment	5 or 2	8.5k / 1.1k	19
IMDb Review	sentiment	2	25k / 25k	271
Yelp Review	sentiment	5 or 2	650k / 50k	179
Amazon Review	sentiment	5 or 2	3m / 650k	79
TREC	question	6	5.5k / 0.5k	10
Yahoo! Answers	question	10	1.4m / 60k	131
AG's News	topic	4	120k / 7.6k	44
Sogou News	topic	6	54k / 6k	737
DBPedia	topic	14	560k / 70k	67

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification

Datasets vary a lot in:

- Type
- Number of labels

Dataset	Type	Number of labels	Size (train/test)	Avg. length (tokens)
SST	sentiment	5 or 2	8.5k / 1.1k	19
IMDb Review	sentiment	2	25k / 25k	271
Yelp Review	sentiment	5 or 2	650k / 50k	179
Amazon Review	sentiment	5 or 2	3m / 650k	79
TREC	question	6	5.5k / 0.5k	10
Yahoo! Answers	question	10	1.4m / 60k	131
AG's News	topic	4	120k / 7.6k	44
Sogou News	topic	6	54k / 6k	737
DBPedia	topic	14	560k / 70k	67

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification

Datasets vary a lot in:

- Type
- Number of labels
- Dataset size

Dataset	Type	Number of labels	Size (train/test)	Avg. length (tokens)
SST	sentiment	5 or 2	8.5k / 1.1k	19
IMDb Review	sentiment	2	25k / 25k	271
Yelp Review	sentiment	5 or 2	650k / 50k	179
Amazon Review	sentiment	5 or 2	3m / 650k	79
TREC	question	6	5.5k / 0.5k	10
Yahoo! Answers	question	10	1.4m / 60k	131
AG's News	topic	4	120k / 7.6k	44
Sogou News	topic	6	54k / 6k	737
DBPedia	topic	14	560k / 70k	67

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification

Datasets vary a lot in:

- Type
- Number of labels
- Dataset size
- Example length

Dataset	Type	Number of labels	Size (train/test)	Avg. length (tokens)
SST	sentiment	5 or 2	8.5k / 1.1k	19
IMDb Review	sentiment	2	25k / 25k	271
Yelp Review	sentiment	5 or 2	650k / 50k	179
Amazon Review	sentiment	5 or 2	3m / 650k	79
TREC	question	6	5.5k / 0.5k	10
Yahoo! Answers	question	10	1.4m / 60k	131
AG's News	topic	4	120k / 7.6k	44
Sogou News	topic	6	54k / 6k	737
DBPedia	topic	14	560k / 70k	67

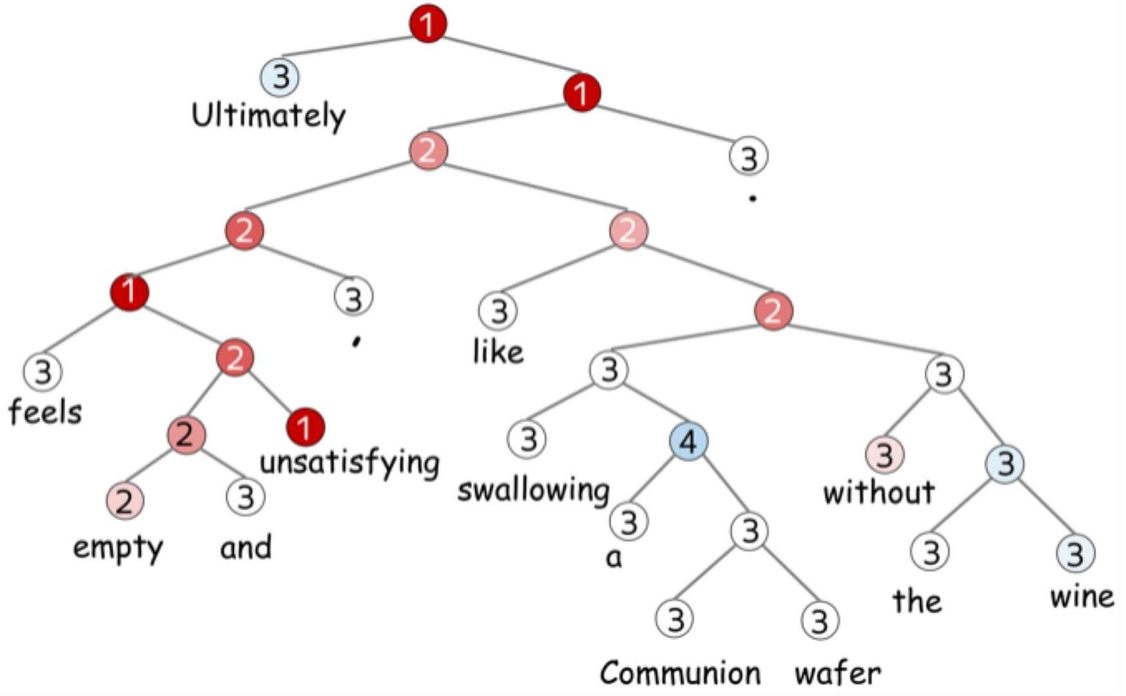
https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification: Sentiment

Pick a dataset

SST | IMDb Review | Yelp Review | Amazon Review
TREC | Yahoo! Answers | AG's News | Sogou News | DBPedia

Label: 1
Review:
Ultimately feels empty and unsatisfying , like swallowing a Communion wafer without the wine .



https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification: Sentiment

Pick a dataset

- SST
- (•) IMDb Review
- Yelp Review
- Amazon Review
- TREC
- Yahoo! Answers
- AG's News
- Sogou News
- DBPedia

Label: negative

Review
Hobgoblins .... Hobgoblins .... where do I begin?!?

This film gives Manos - The Hands of Fate and Future War a run for their money as the worst film ever made . This one is fun to laugh at , where as Manos was just painful to watch . Hobgoblins will end up in a time capsule somewhere as the perfect movie to describe the term : " 80 's cheeze " . The acting (and I am using this term loosely) is atrocious , the Hobgoblins are some of the worst puppets you will ever see , and the garden tool fight has to be seen to be believed . The movie was the perfect vehicle for MST3 K , and that version is the only way to watch this mess . This movie gives Mike and the bots lots of ammunition to pull some of the funniest one - liners they have ever done . If you try to watch this without the help of Mike and the bots ..... God help you ! !

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification: Sentiment

Pick a dataset

○ SST ○ IMDb Review ◉ Yelp Review ○ Amazon Review
○ TREC ○ Yahoo! Answers ○ AG's News ○ Sogou News ○ DBPedia

Label: 4

Review
I had a serious craving for Roti. So glad I found this place. A very small menu selection but it had exactly what I wanted. The serving for $8.20 after tax is enough for 2 meals. I know where to go from now on for a great meal with leftovers. This is a noteworthy place to bring my Uncle T.J. who's a Trini when he comes to visit.

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification: Sentiment

Pick a dataset

- SST
- IMDb Review
- Yelp Review
- (•) Amazon Review
- TREC
- Yahoo! Answers
- AG's News
- Sogou News
- DBPedia

Label: 3

Review Title: Simple

Review Content:
This book was not anything special. Although I love romances, it was too simple. The symbolism was spelled out to the readers in a blunt manner. The less educated readers may appreciate it. The wording was quite beautiful at times and the plot was enchanting (perfect for a movie) but it is not heart wrenching like the movie Titantic (which was a must see!) ;)

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification: Questions

Pick a dataset

○ SST ○ IMDb Review ○ Yelp Review ○ Amazon Review
○ TREC ◉ Yahoo! Answers ○ AG's News ○ Sogou News ○ DBPedia

Label: Society & Culture

Question Title: Why do people have the bird, turkey for thanksgiving?

Question Content: Why this bird? Any Significance?

Best Answer
It is believed that the pilgrims and indians shared wild turkey and venison on the original Thanksgiving.

Turkey's "Americanness" was established by Benjamin Franklin, who had advocated for the turkey, not the bald eagle, becoming the national bird.

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification: Topic

Pick a dataset

SST | IMDb Review | Yelp Review | Amazon Review
TREC | Yahoo! Answers | (•) AG's News | Sogou News | DBPedia

Label: Sports

Title: Schumacher Triumphs as Ferrari Seals Formula One Title

Description
BUDAPEST (Reuters) - Michael Schumacher cruised to a record 12th win of the season in the Hungarian Grand Prix on Sunday to hand his Ferrari team a sixth successive constructors' title.

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

Datasets for Text Classification: Topic

Pick a dataset

○ SST ○ IMDb Review ○ Yelp Review ○ Amazon Review
○ TREC ○ Yahoo! Answers ○ AG's News ○ Sogou News ◉ DBPedia

Label: Artist

Title: Esfandiar Monfaredzadeh

Abstract
Esfandiar Monfaredzadeh (Persian : اسفندیار منفردزاده) is an Iranian composer and director. He was born in 1941 in Tehran His major works are Gheisar Dash Akol Tangna Gavaznha. He has 2 daughters Bibinaz Monfaredzadeh and Sanam Monfaredzadeh Woods (by marriage).

https://lena-voita.github.io/nlp_course/text_classification.html#dataset_examples

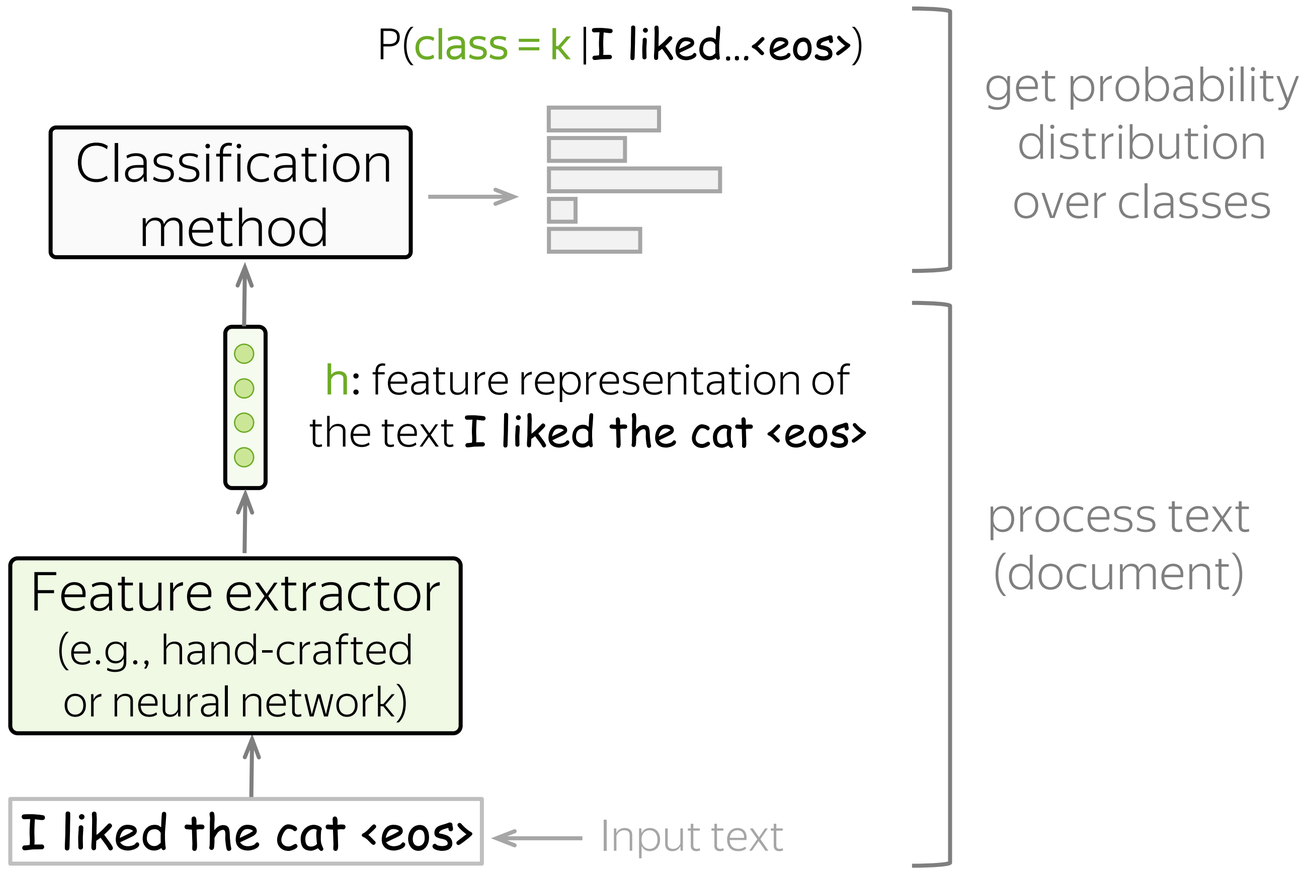
What is going to happen:

- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

What is going to happen:

- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

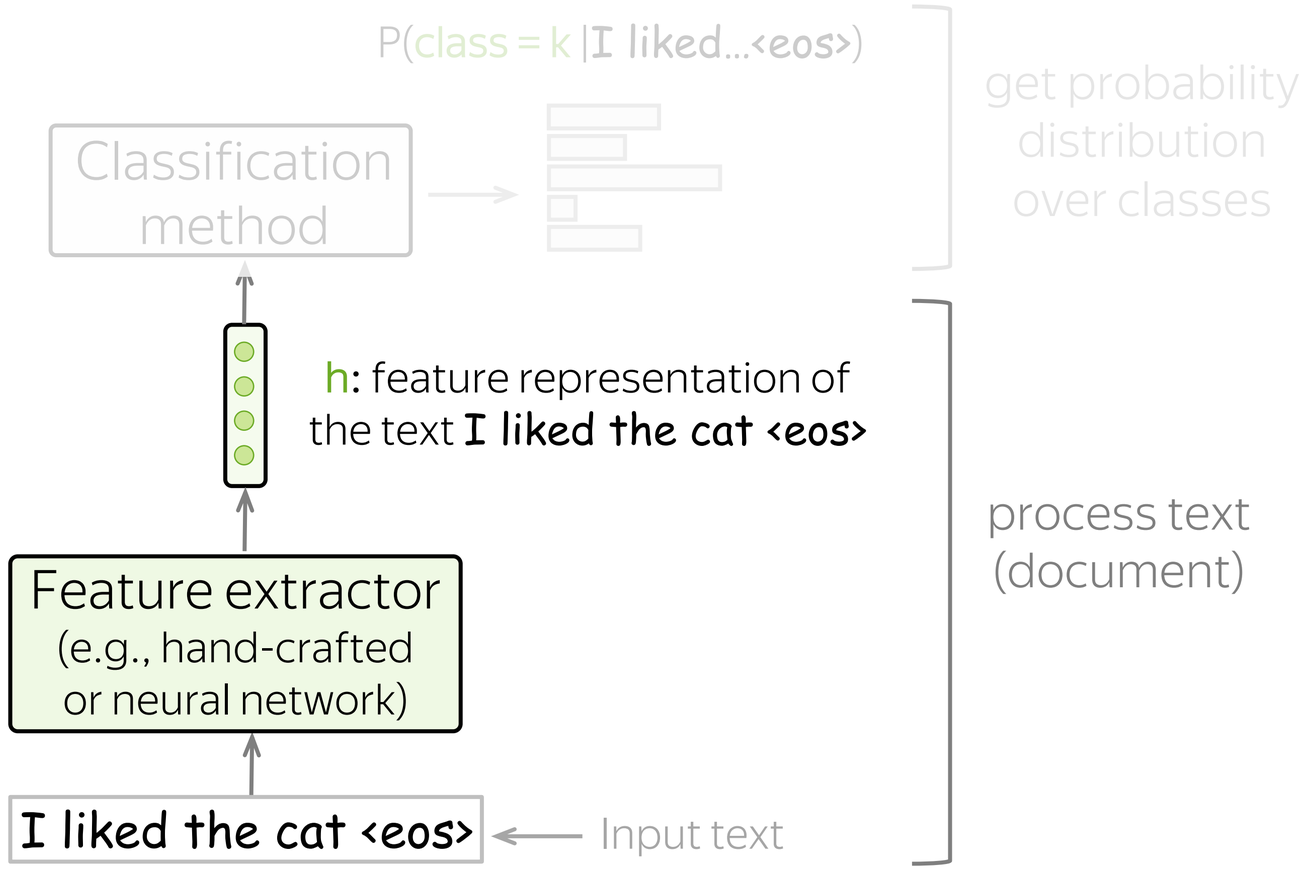
Get Feature Representation and Classify



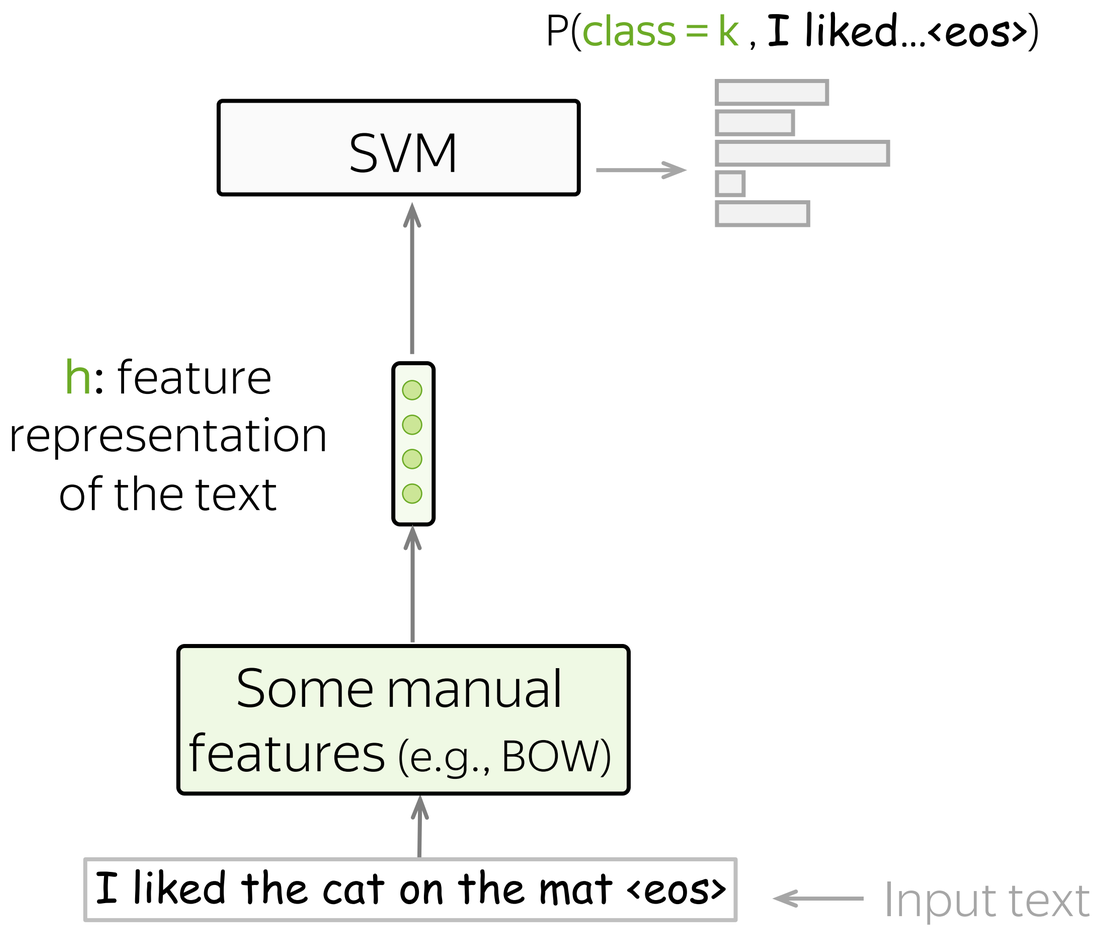
Get Feature Representation and Classify

Feature extractor:

- **classical methods** – features are extracted manually by a human
- **neural methods** – features are extracted by a network: a network **learns** what is important

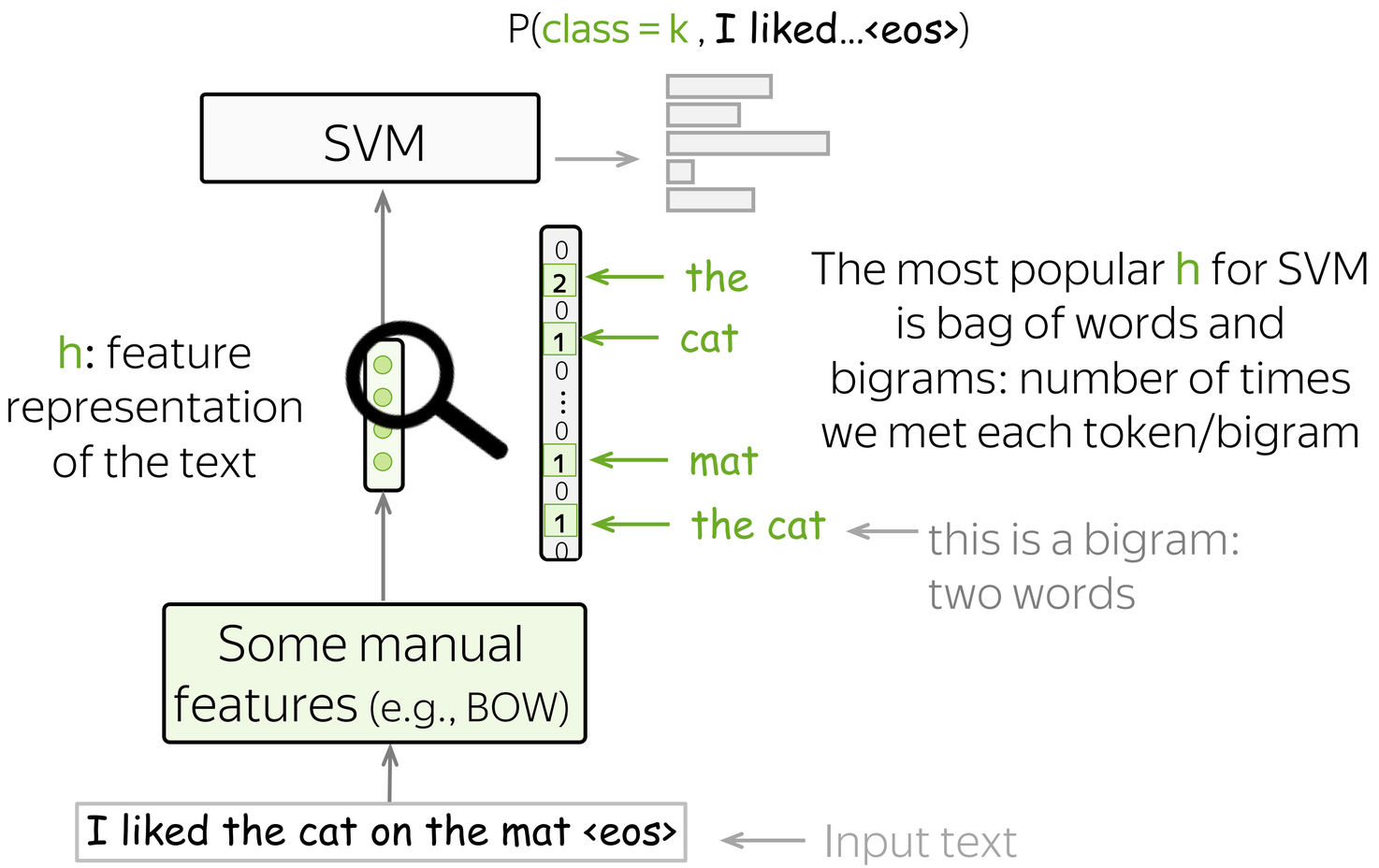


SVM: Get Features and Apply SVM



SVM: Get Features and Apply SVM

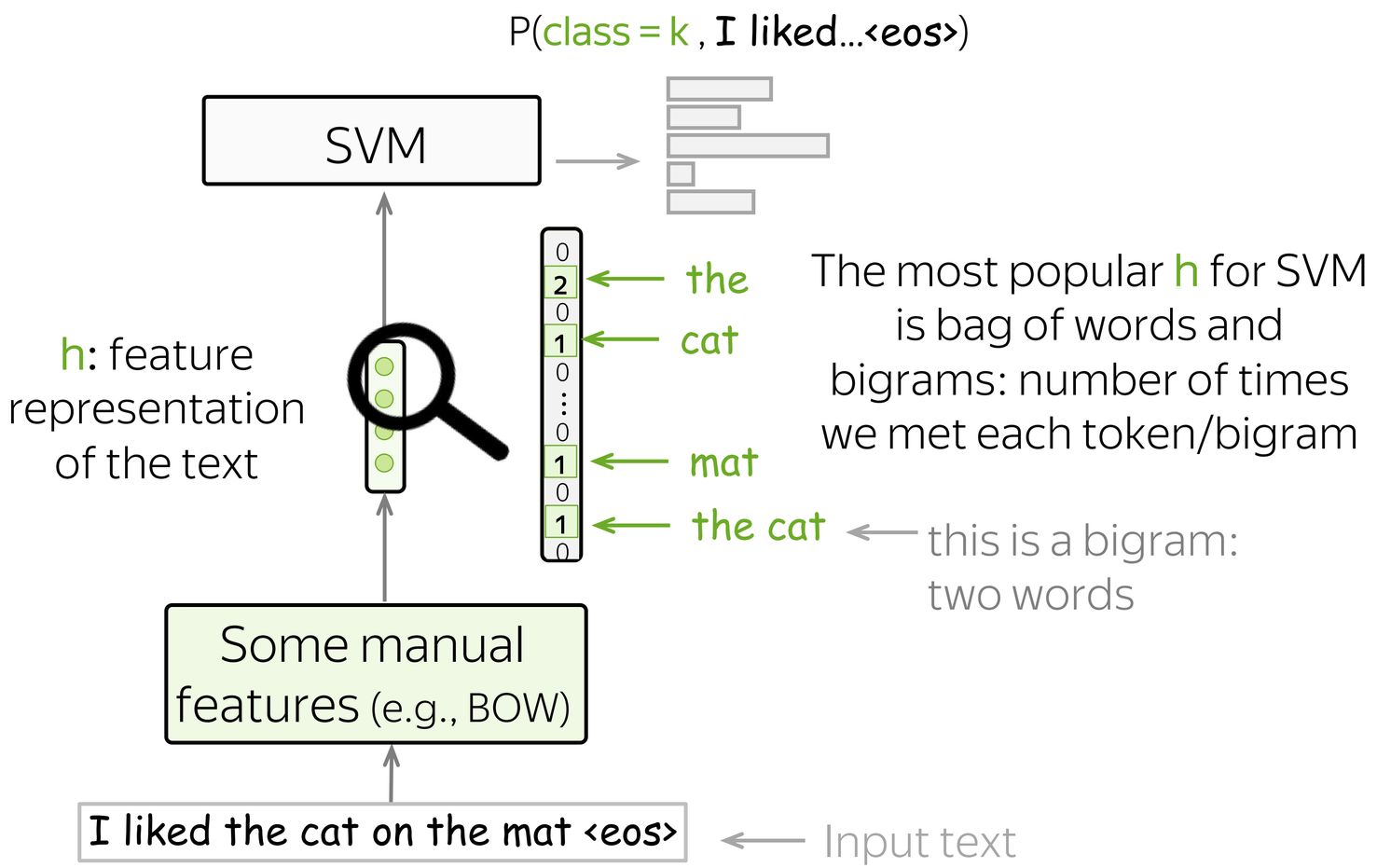
- Features: bag-of-words and ngrams



SVM: Get Features and Apply SVM

- Features: bag-of-words and ngrams
- Result: SVMs are better than Naïve Bayes

  (e.g., see the paper Question Classification using Support Vector Machines)



What is going to happen:

- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

What is going to happen:

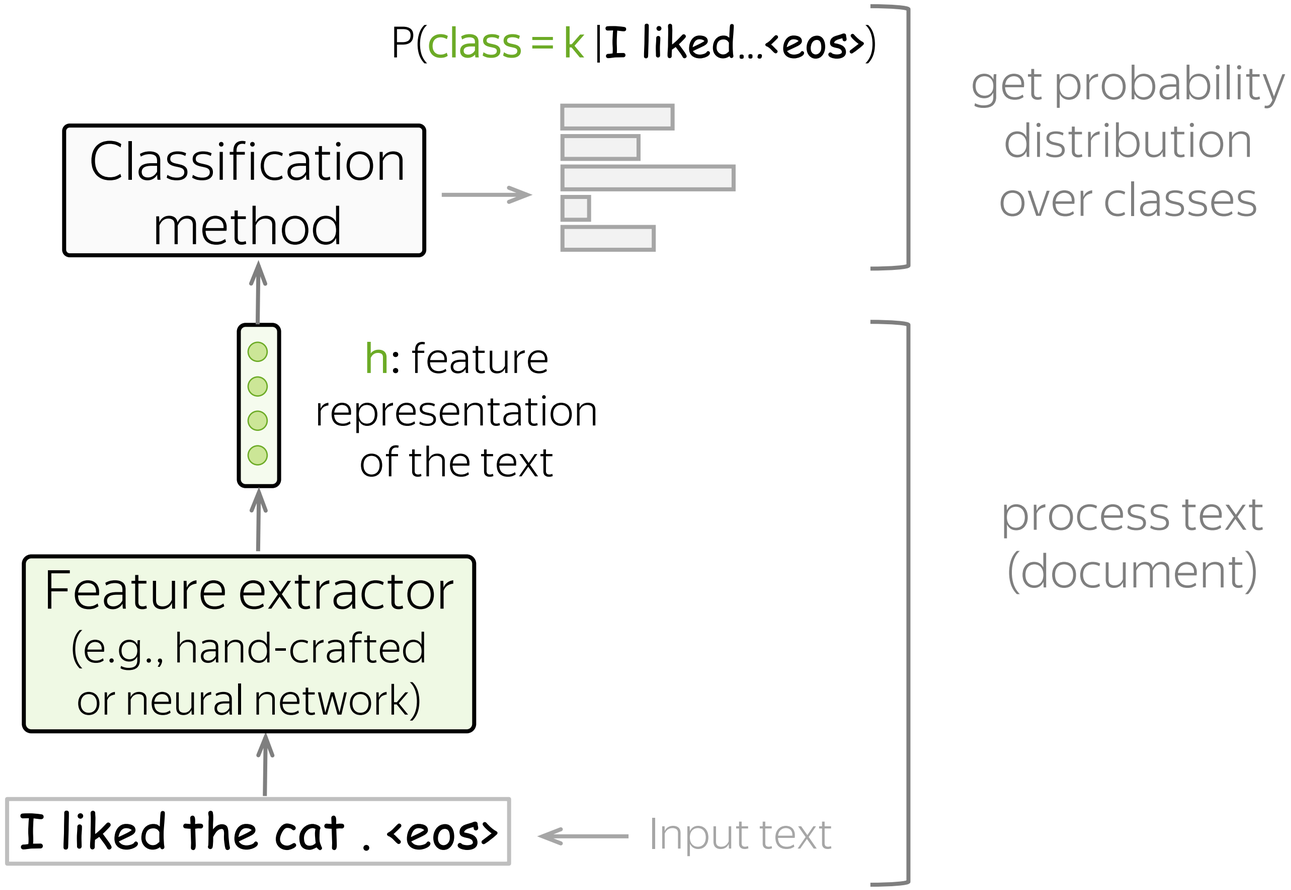
- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

What is going to happen:

- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
  - High-Level View
  - Training: Cross-Entropy
  - Models: (Weighted) BOW
  - Models: Convolutional
  - Models: Recurrent
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

Classification with Neural Networks

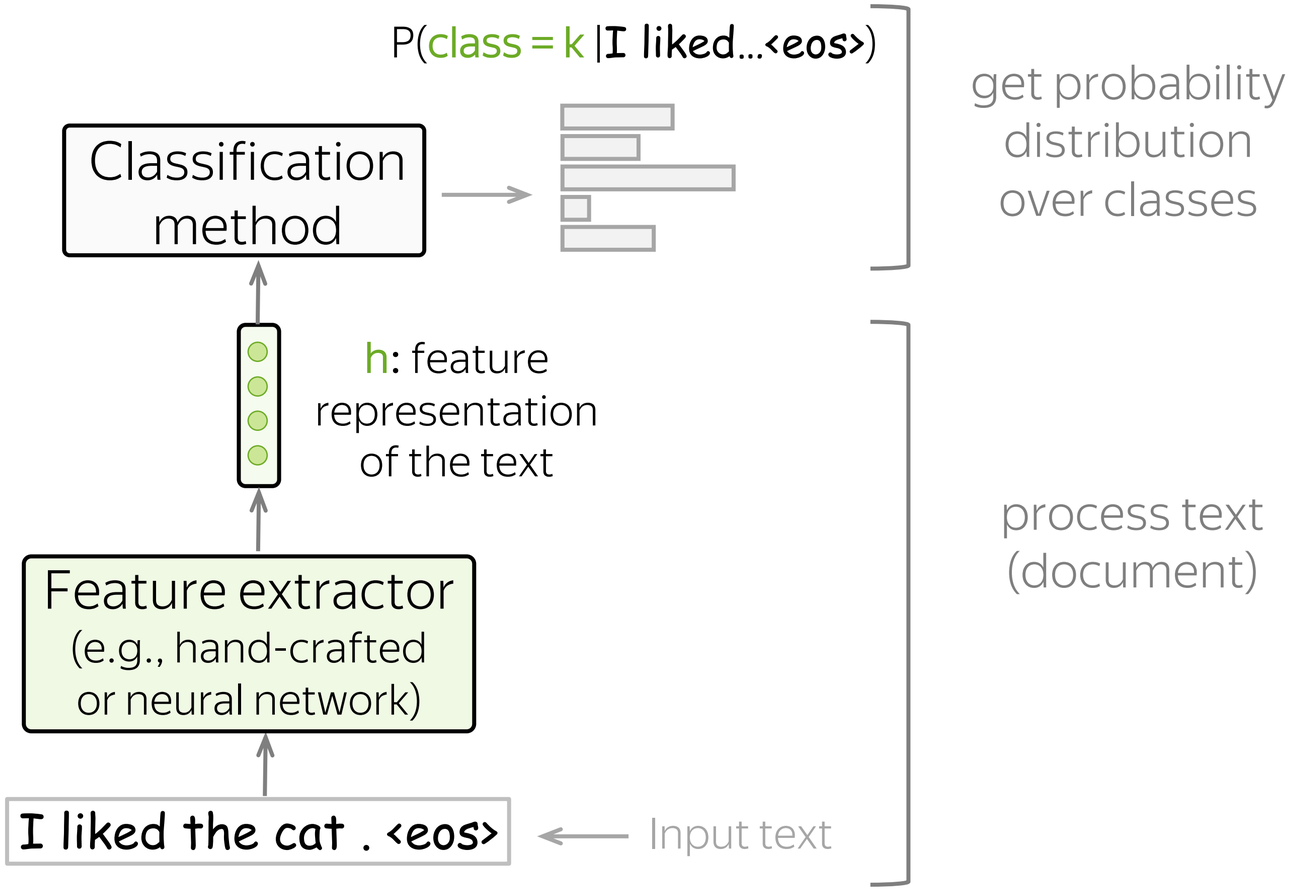
- General Classification Pipeline



Classification with Neural Networks

> Instead of manually defined features, let a neural network to learn useful features.

- General Classification Pipeline

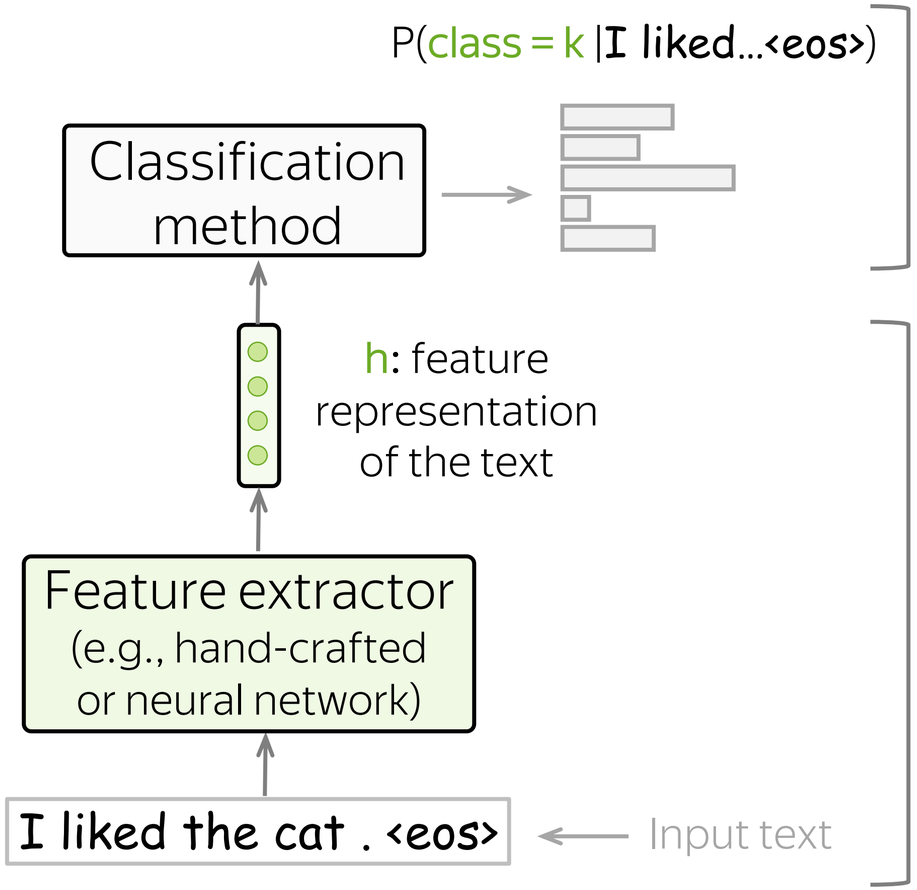


- Classification with Neural Networks

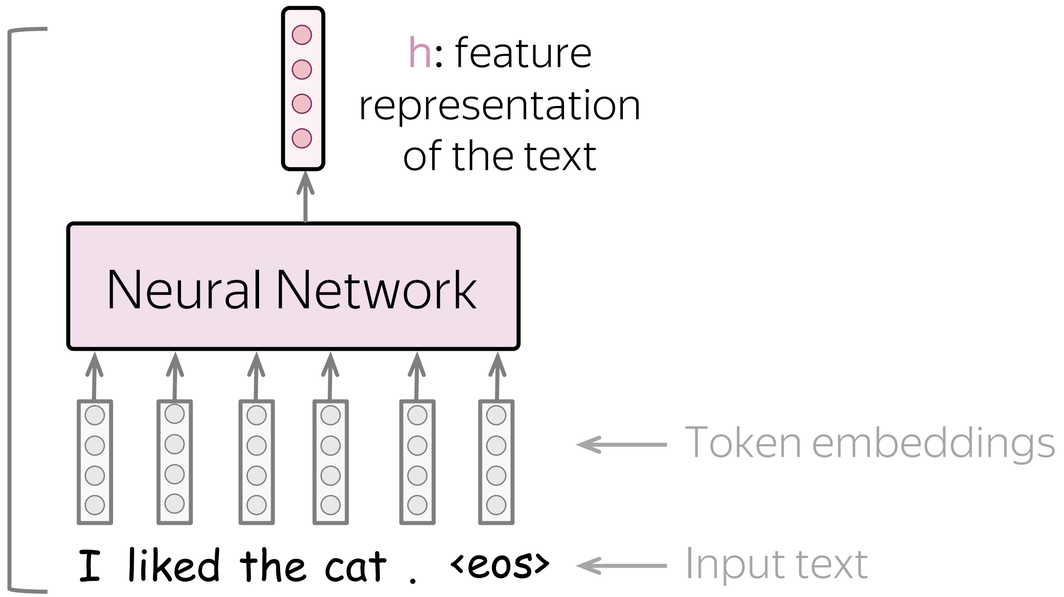
Classification with Neural Networks

> Instead of manually defined features, let a neural network to **learn useful features**.

- General Classification Pipeline



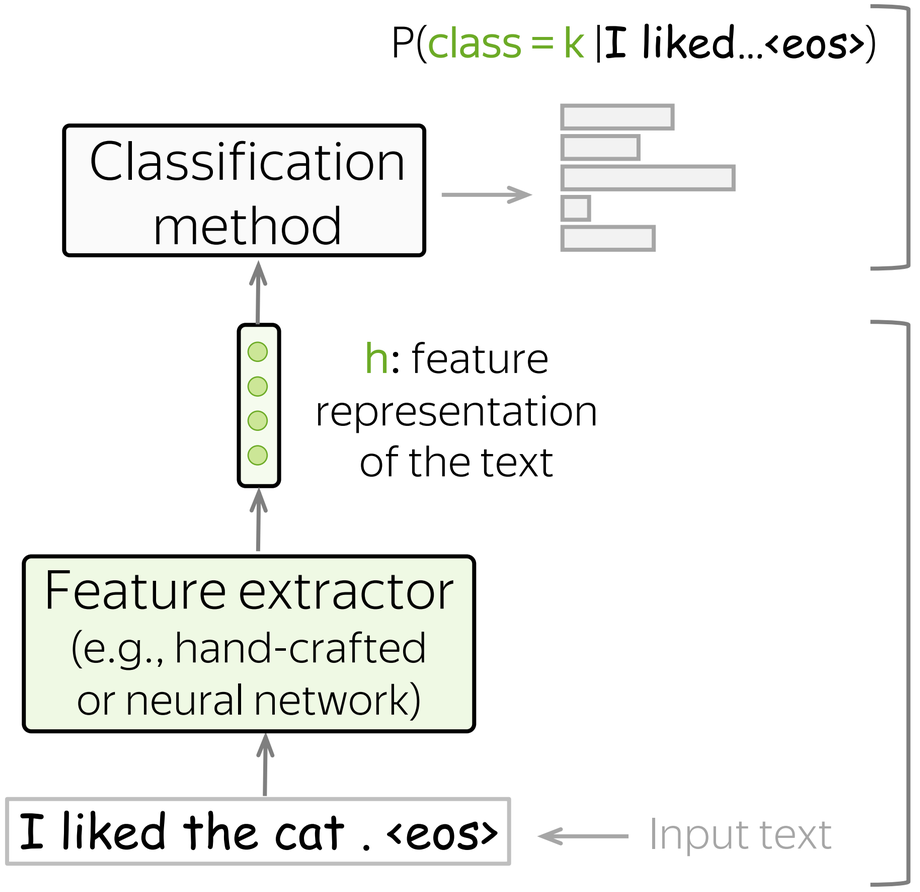
- Classification with Neural Networks



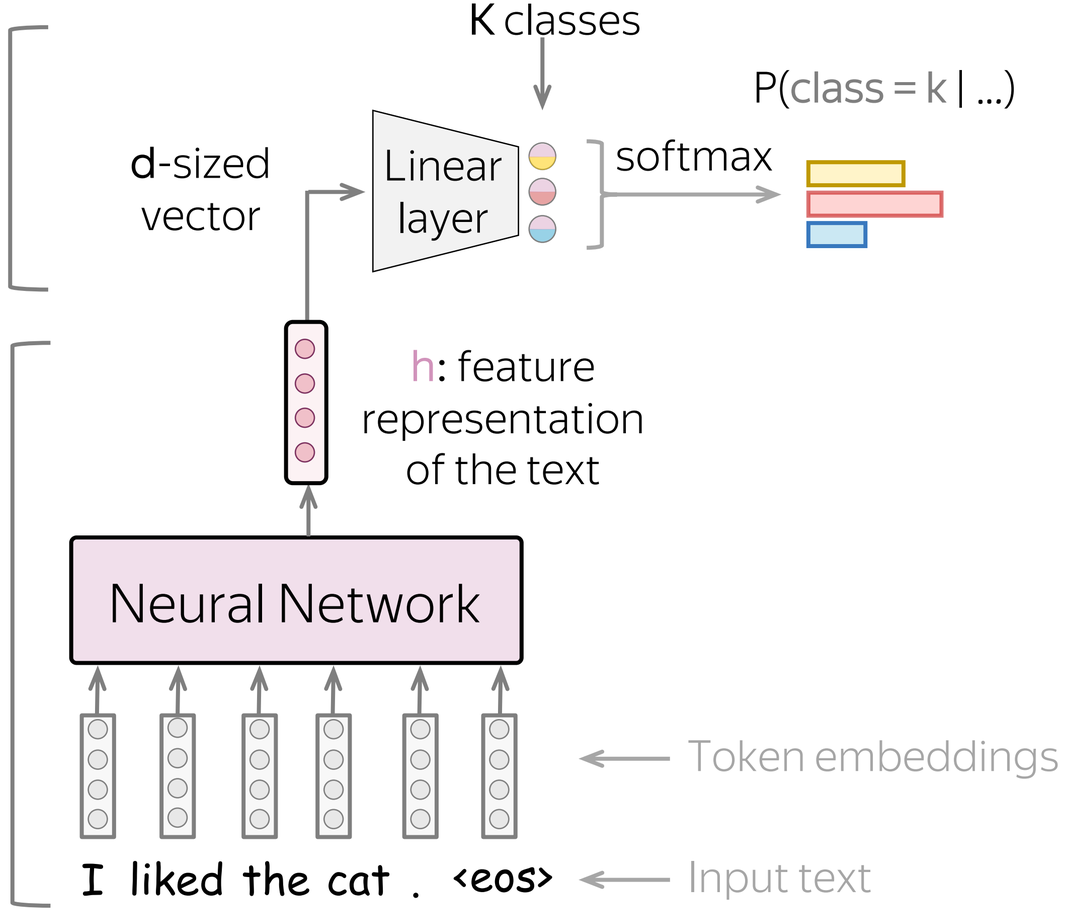
Classification with Neural Networks

> Instead of manually defined features, let a neural network to **learn useful features**.

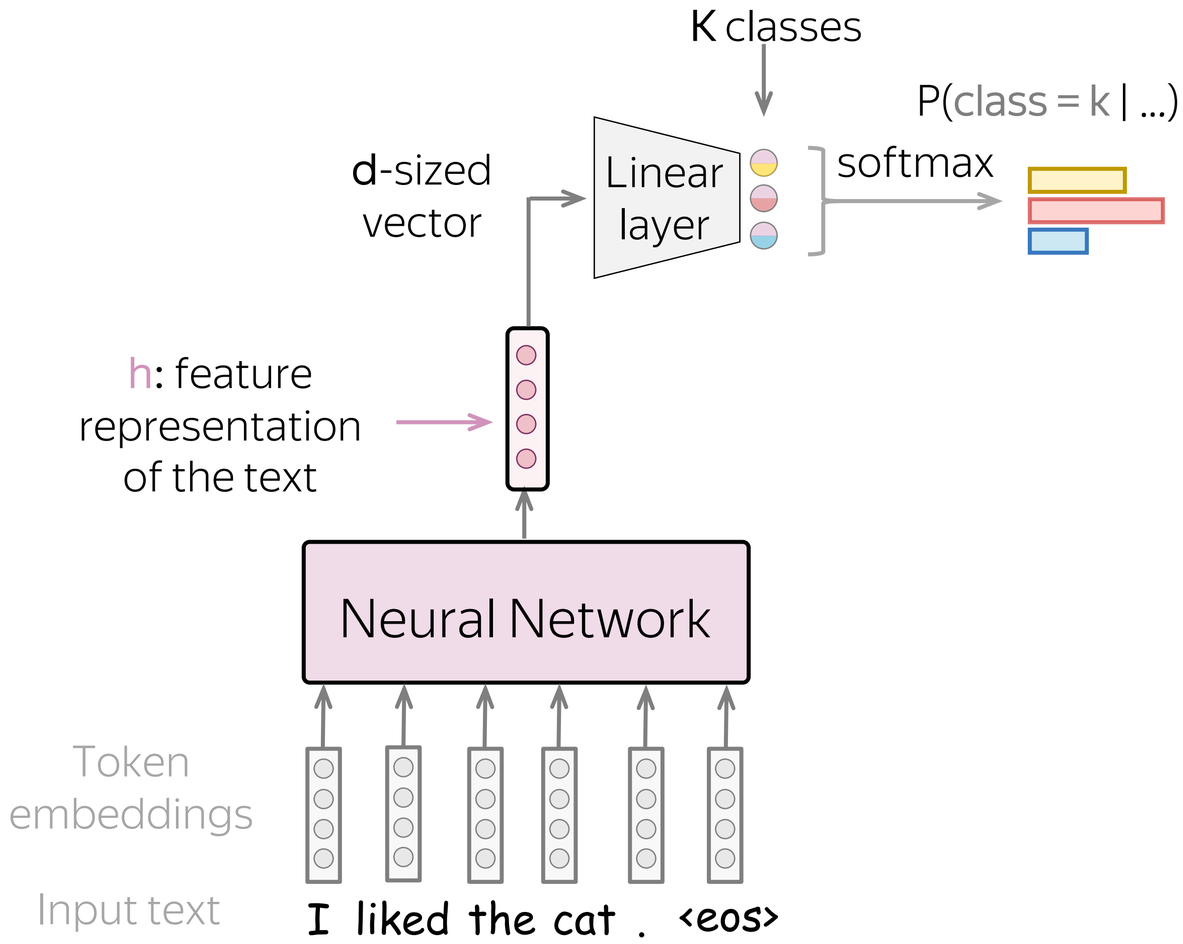
- General Classification Pipeline



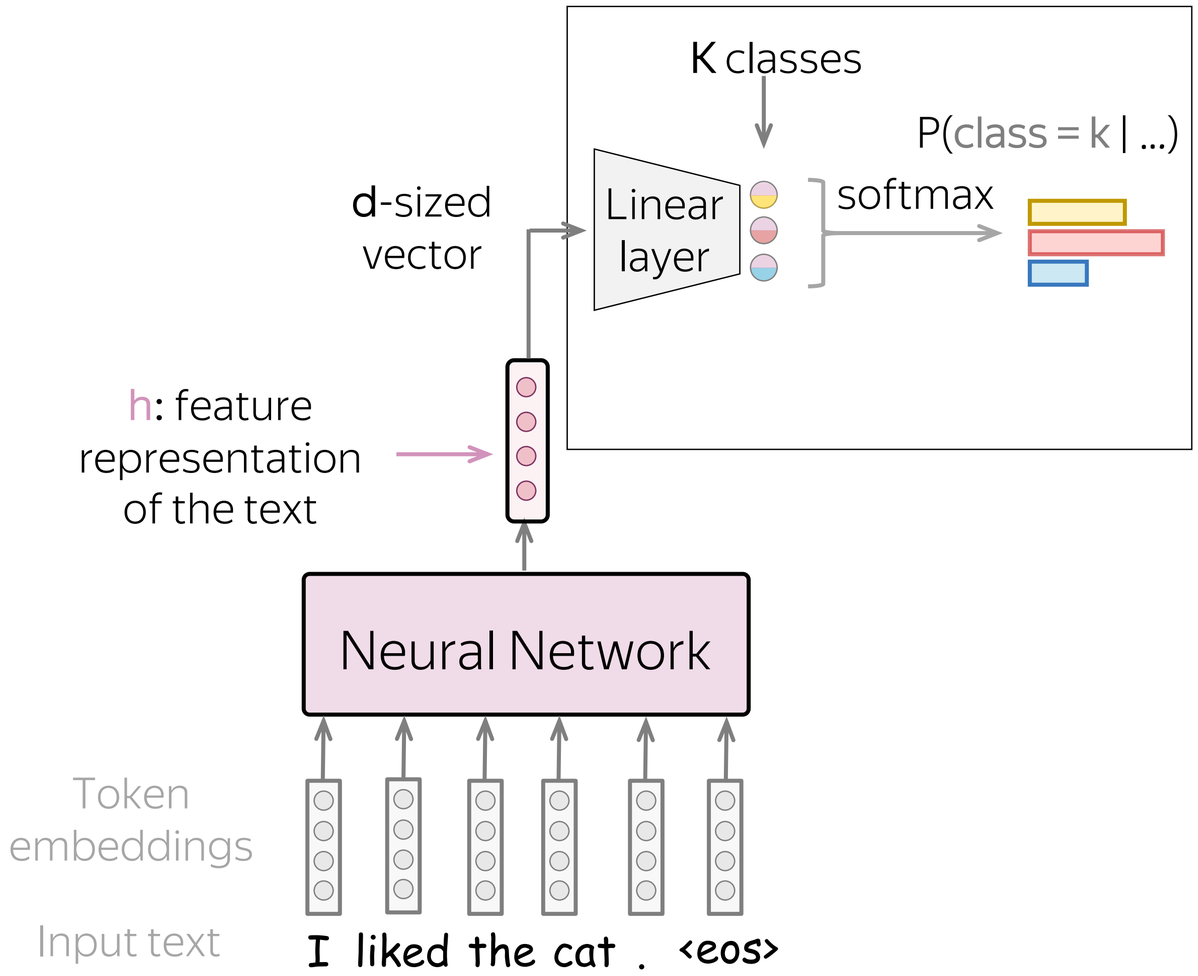
- Classification with Neural Networks



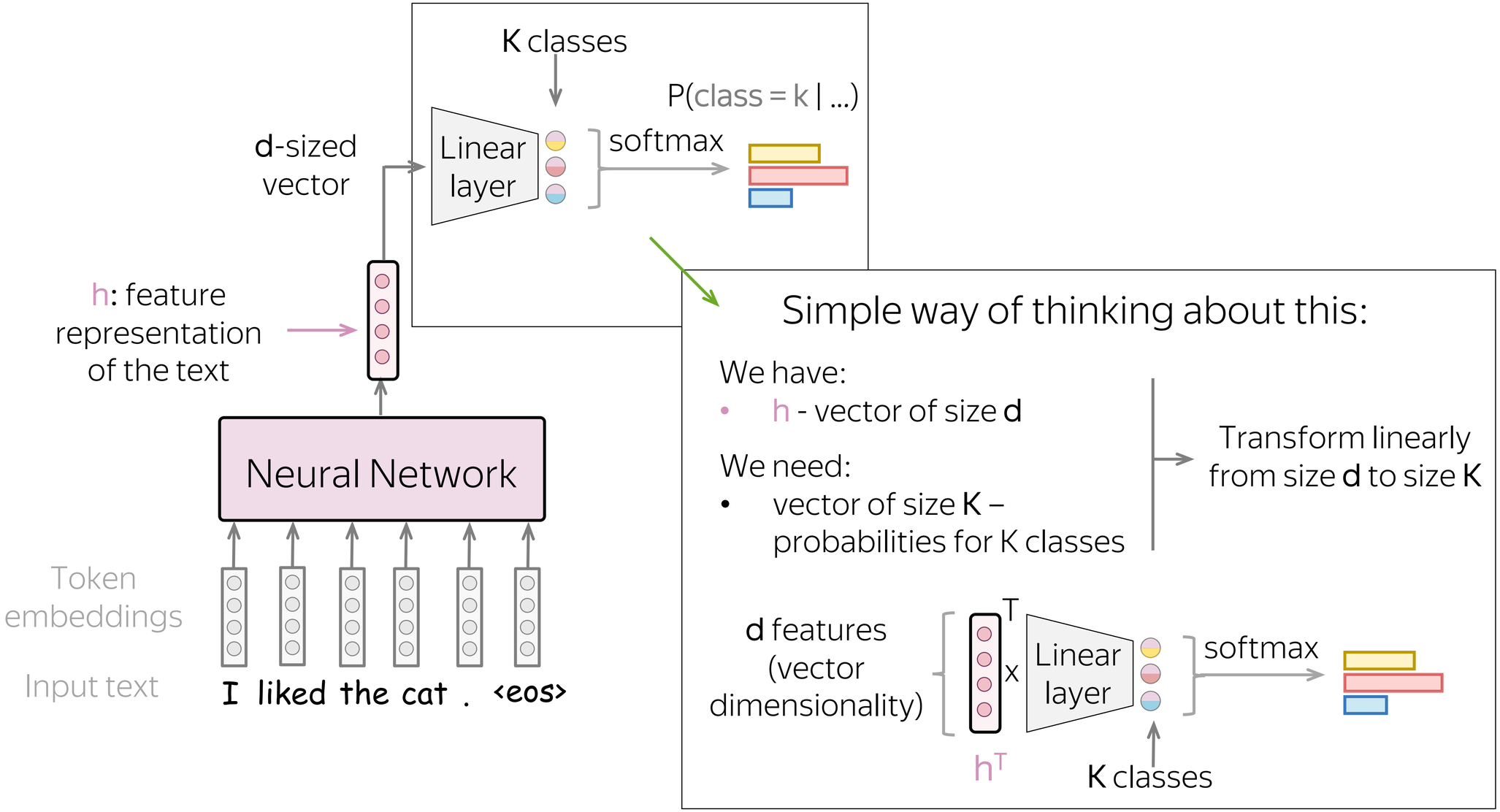
Classification with Neural Networks



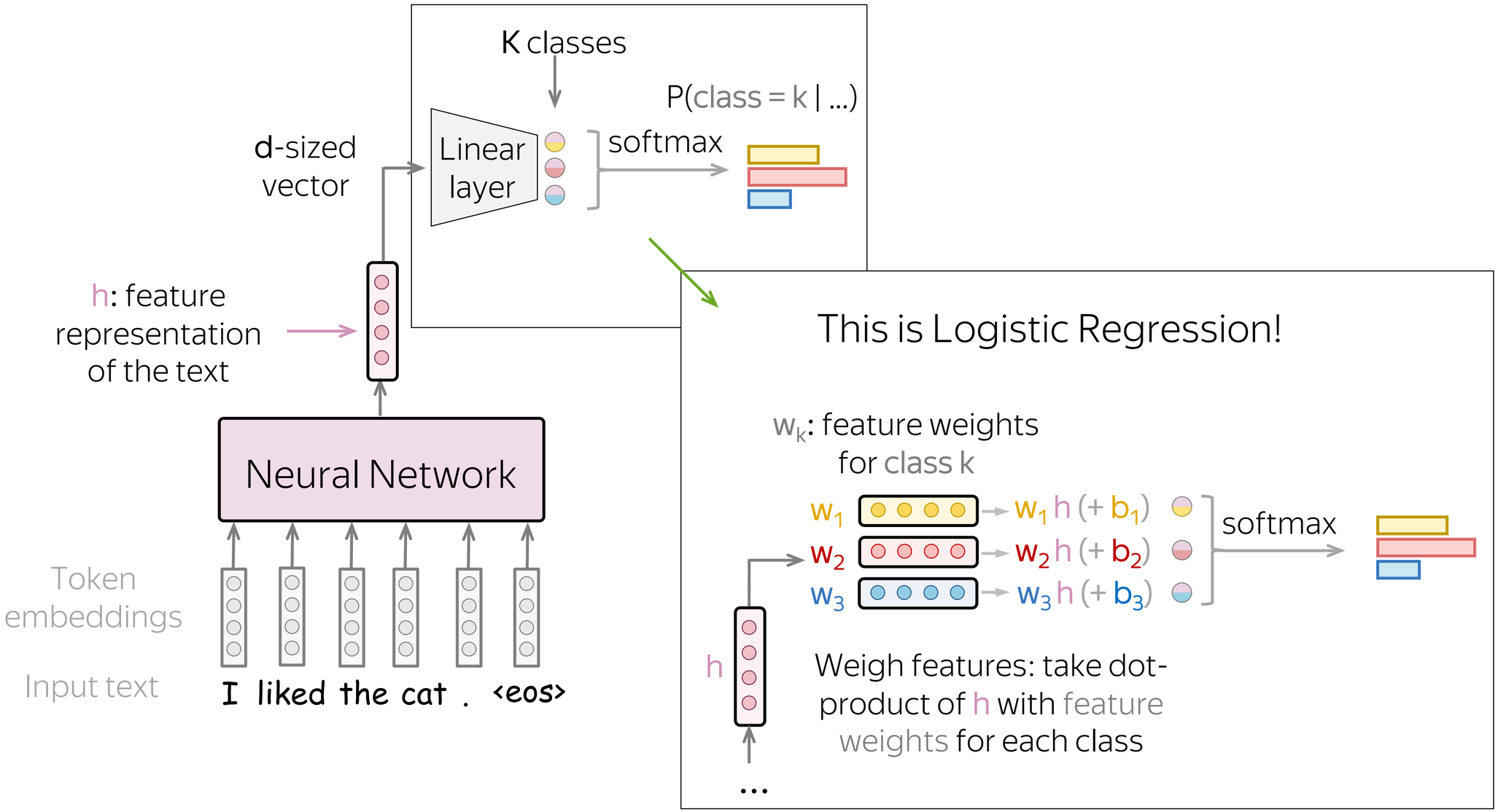
Classification with Neural Networks



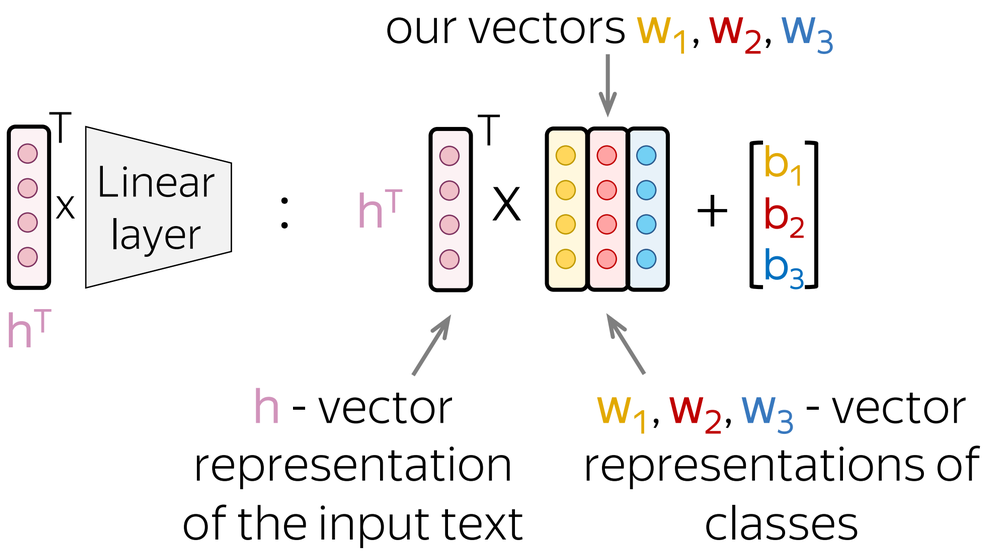
Classification with Neural Networks



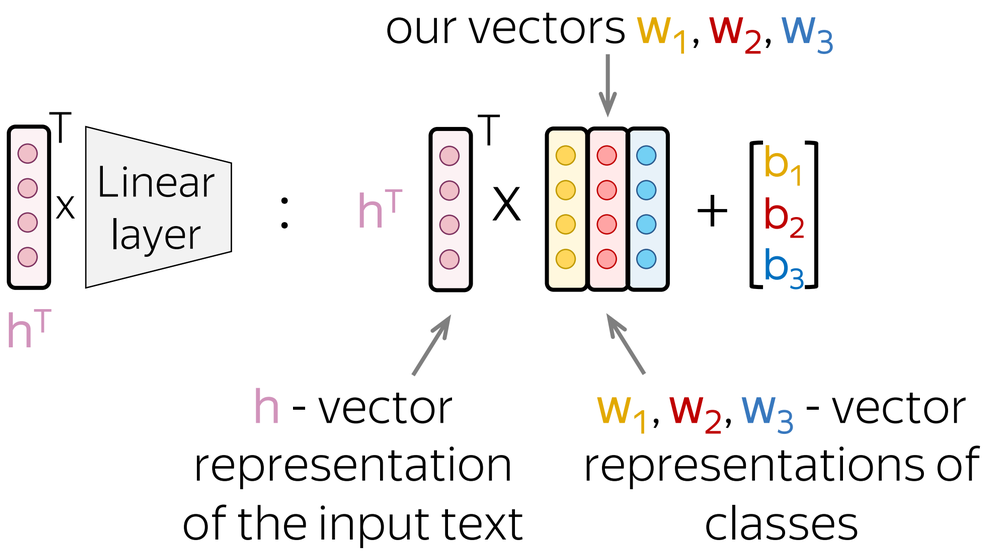
Classification with Neural Networks



Text Representation and Class Representation

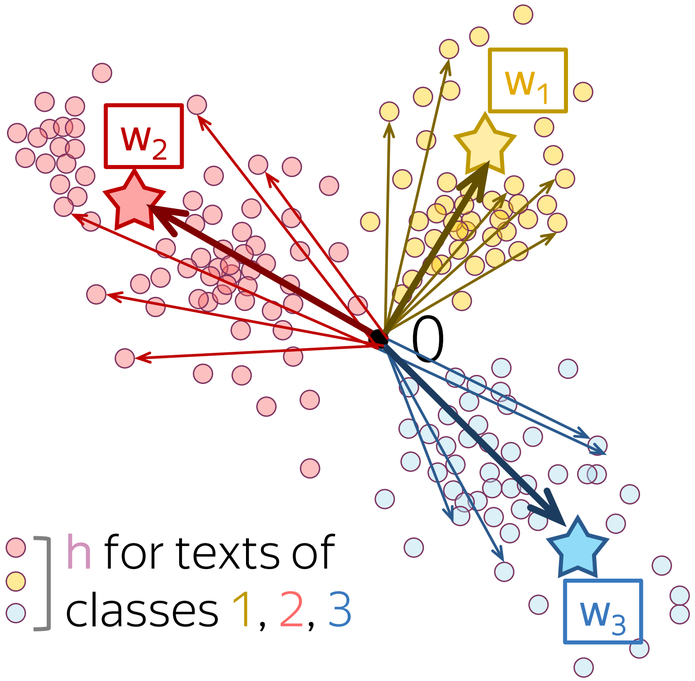


Text Representation and Class Representation



What NN learns (hopefully):

Text vectors point in the direction of the corresponding class vectors



What is going to happen:

- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
  - High-Level View
  - Training: Cross-Entropy
  - Models: (Weighted) BOW
  - Models: Convolutional
  - Models: Recurrent
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

What is going to happen:

- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
  - High-Level View
  - Training: Cross-Entropy
  - Models: (Weighted) BOW
  - Models: Convolutional
  - Models: Recurrent
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

Training: Cross-Entropy

Training example: I liked the cat on the mat <eos>
Label: k

Training: Cross-Entropy

Training example: I liked the cat on the mat <eos>

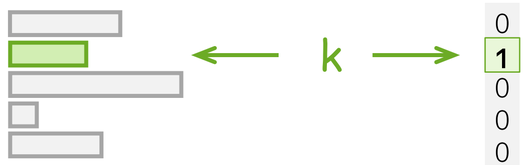
Label: k

Model prediction:

P(class = i |I liked…<eos>)

Target:

p^*



Training: Cross-Entropy

Training example: I liked the cat on the mat <eos>

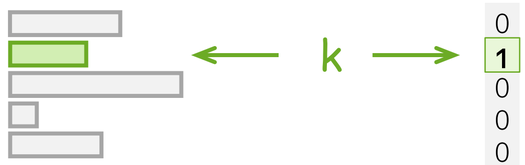
Label: k

Model prediction:

P(class = i |I liked…<eos>)

Target:

p^*



Cross-entropy loss:

$$-\sum_{i=1}^{K} p_i^* \cdot \log P(y = i|x) \to min \quad (p_k^* = 1, p_i^* = 0, i \neq k)$$

For one-hot targets, this is equivalent to

$$-\log P(y = k|x) \to min$$

Training: Cross-Entropy

Training example: I liked the cat on the mat <eos>

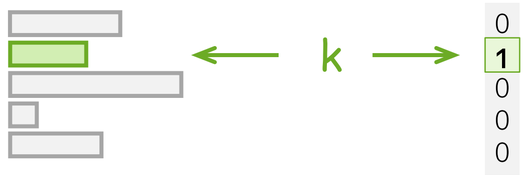
Label: k

Model prediction:

P(class = i |I liked…<eos>)

Target:

p^*

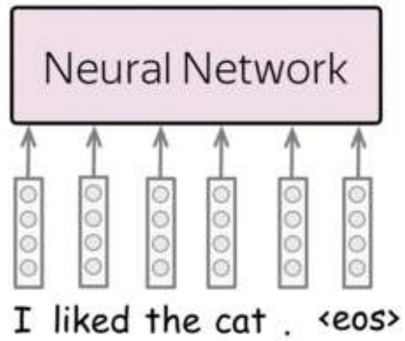


Cross-entropy loss:

$$-\sum_{i=1}^{K} p_i^* \cdot \log P(y=i|x) \to min \quad (p_k^* = 1, p_i^* = 0, i \neq k)$$

For one-hot targets, this is equivalent to

$$-\log P(y=k|x) \to min$$



Feed a text to the network

Correct label: 4 ← we want the model to predict this

What is going to happen:

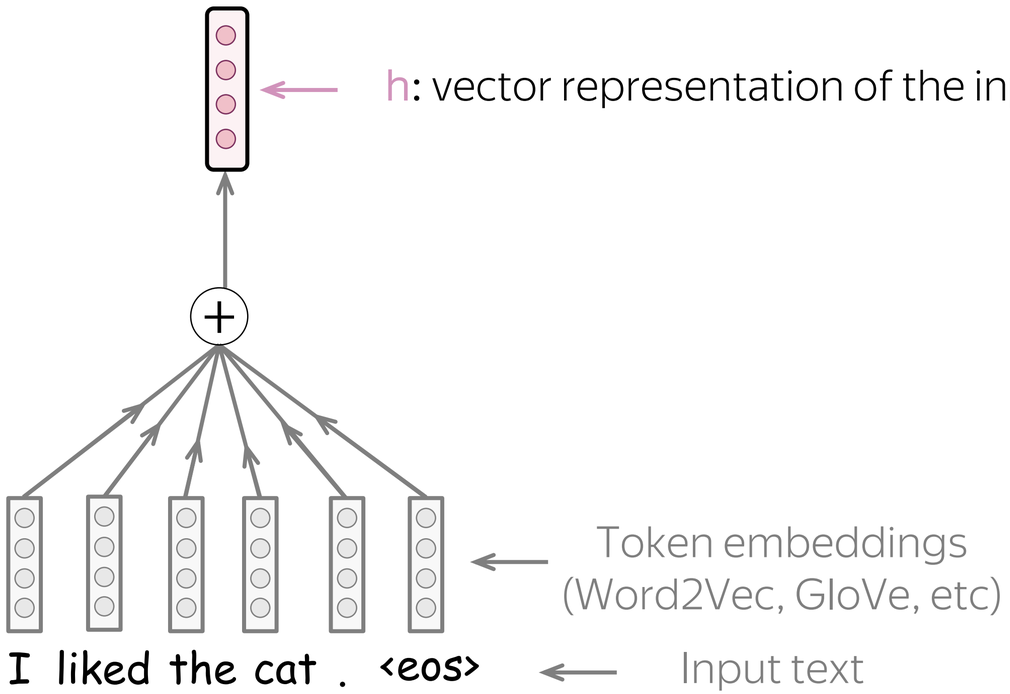
- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
  - High-Level View
  - Training: Cross-Entropy
  - Models: (Weighted) BOW
  - Models: Convolutional
  - Models: Recurrent
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

What is going to happen:

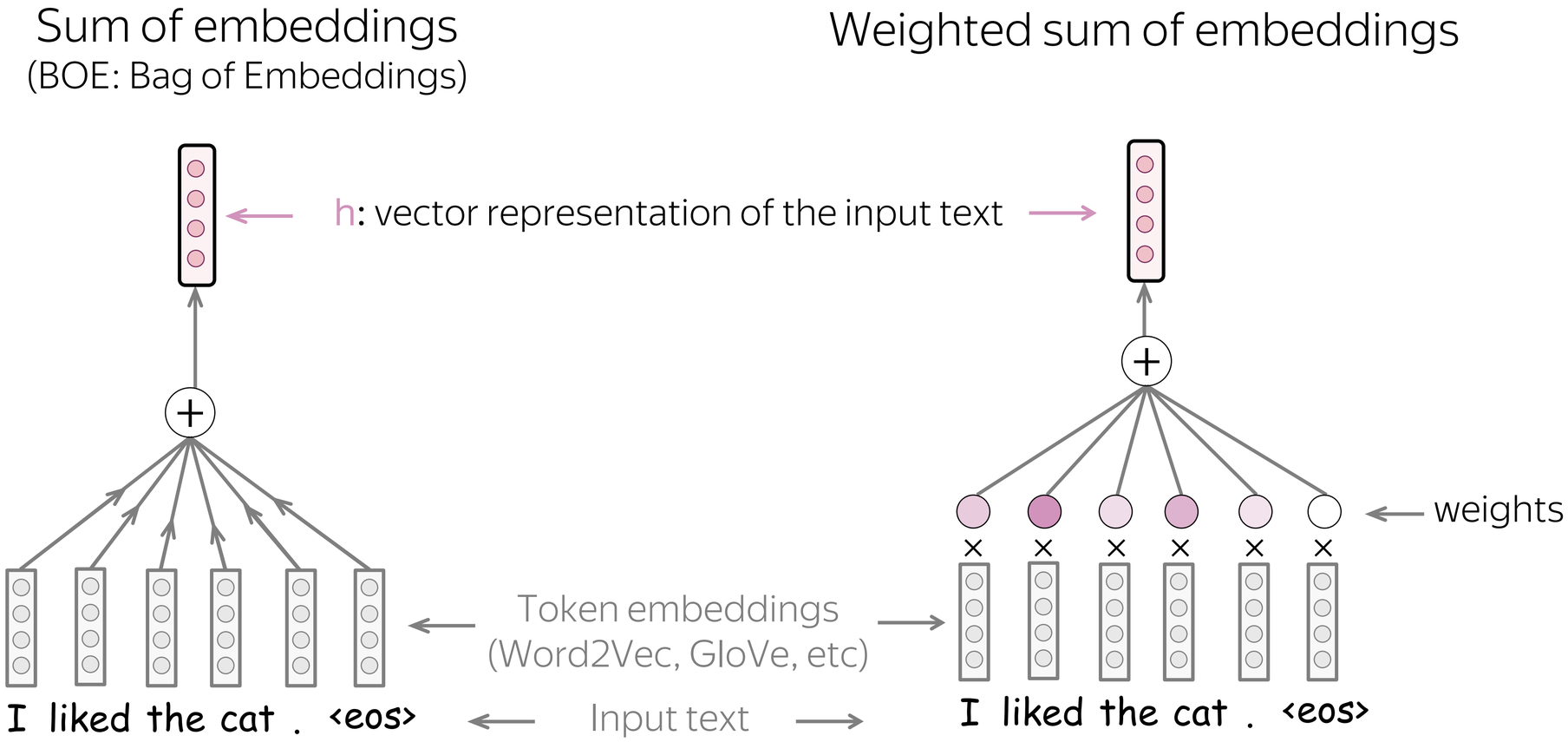
- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
  - High-Level View
  - Training: Cross-Entropy
  - Models: (Weighted) BOW
  - Models: Convolutional
  - Models: Recurrent
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

The Simplest Models: BOE and Weighted BOE

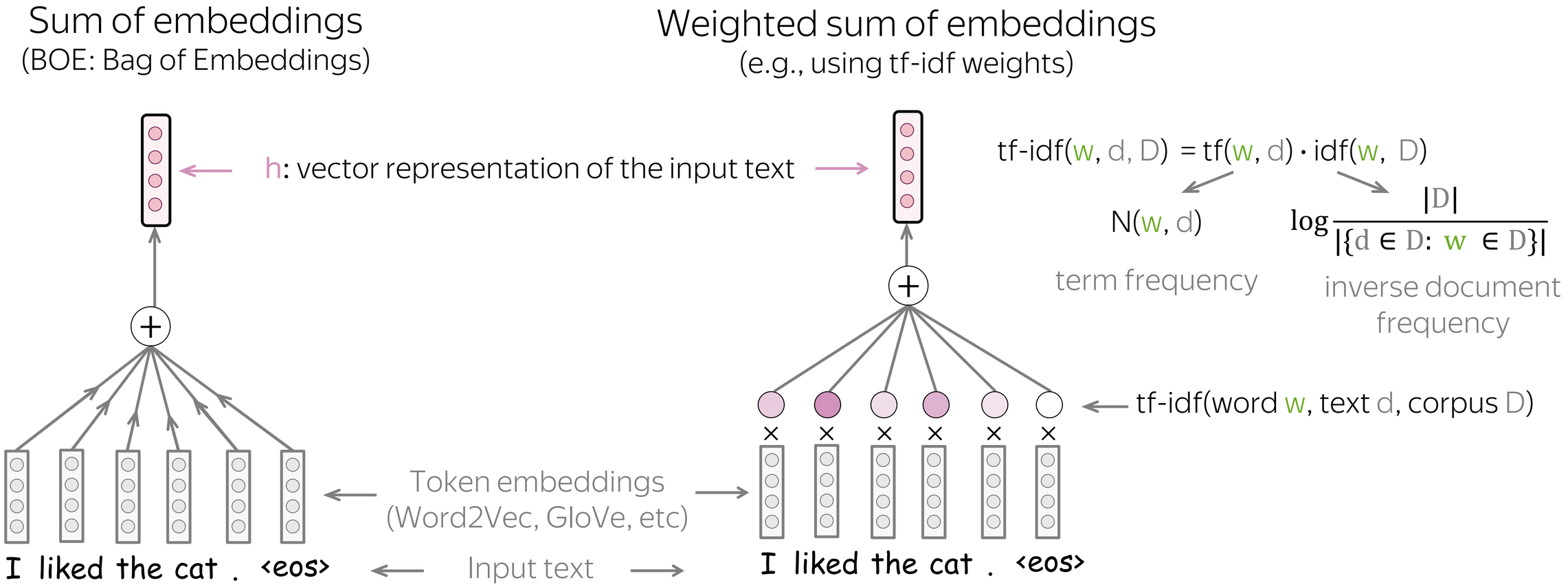
Sum of embeddings
(BOE: Bag of Embeddings)



The Simplest Models: BOE and Weighted BOE



The Simplest Models: BOE and Weighted BOE



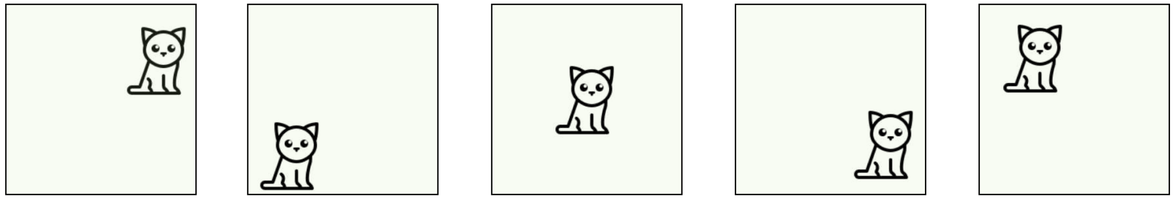
What is going to happen:

- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
  - High-Level View
  - Training: Cross-Entropy
  - Models: (Weighted) BOW
  - Models: Convolutional
  - Models: Recurrent
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

What is going to happen:

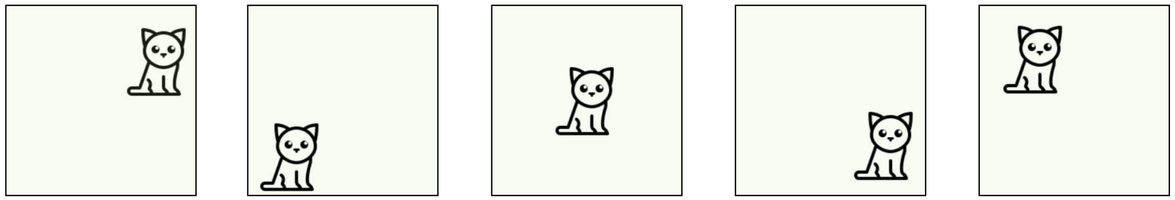
- Examples of classification tasks
- General View: Features + Classifier
- Models: Generative vs Discriminative
- Classical Methods
- Neural Methods
  - High-Level View
  - Training: Cross-Entropy
  - Models: (Weighted) BOW
  - Models: Convolutional
  - Models: Recurrent
- Multi-Label Classification
- Practical Tips
-  Analysis and Interpretability

Convolutions for Images and Translation Invariance



Label: cat Label: cat Label: cat Label: cat Label: cat

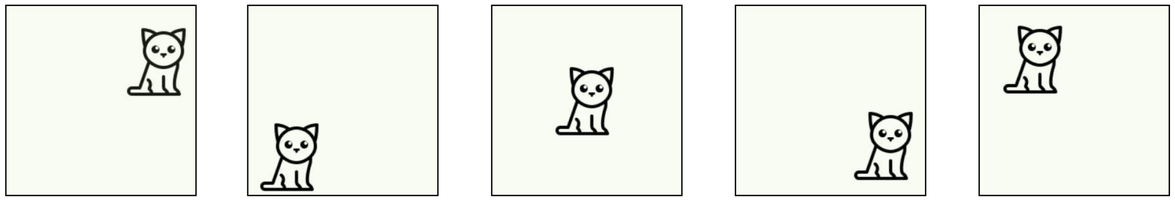
Convolutions for Images and Translation Invariance



Label: **cat** Label: **cat** Label: **cat** Label: **cat** Label: **cat**

We don’t care where the cat is, we
care that it is somewhere.

Convolutions for Images and Translation Invariance

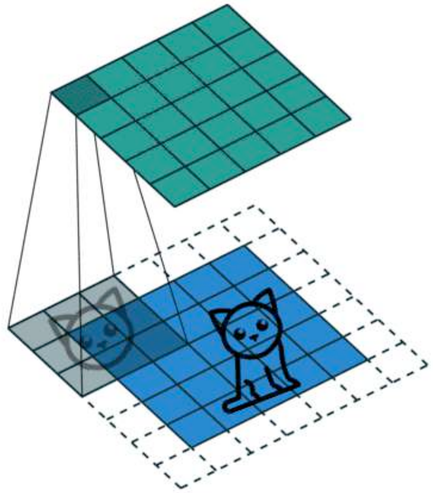


Label: **cat** Label: **cat** Label: **cat** Label: **cat** Label: **cat**

We don’t care where the cat is, we care that it is somewhere.

Then why don’t we process all these cats similarly?

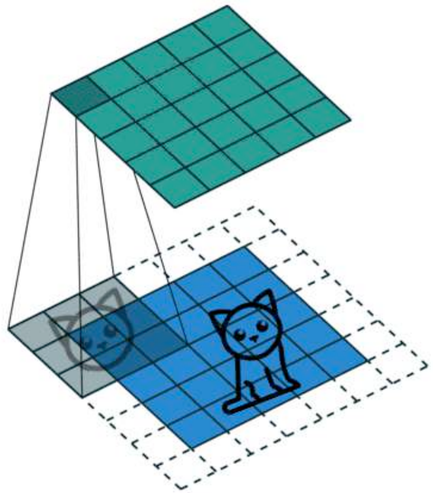
Convolutions for Images and Translation Invariance



The gif is adapted from the one taken from the repo https://github.com/vdumoulin/conv_arithmetic

- apply the same operation to small parts of an input
- find “matches” with patterns

Convolutions for Images and Translation Invariance

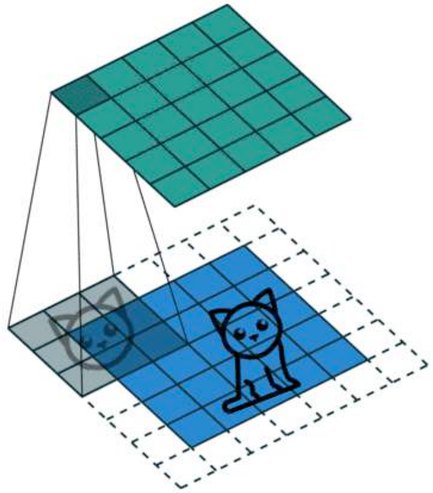


The gif is adapted from the one taken from the repo https://github.com/vdumoulin/conv_arithmetic

- apply the same operation to small parts of an input
- find “matches” with patterns

this is how CNNs extract features

Convolutions for Images and Translation Invariance



The gif is adapted from the one taken from the repo https://github.com/vdumoulin/conv_arithmetic

- apply the same operation to small parts of an input
- find “matches” with patterns

this is how CNNs extract features

- a network learns which patterns are useful
- from bottom to top of a network, patterns evolve from simple to complicated

We’ll see this in the analysis section

What About Texts?

An absolutely great movie! I watched the premiere with my friends.

The movie about cats was absolutely great, and the cats were cute.

The movie is about cats running around, and it is absolutely great.

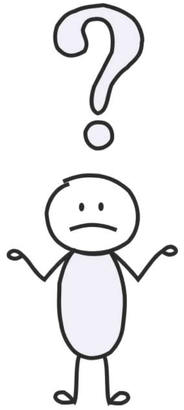
What About Texts?

An absolutely great movie! I watched the premiere with my friends.

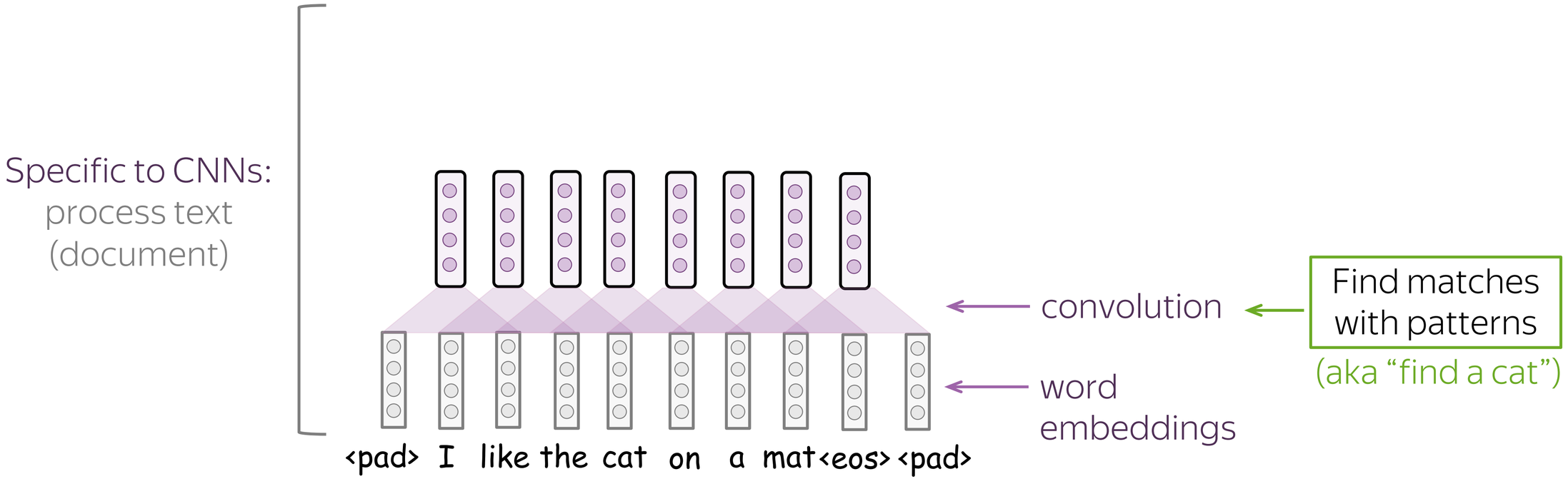
The movie about cats was absolutely great, and the cats were cute.

The movie is about cats running around, and it is absolutely great.

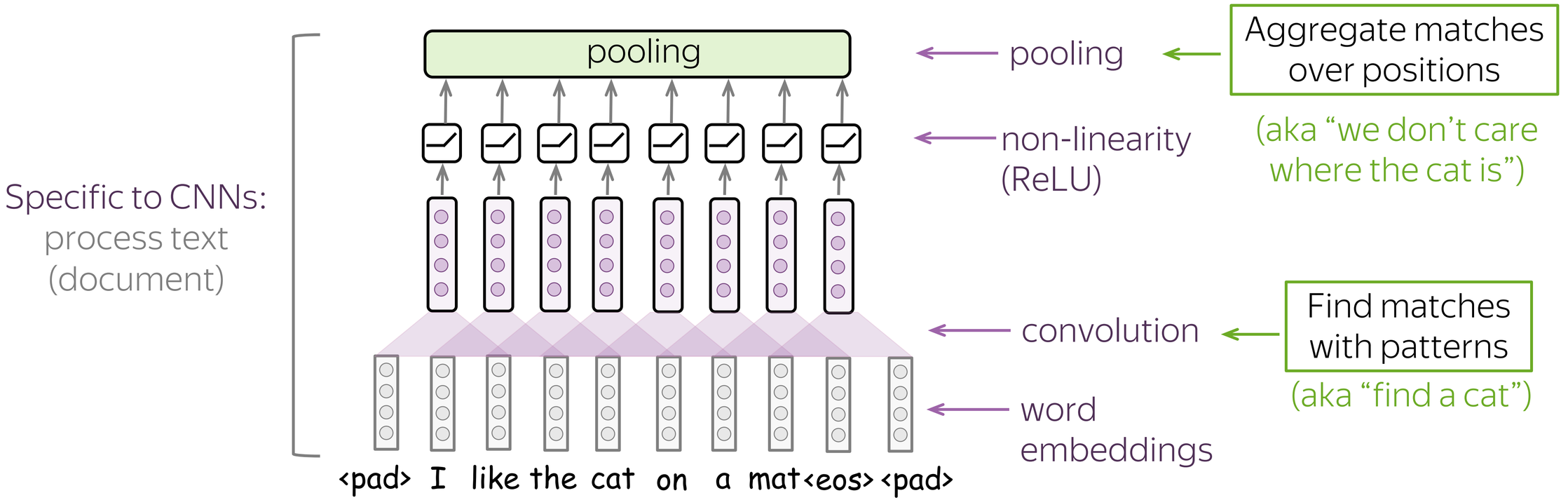
If a clue is very informative, maybe we don't care much where in a text it appears?



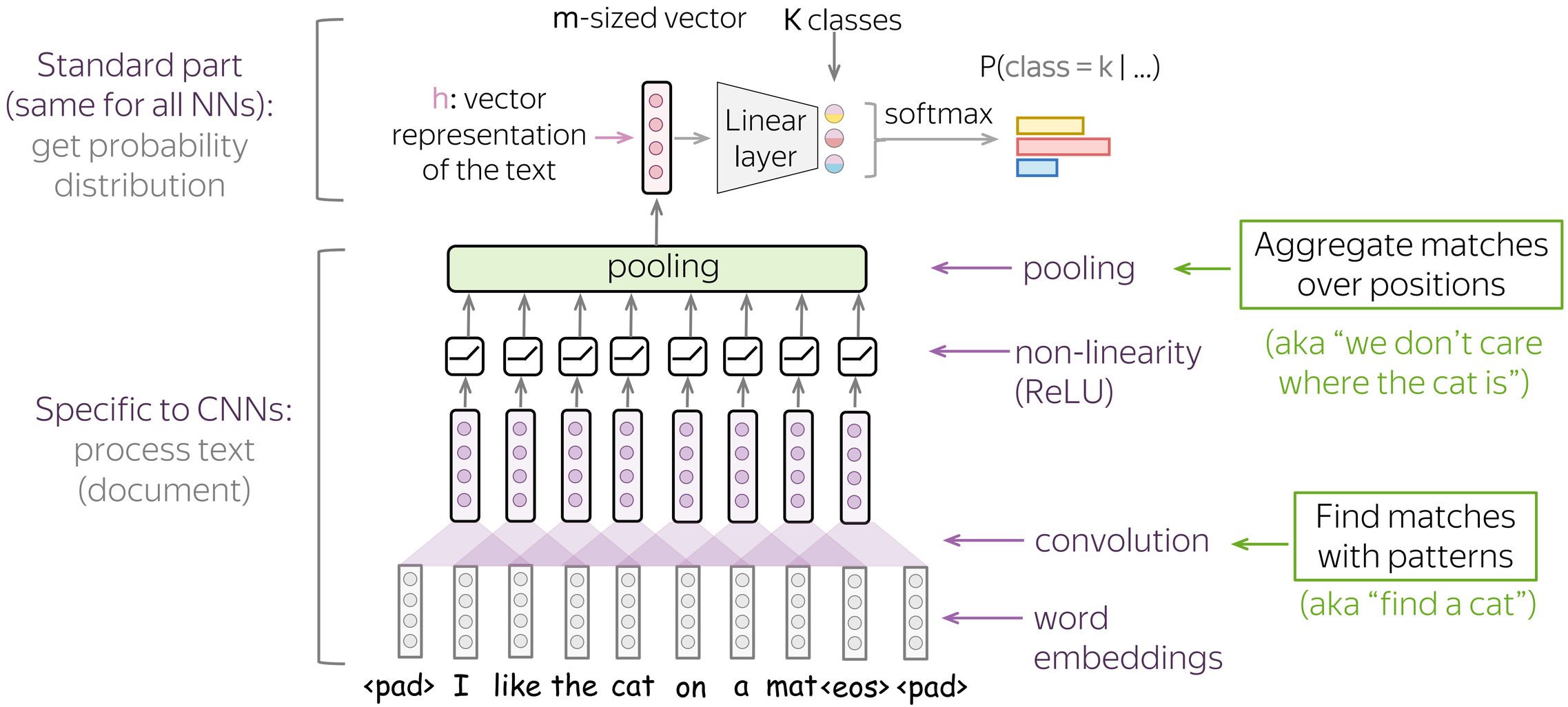
A Typical Model: Convolution + Pooling



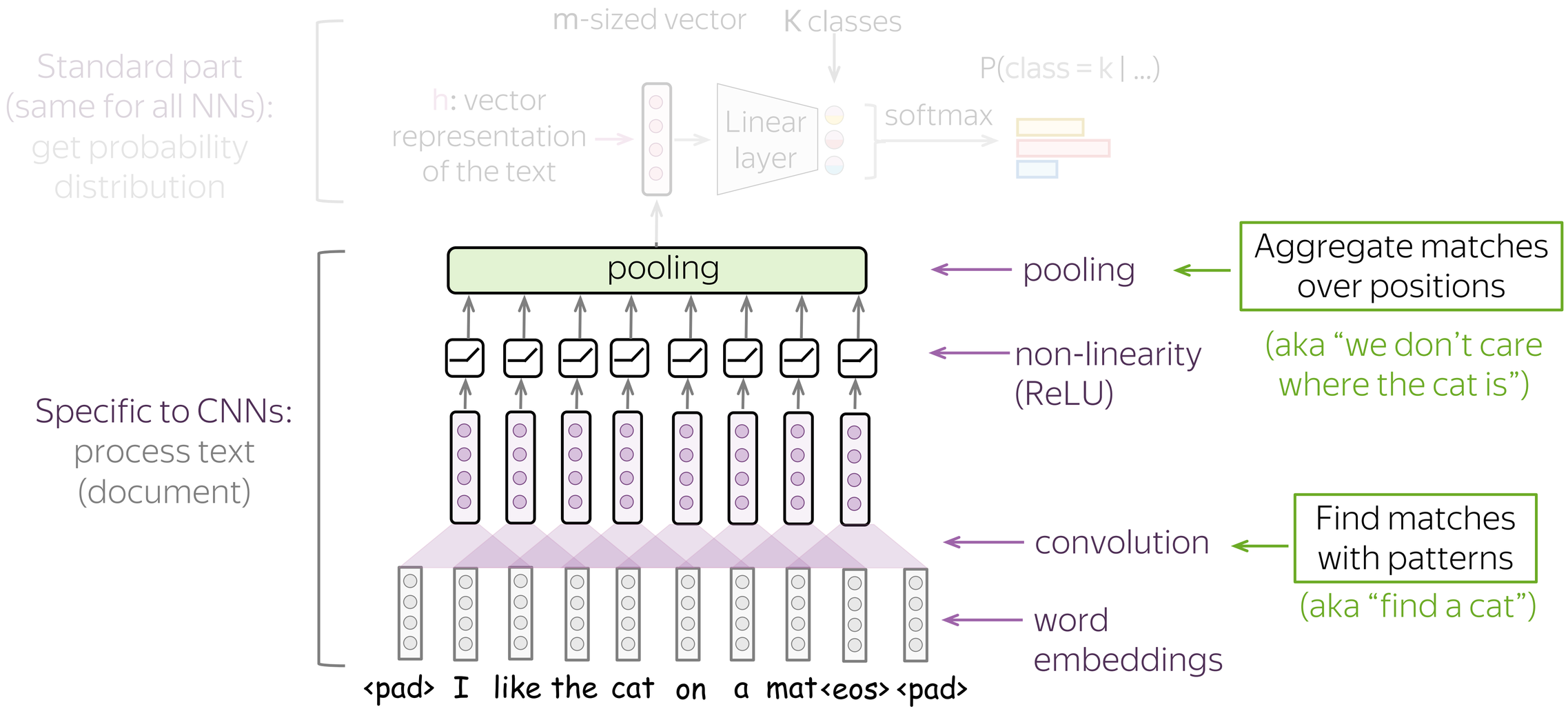
A Typical Model: Convolution + Pooling



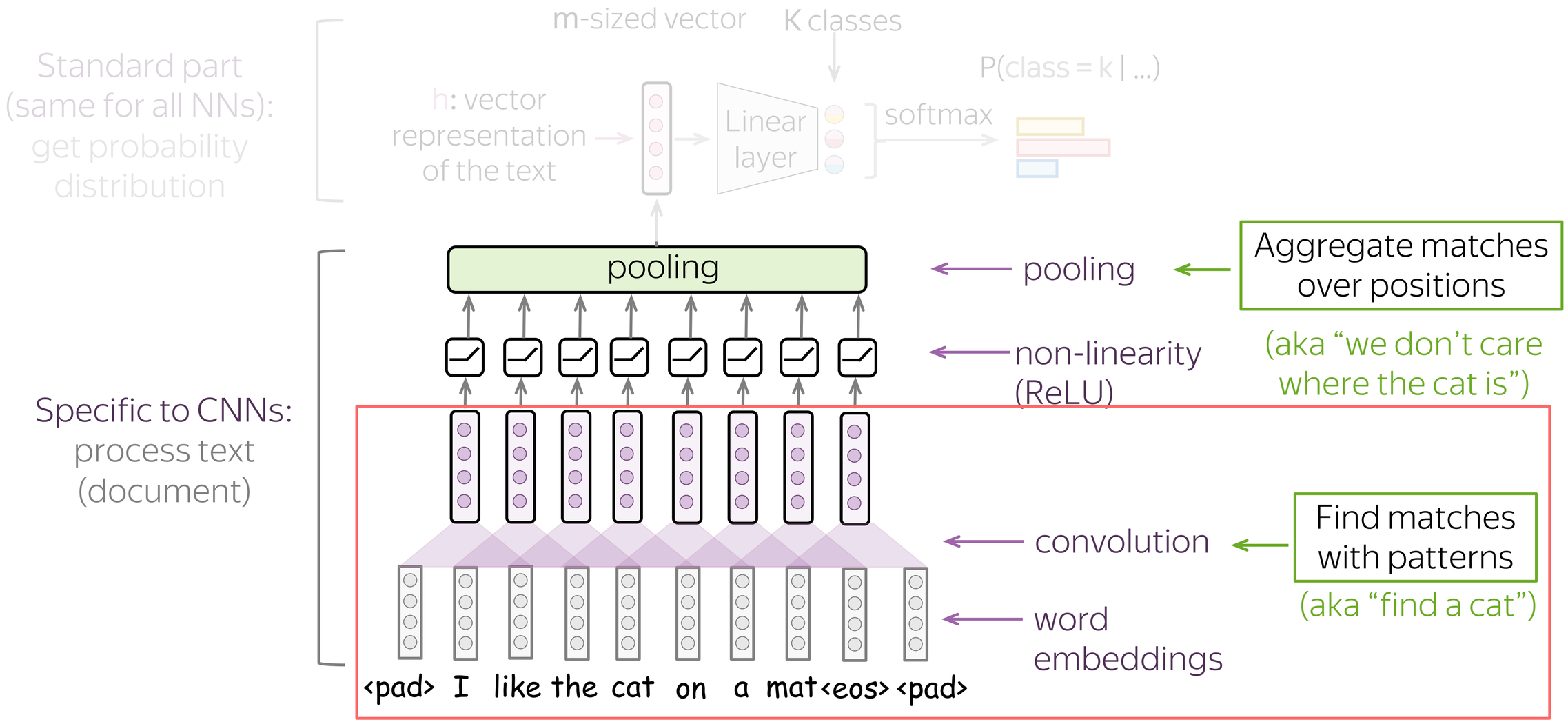
A Typical Model: Convolution + Pooling



A Typical Model: Convolution + Pooling

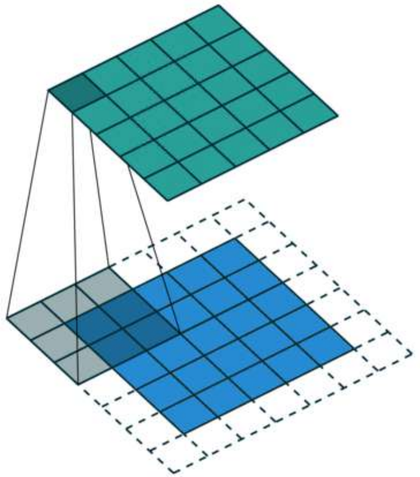


A Typical Model: Convolution + Pooling



Building Blocks: Convolution

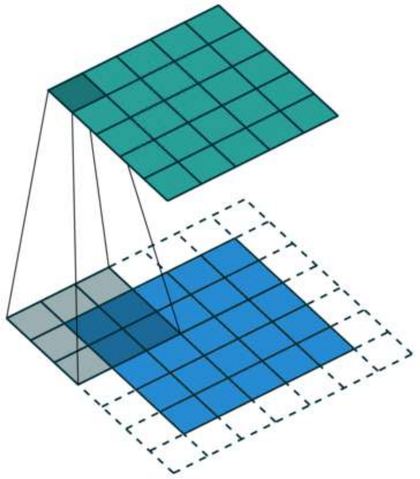
Convolution filter for an image



This gif is from the repo
https://github.com/vdumoulin/conv_arithmetic

Building Blocks: Convolution

Convolution filter for an image

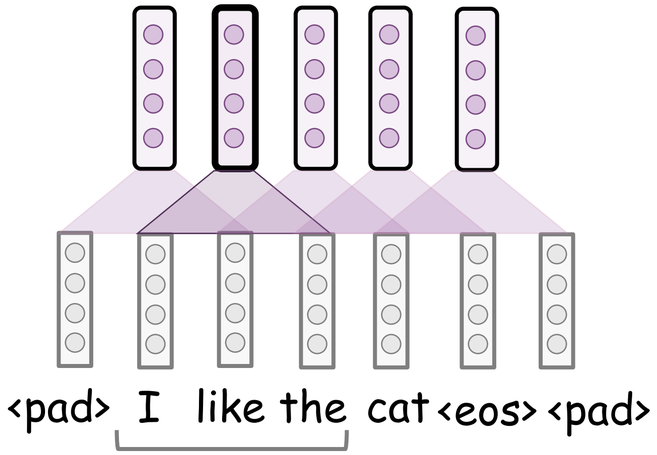


This gif is from the repo
https://github.com/vdumoulin/conv_arithmetic

Convolution filter for a text

Convolution is a Linear Operation Applied to Each Window

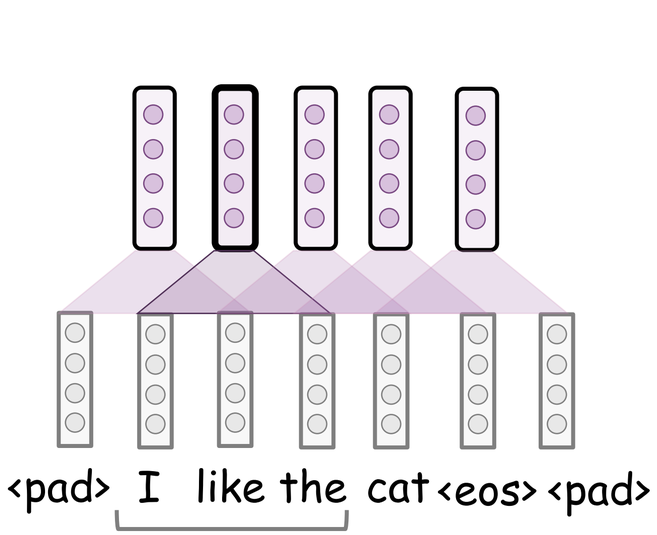
(except for a non-linearity)

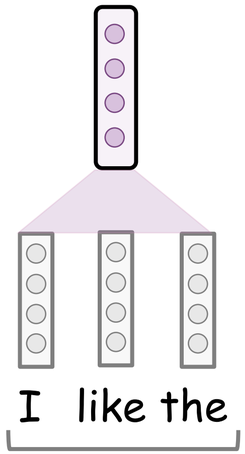


- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)

Convolution is a Linear Operation Applied to Each Window

(except for a non-linearity)

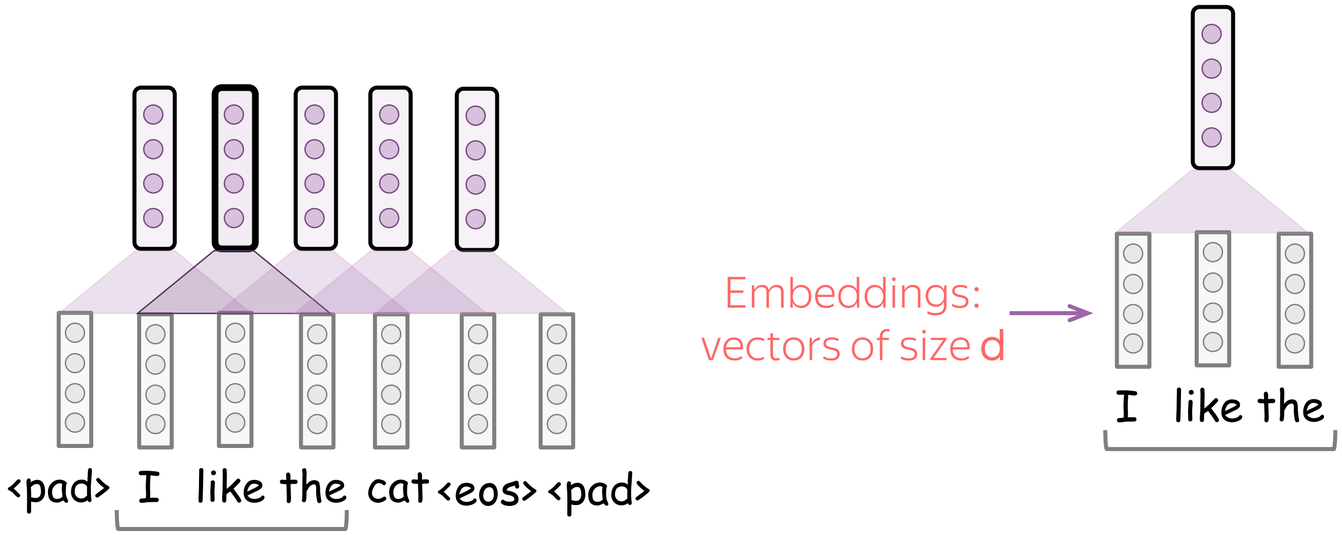




- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)

Convolution is a Linear Operation Applied to Each Window

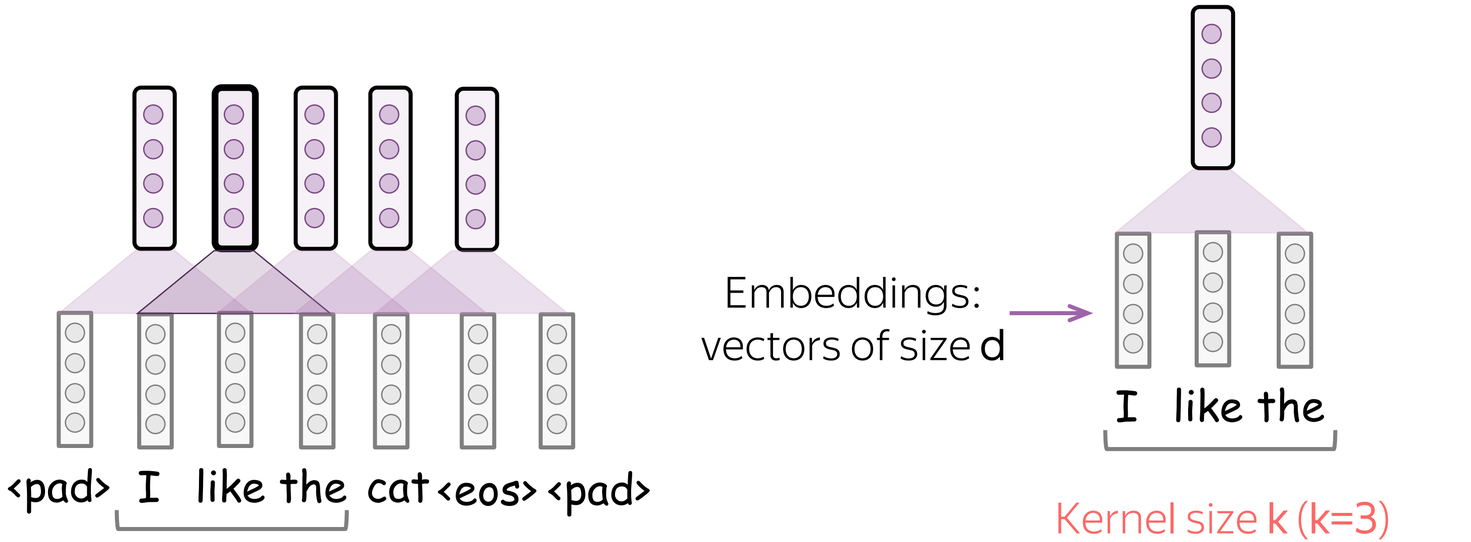
(except for a non-linearity)



- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)
- d (input channels) – input vector size

Convolution is a Linear Operation Applied to Each Window

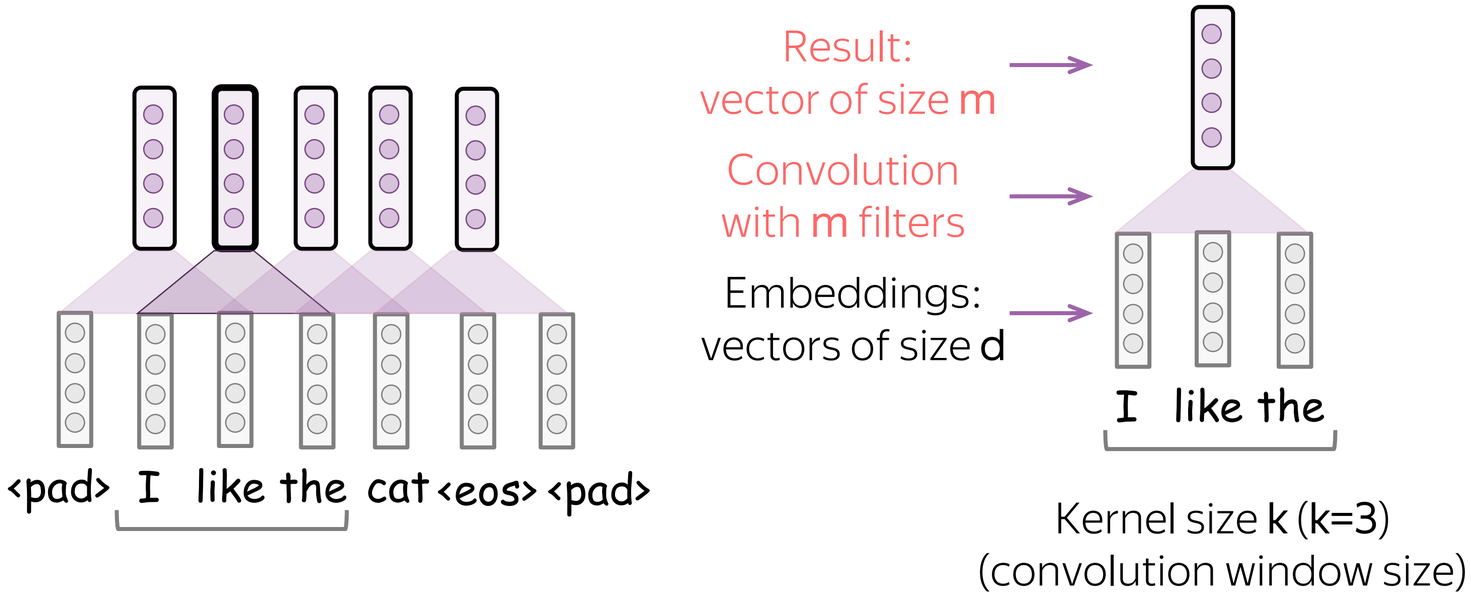
(except for a non-linearity)



- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)
- d (input channels) – input vector size
- k (kernel size) – conv. window length

Convolution is a Linear Operation Applied to Each Window

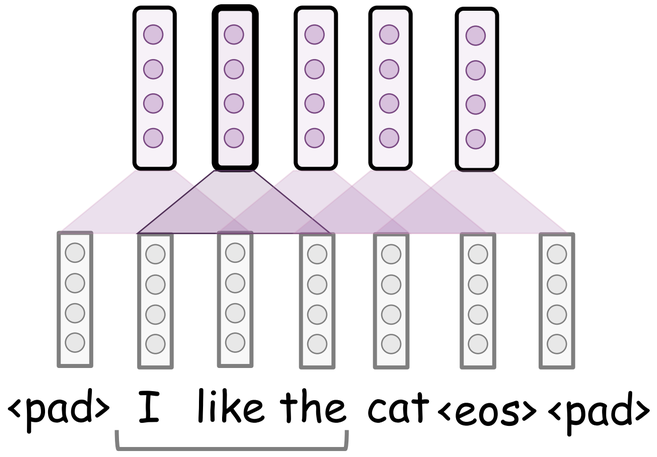
(except for a non-linearity)

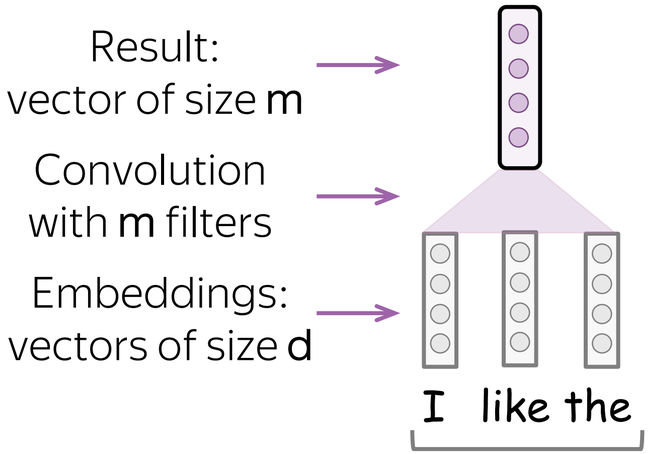


- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)
- d (input channels) – input vector size
- k (kernel size) – conv. window length
- m (output channels) – number of filters

Convolution is a Linear Operation Applied to Each Window

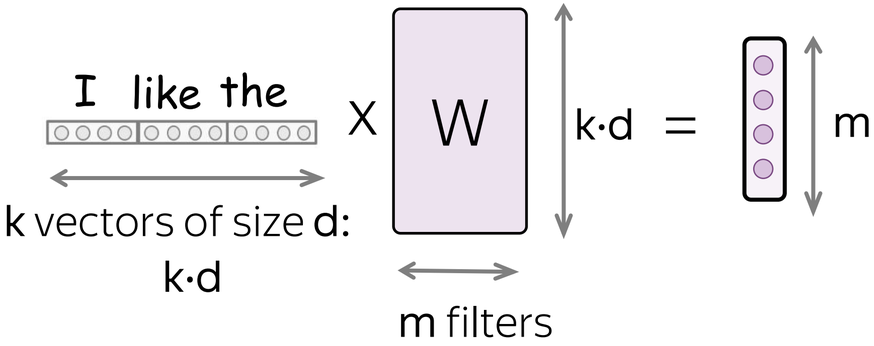
(except for a non-linearity)





Kernel size k (k=3)
(convolution window size)

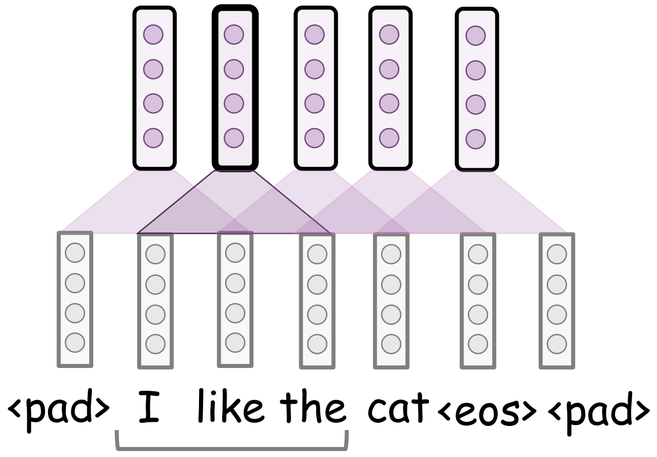
Convolution is a linear layer
mapping from k·d to m

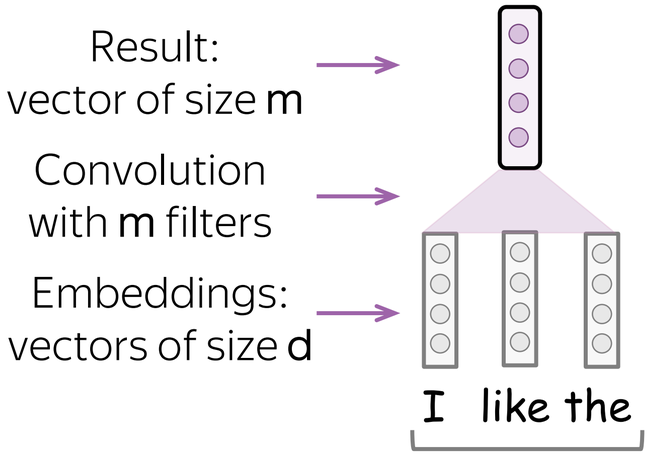


- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)
- d (input channels) – input vector size
- k (kernel size) – conv. window length
- m (output channels) – number of filters

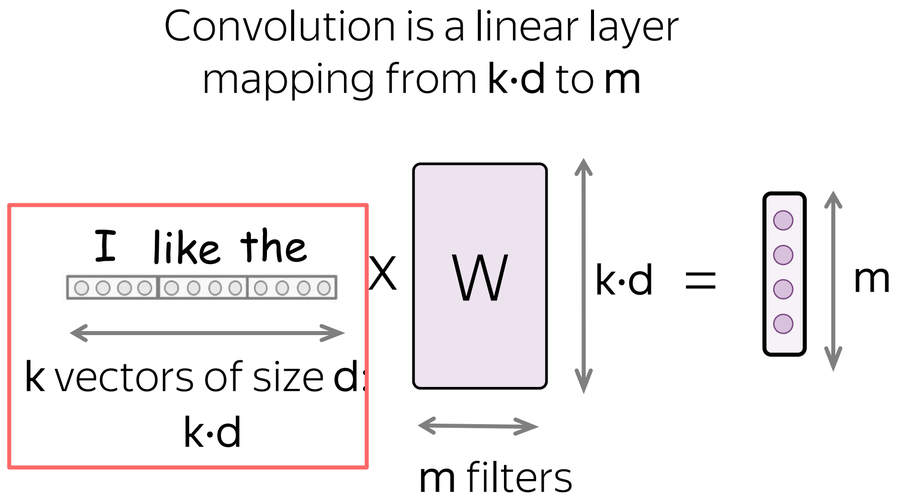
Convolution is a Linear Operation Applied to Each Window

(except for a non-linearity)





Kernel size k (k=3)
(convolution window size)

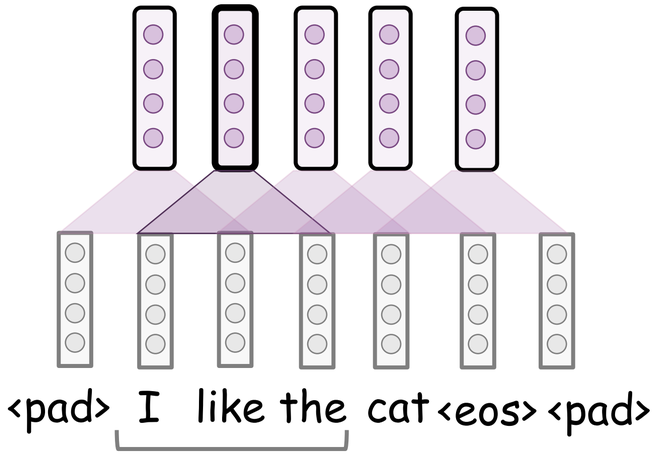


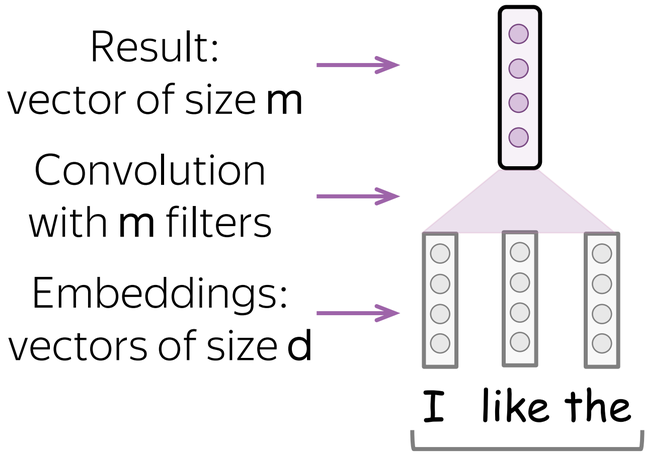
- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)
- d (input channels) – input vector size
- k (kernel size) – conv. window length
- m (output channels) – number of filters

$u_i = [x_i, \ldots, x_{i+k-1}] \in \mathbb{R}^{k \cdot d}$ - concatenate representations in the i-th window

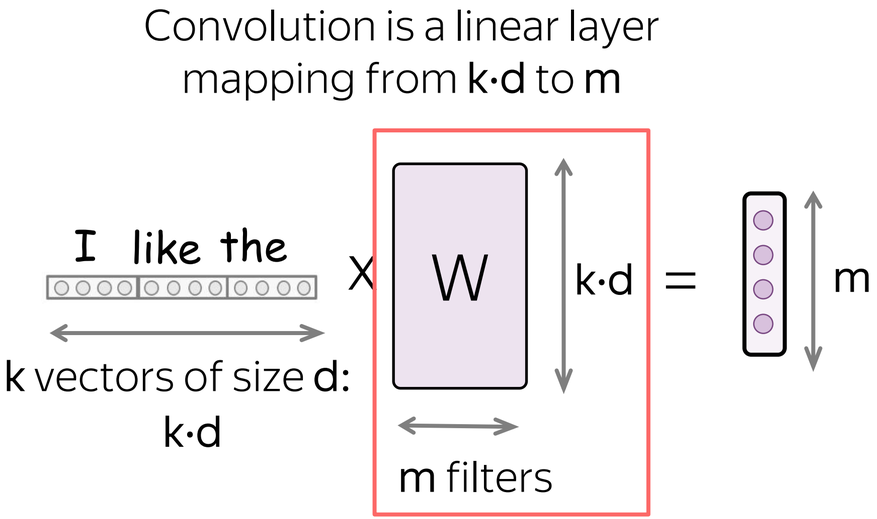
Convolution is a Linear Operation Applied to Each Window

(except for a non-linearity)





Kernel size k (k=3)
(convolution window size)



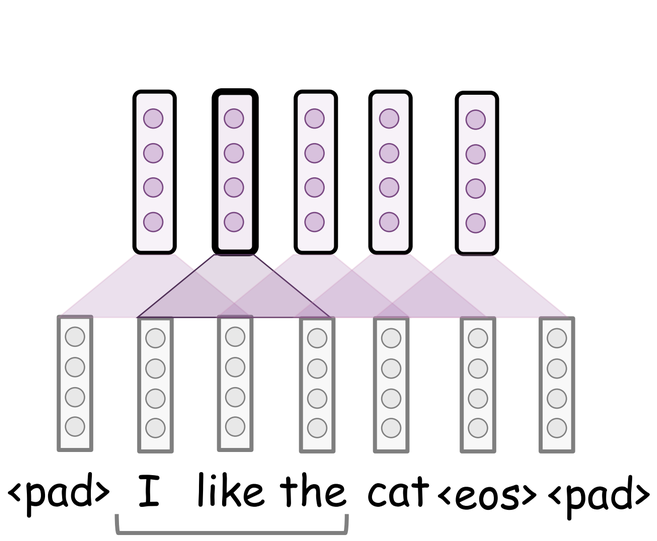
- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)
- d (input channels) – input vector size
- k (kernel size) – conv. window length
- m (output channels) – number of filters

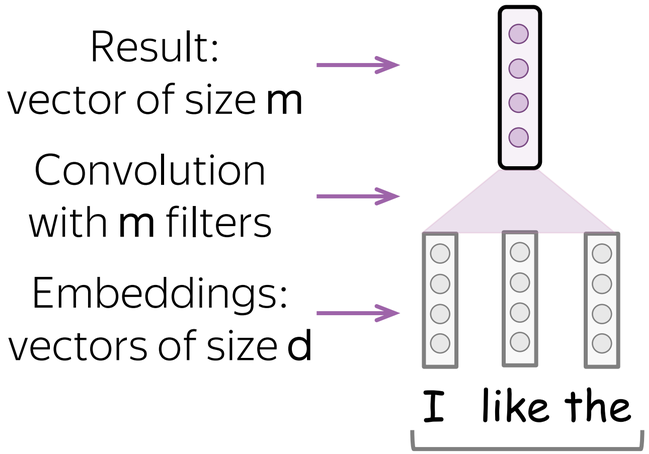
$u_i = [x_i, \ldots, x_{i+k-1}] \in \mathbb{R}^{k \cdot d}$ - concatenate representations in the i-th window

$W \in \mathbb{R}^{(k \cdot d) \times m}$ - convolution

Convolution is a Linear Operation Applied to Each Window

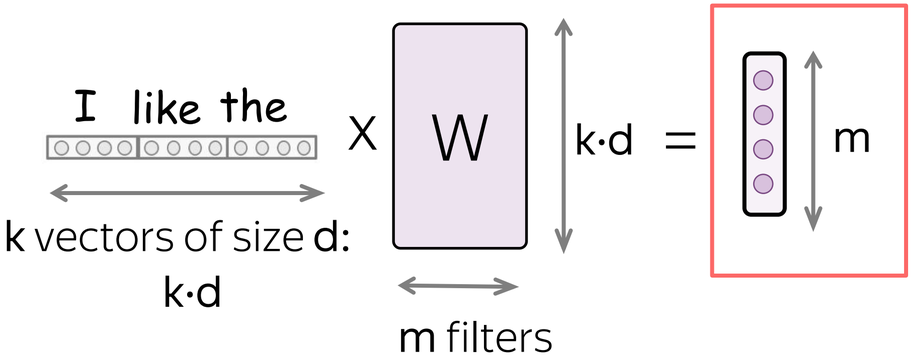
(except for a non-linearity)





Kernel size k (k=3)
(convolution window size)

Convolution is a linear layer
mapping from k·d to m



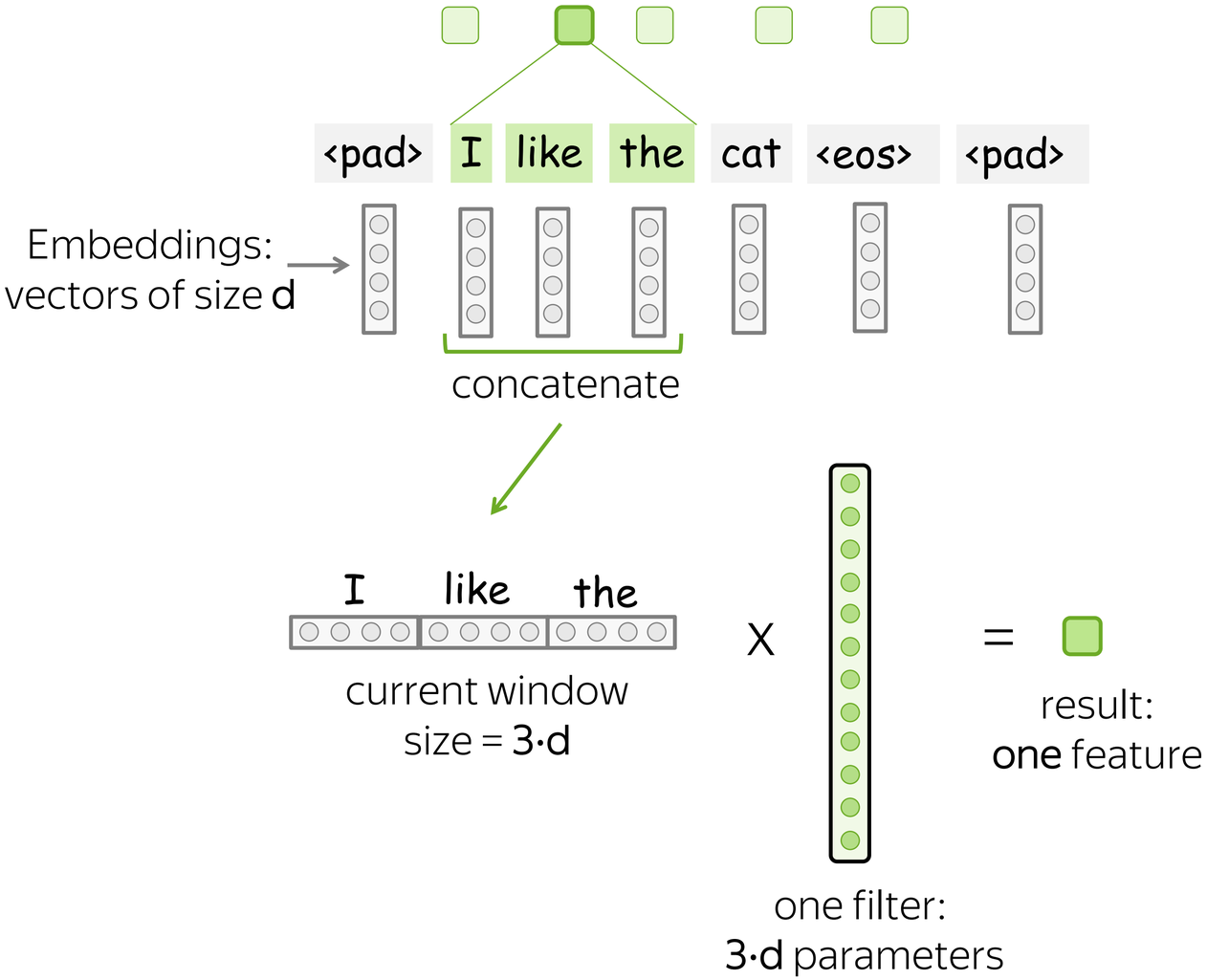
- $x_1, x_2, \ldots, x_n$ - input vectors (e.g., word embeddings)
- d (input channels) – input vector size
- k (kernel size) – conv. window length
- m (output channels) – number of filters

$u_i = [x_i, \ldots, x_{i+k-1}] \in \mathbb{R}^{k \cdot d}$ - concatenate representations in the i-th window

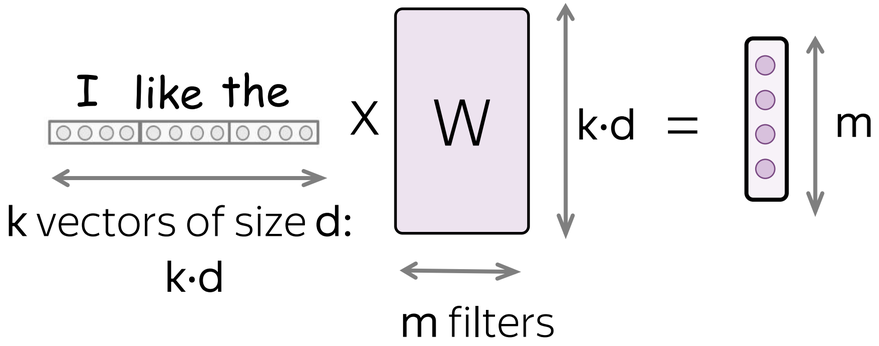
$W \in \mathbb{R}^{(k \cdot d) \times m}$ - convolution

$F_i = u_i \times W$ – convolution applied to the i-th window

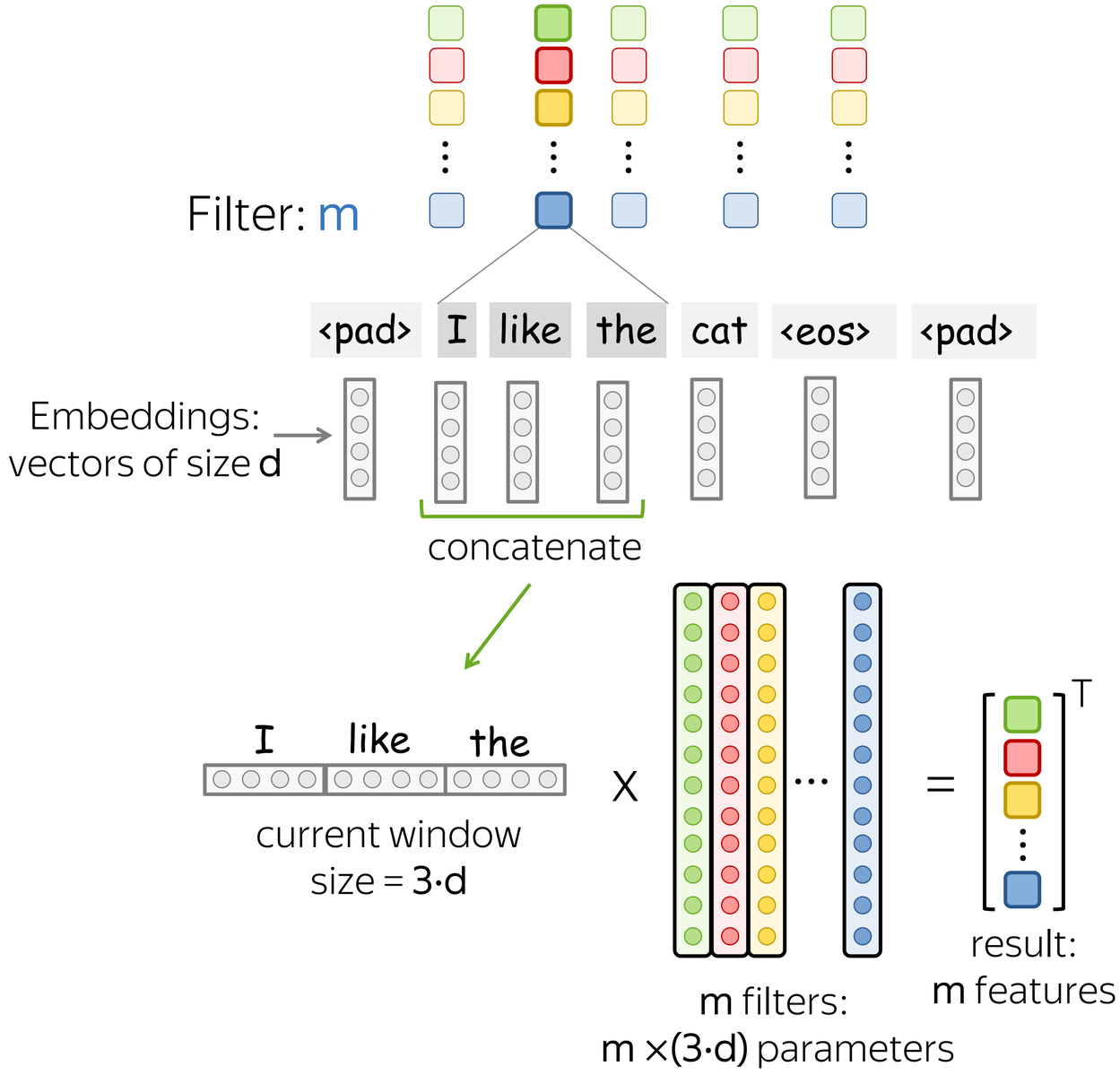
Intuition: A Filter is a Feature Extractor



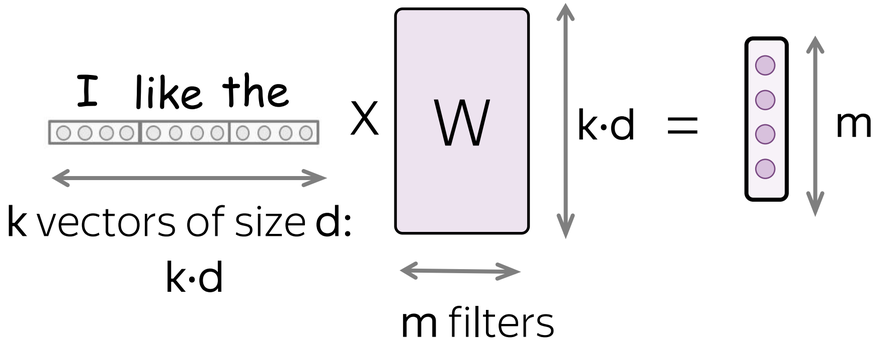
Convolution is a linear layer
mapping from **k·d** to **m**



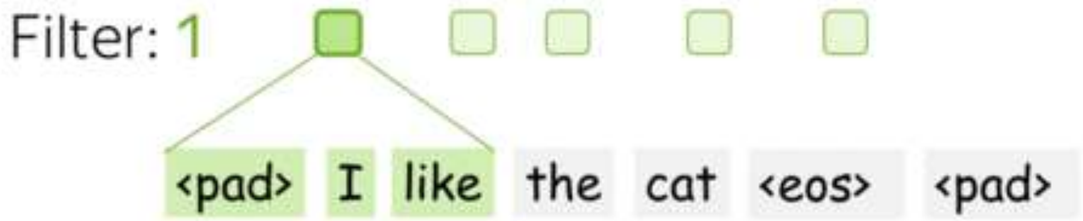
Several Filters – Several Features



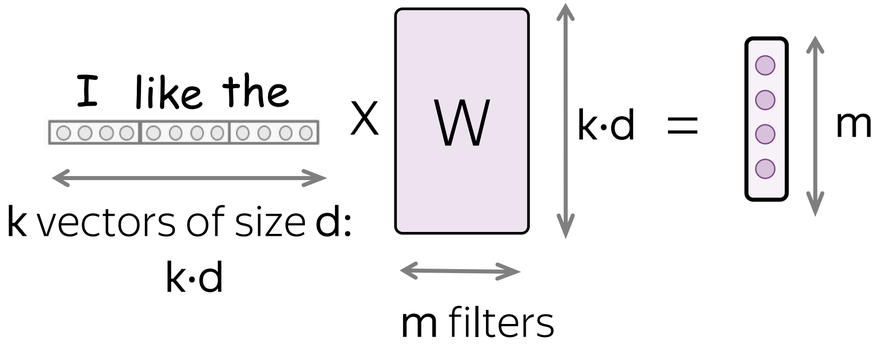
Convolution is a linear layer
mapping from **k·d** to **m**



Several Filters – Several Features



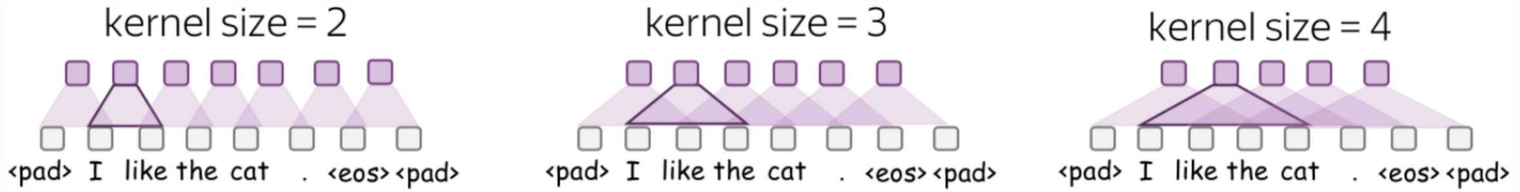
Convolution is a linear layer
mapping from **k·d** to **m**



Convolution: Parameters

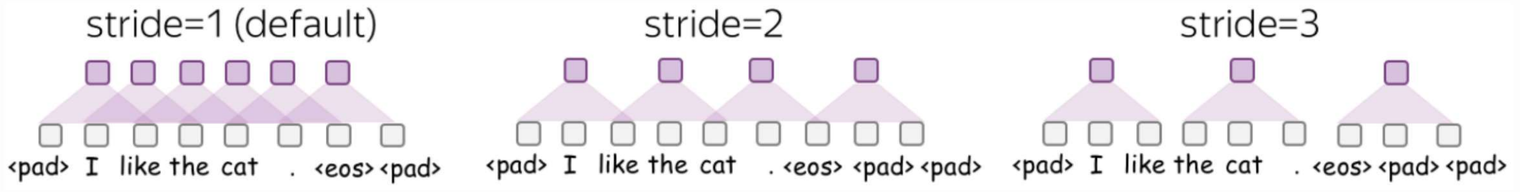
- **Kernel size**: How far to look

Kernel size is the number of input elements (tokens) a convolution looks at each step. For text, typical values are 2-5.

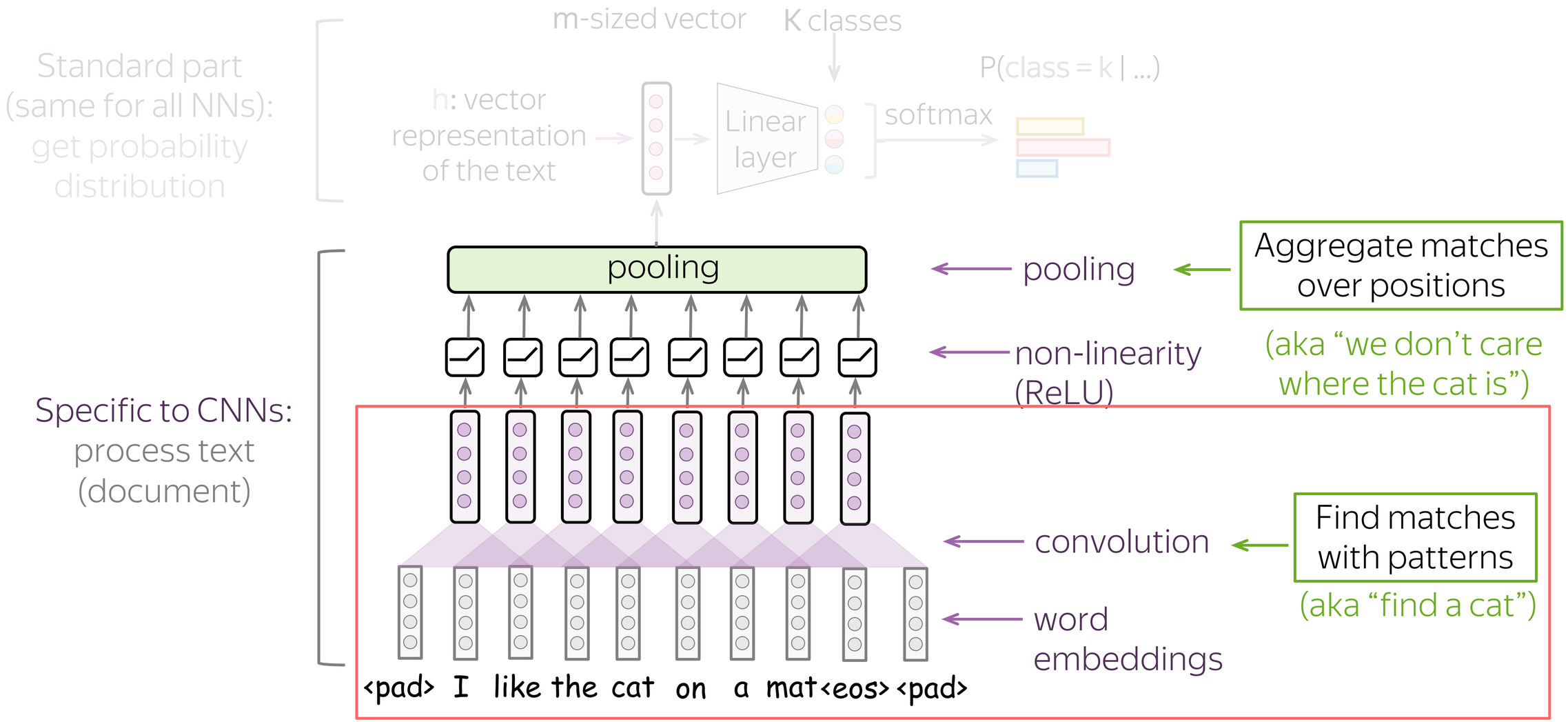


- **Stride**: How much move a filter at each step

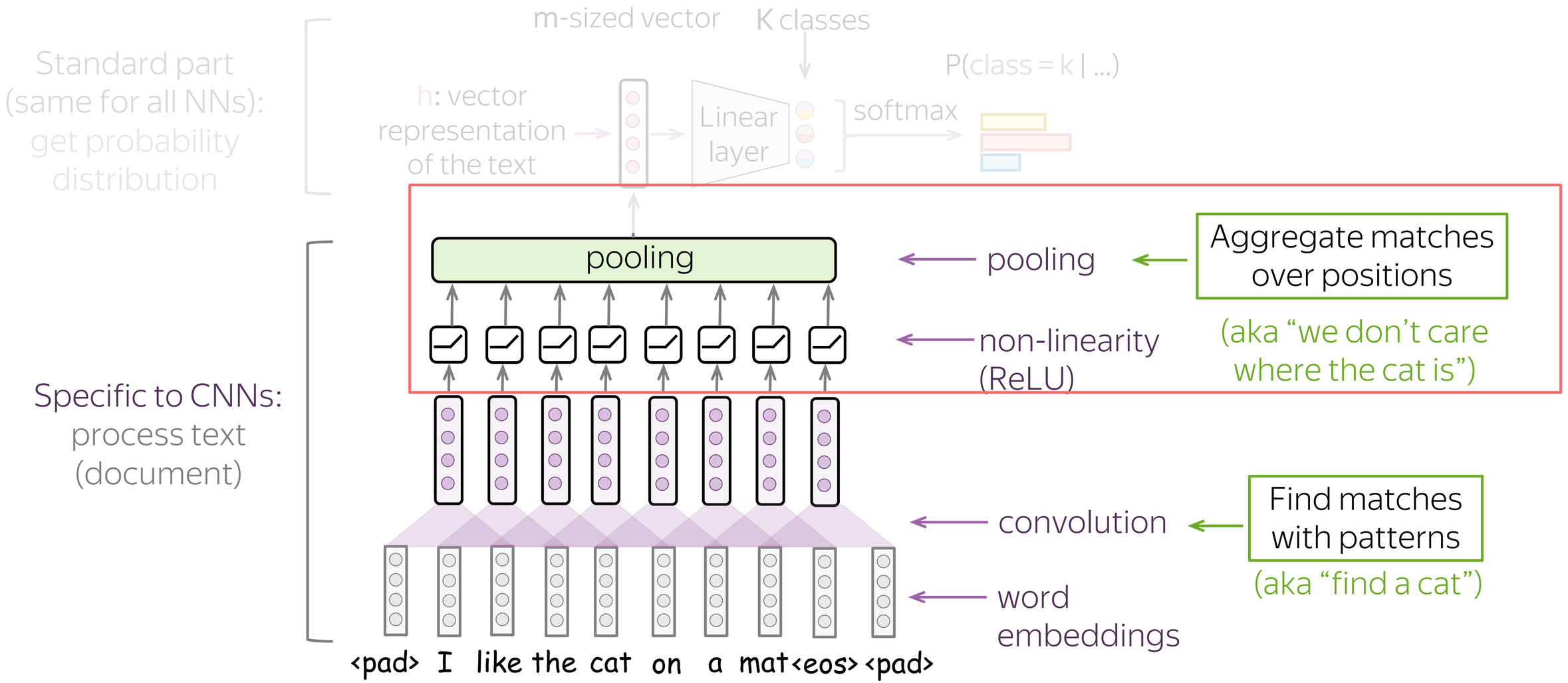
Stride tells how much to move filter at each step. For example, stride equal to 1 means that we move the filter by 1 input element (pixel for images, token for texts) at each step.



A Typical Model: Convolution + Pooling

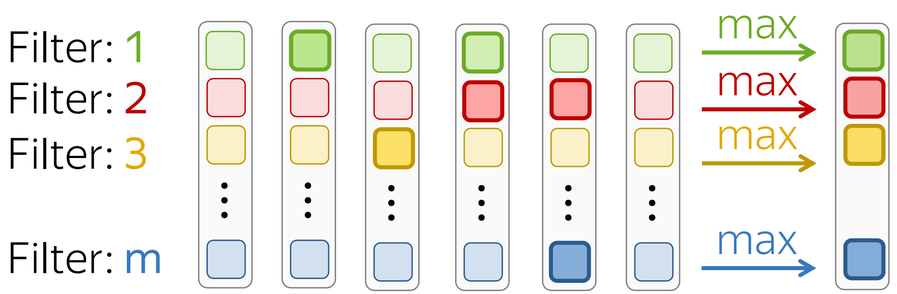


A Typical Model: Convolution + Pooling



Building Blocks: Pooling (max, mean, etc)

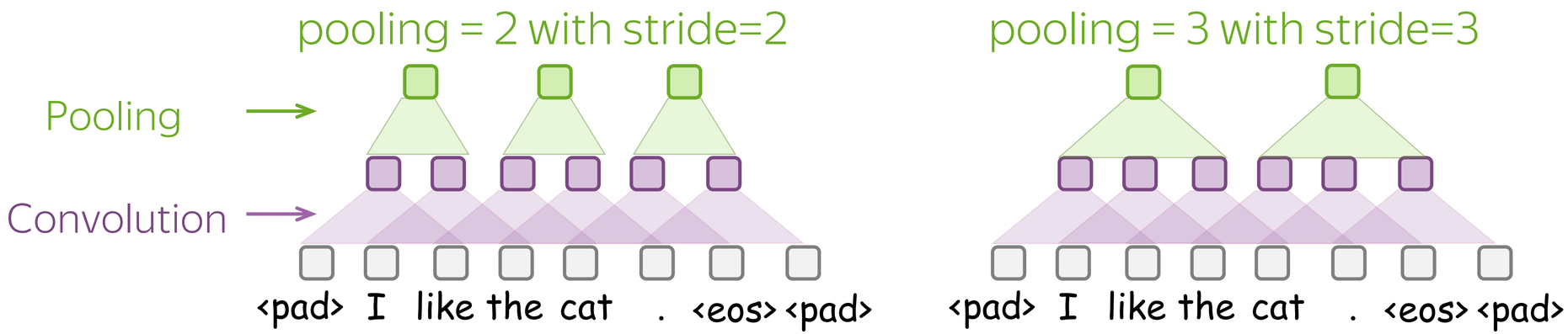
- **Max pooling**: maximum for each dimension (feature)
- **Mean pooling**: mean value for each dimension (feature)



Max pooling: maximum for each dimension (feature)

Building Blocks: Pooling and Global Pooling

- **Pooling**: aggregate features in some area

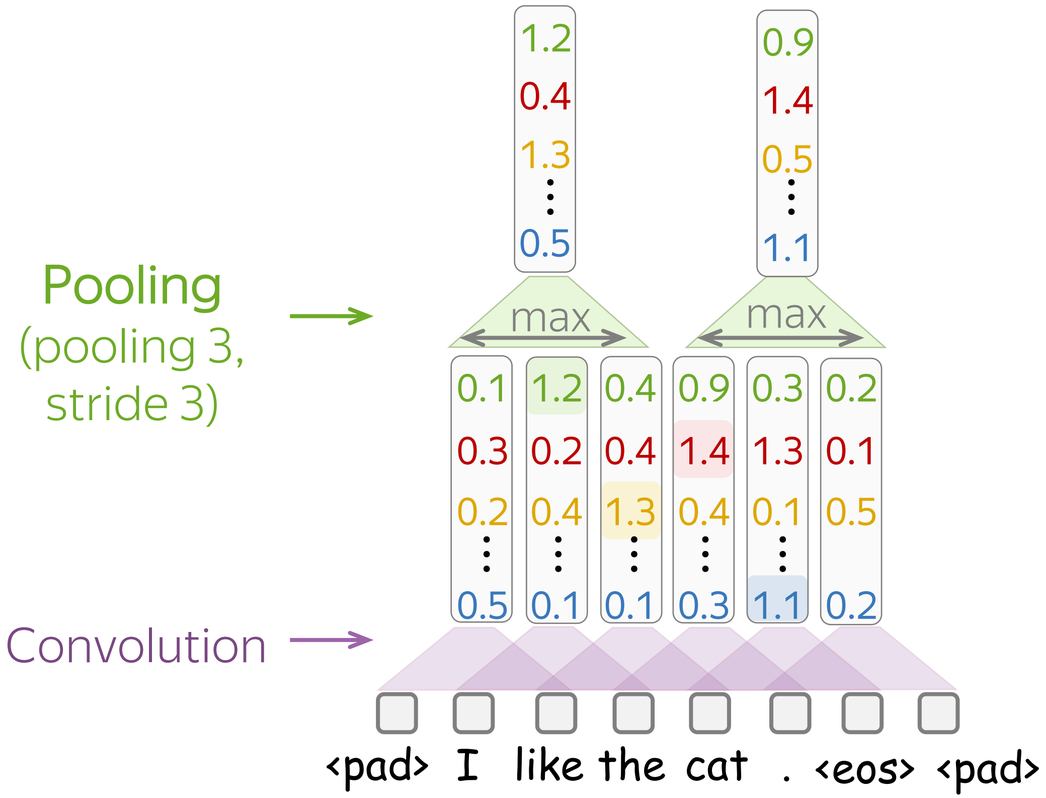


Building Blocks: Pooling and Global Pooling

- Pooling: aggregate features in some area
- Global pooling: aggregate features over all input

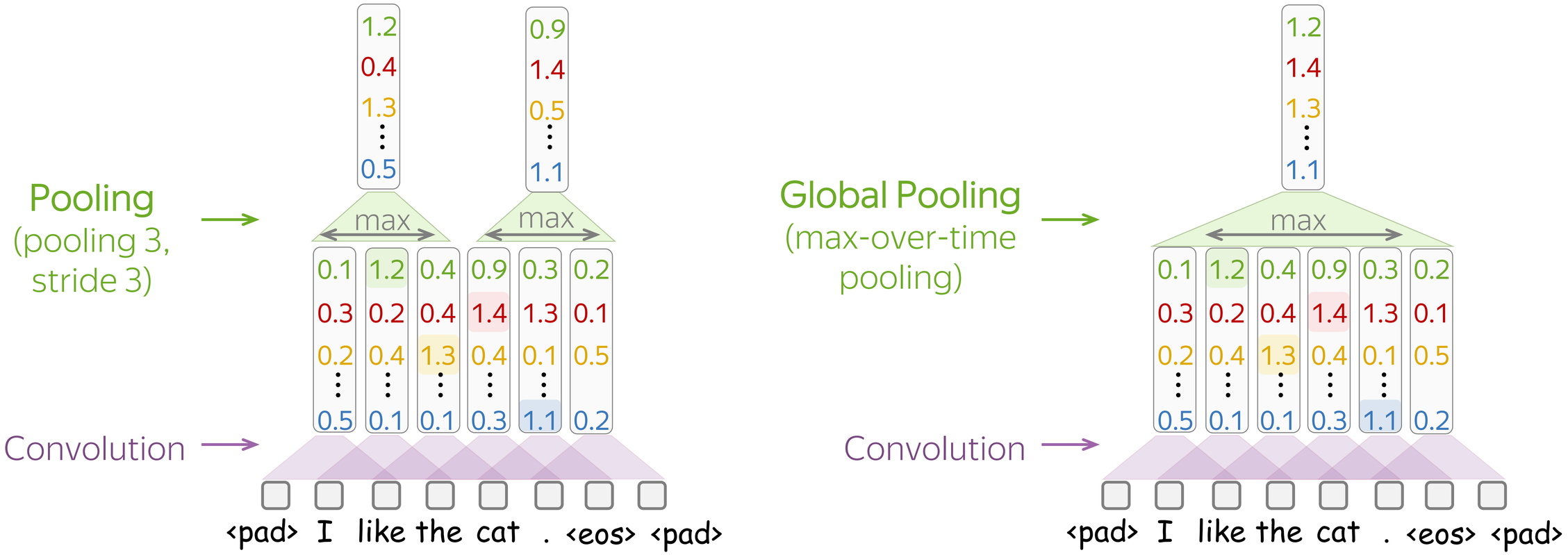
Building Blocks: Pooling and Global Pooling

- **Pooling**: aggregate features in some area
- **Global pooling**: aggregate features over all input

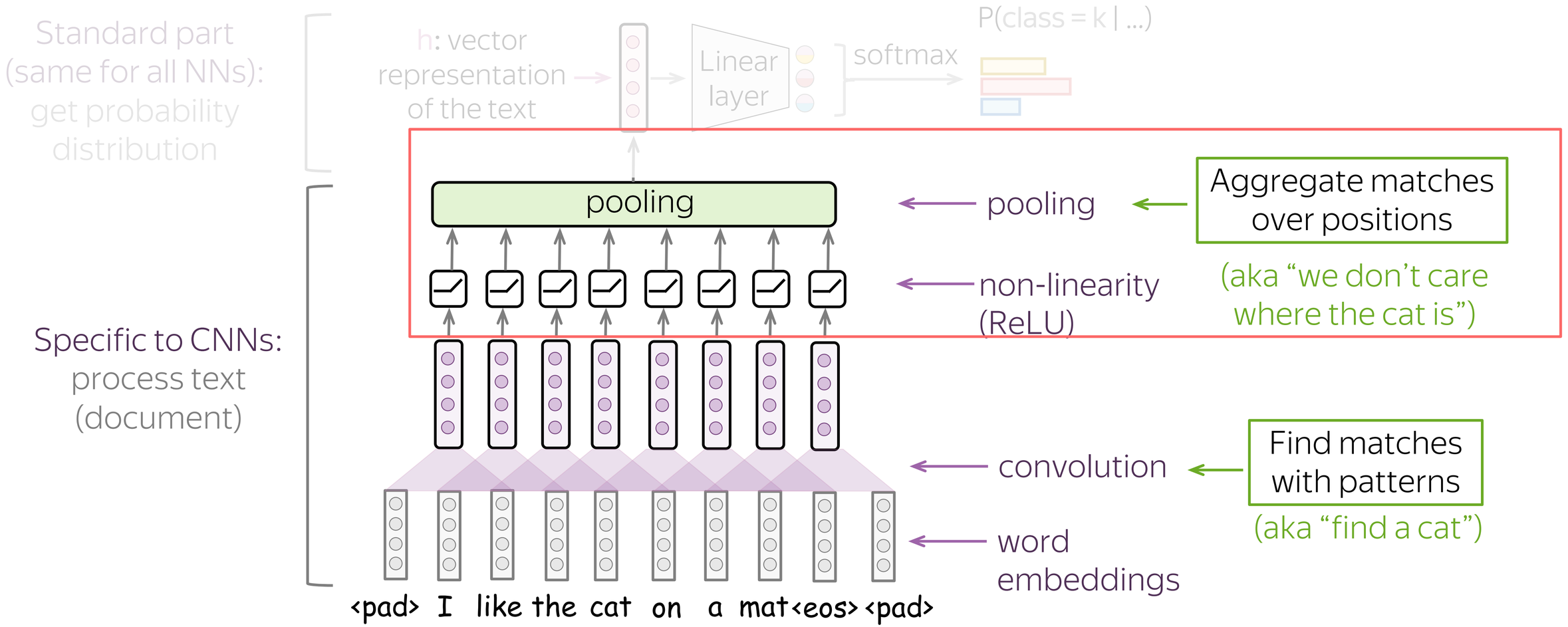


Building Blocks: Pooling and Global Pooling

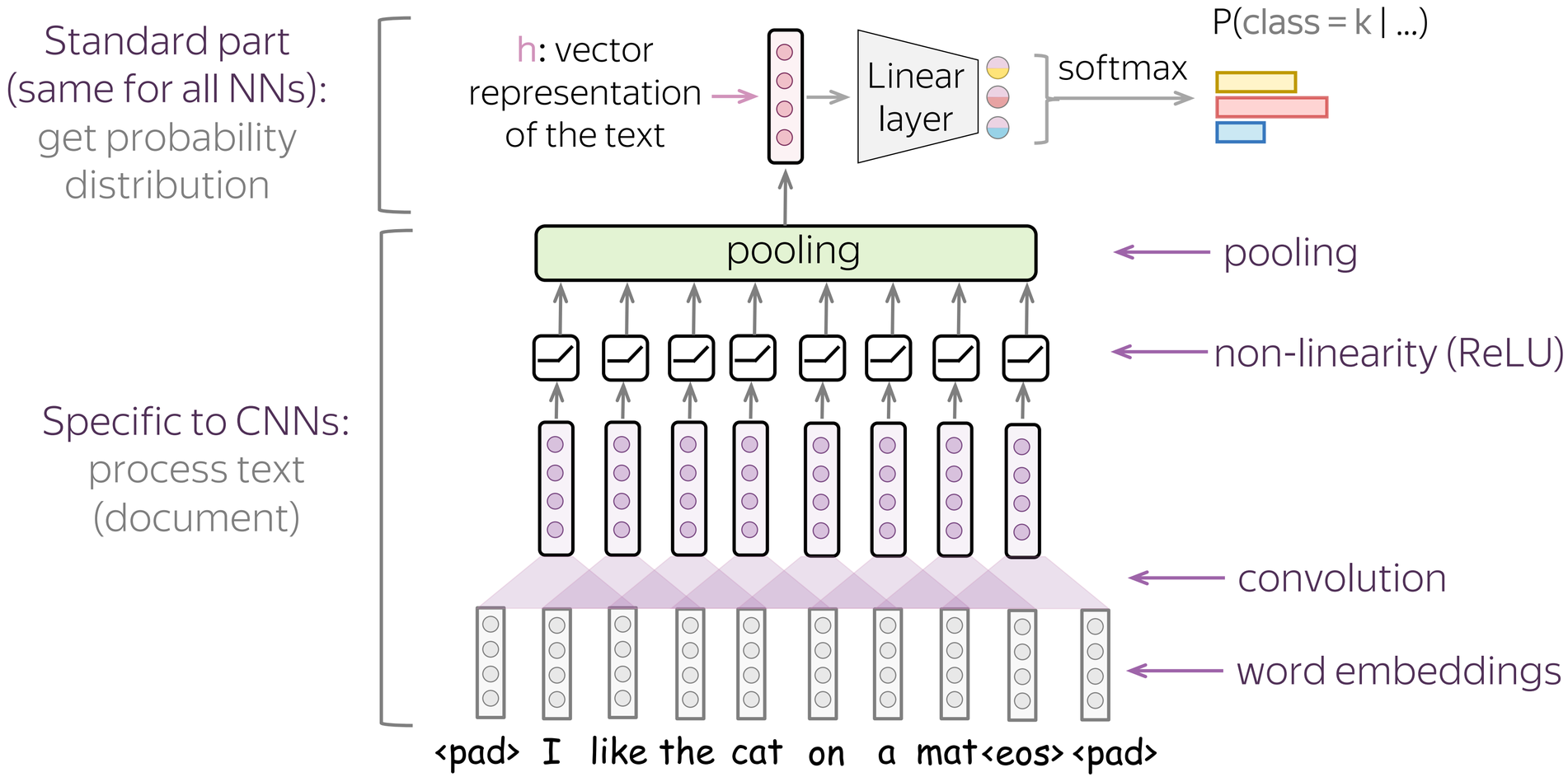
- **Pooling**: aggregate features in some area
- **Global pooling**: aggregate features over all input



A Typical Model: Convolution + Pooling

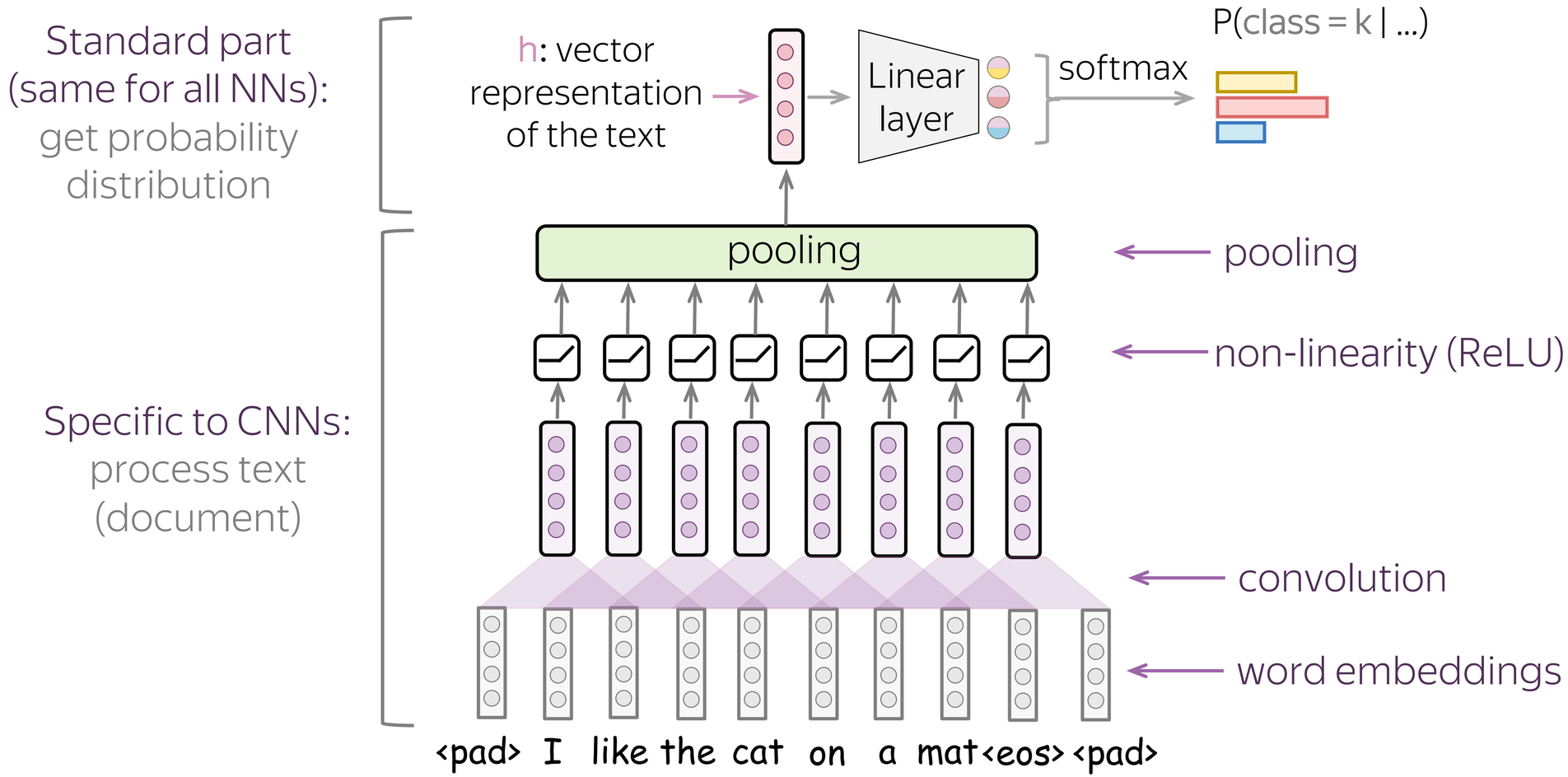


A Typical Model: Convolution + Pooling



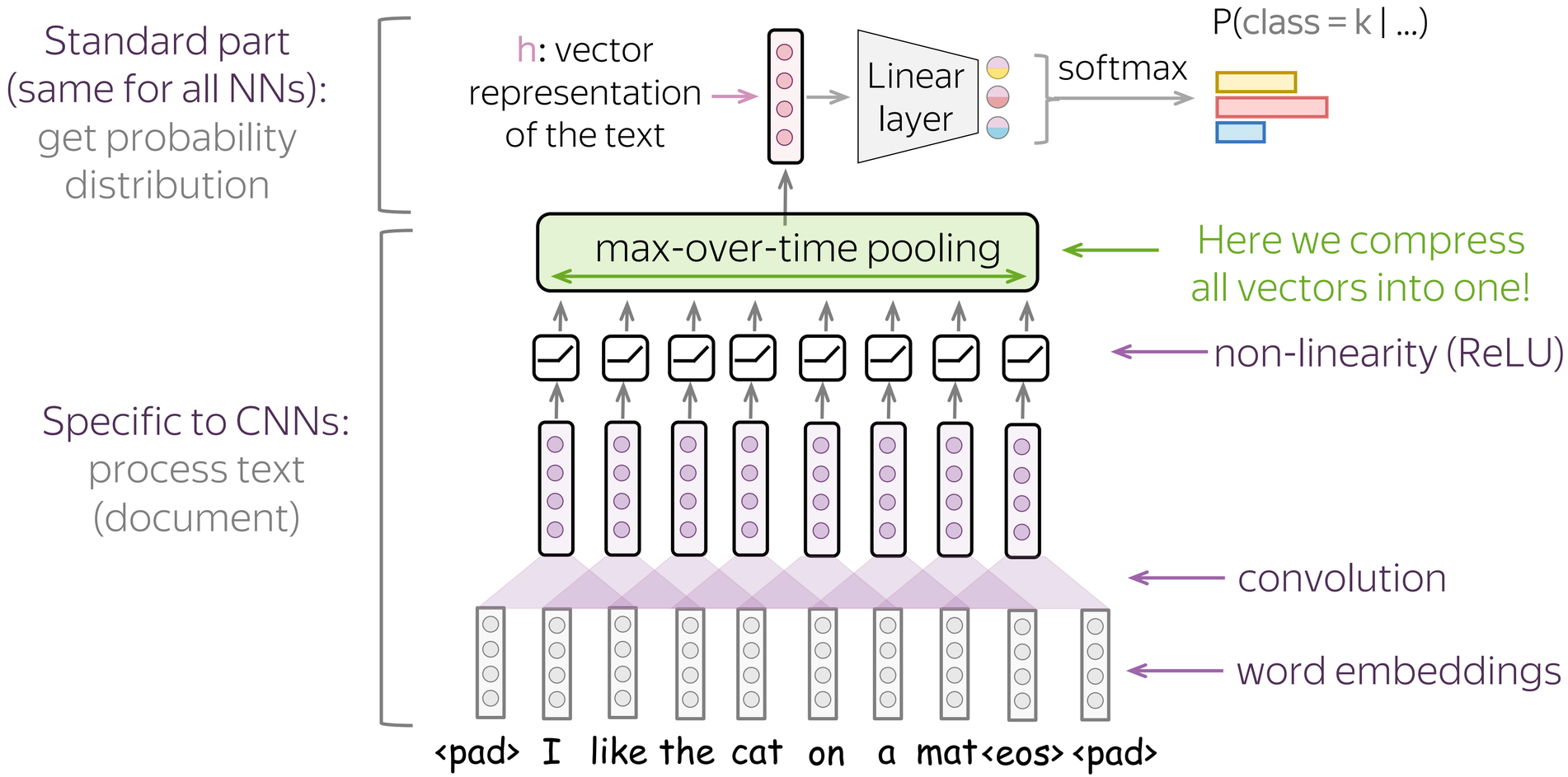
A Typical Model: Convolution + Pooling

> We need a model that can produce a **fixed-sized** vector for inputs of **different** lengths.



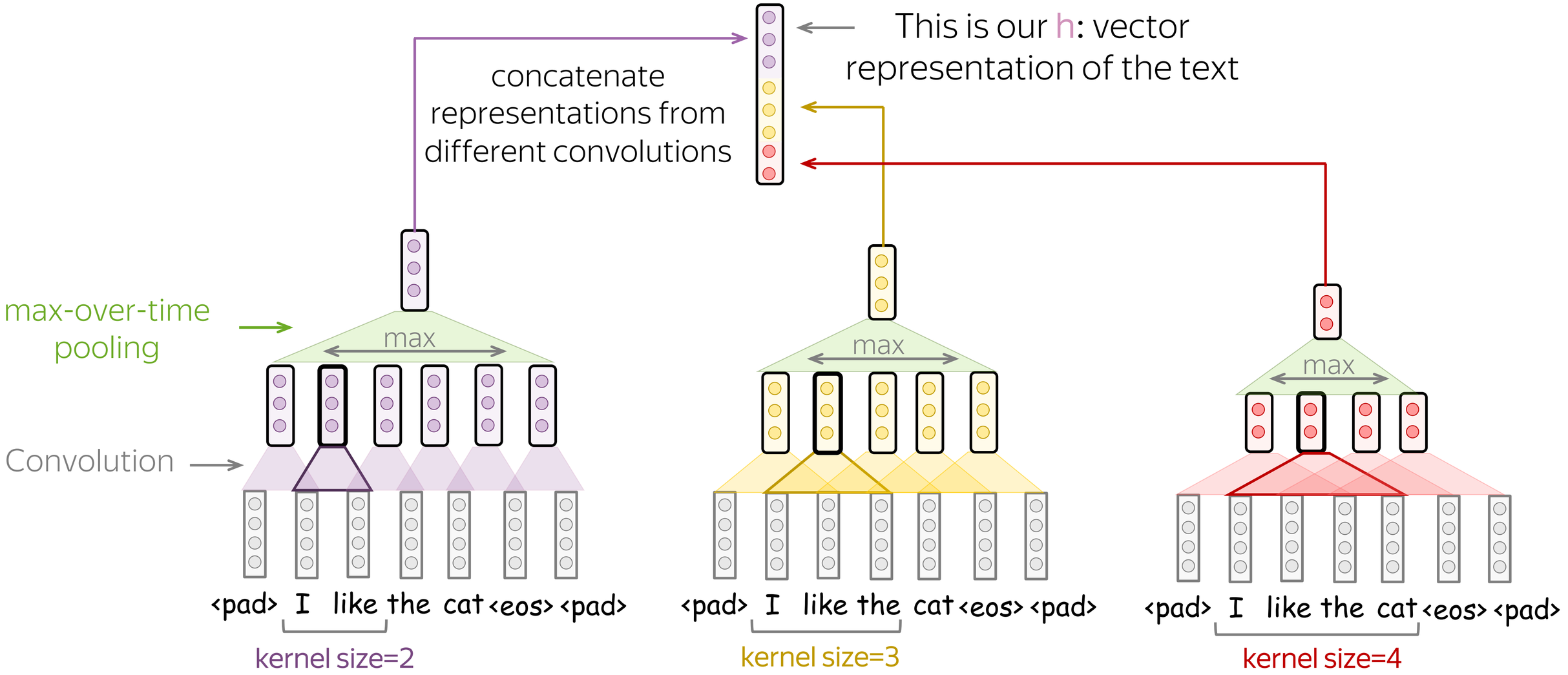
A Typical Model: Convolution + Pooling

> We need a model that can produce a **fixed-sized** vector for inputs of **different** lengths.



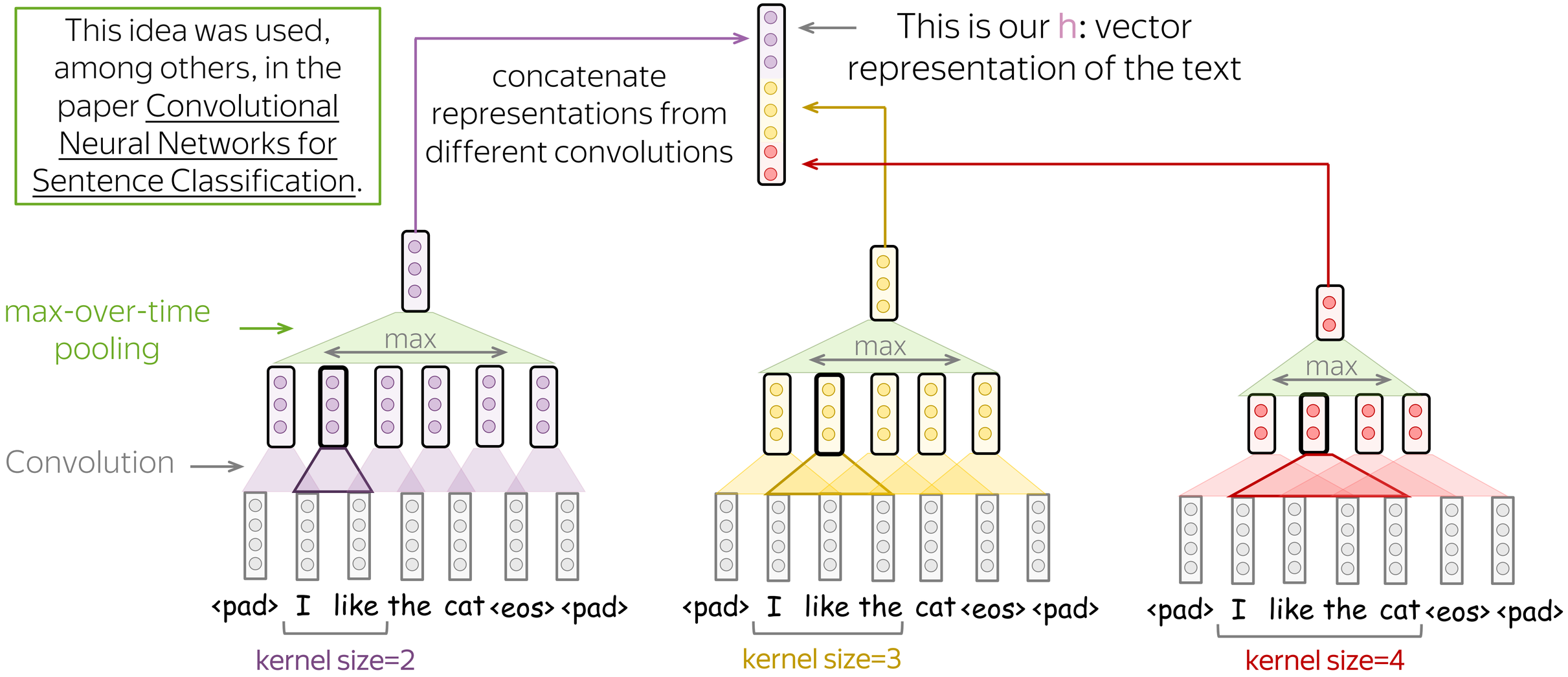
Convolutional Models for Text Classification

- Several Convolutions with Different Kernel Sizes



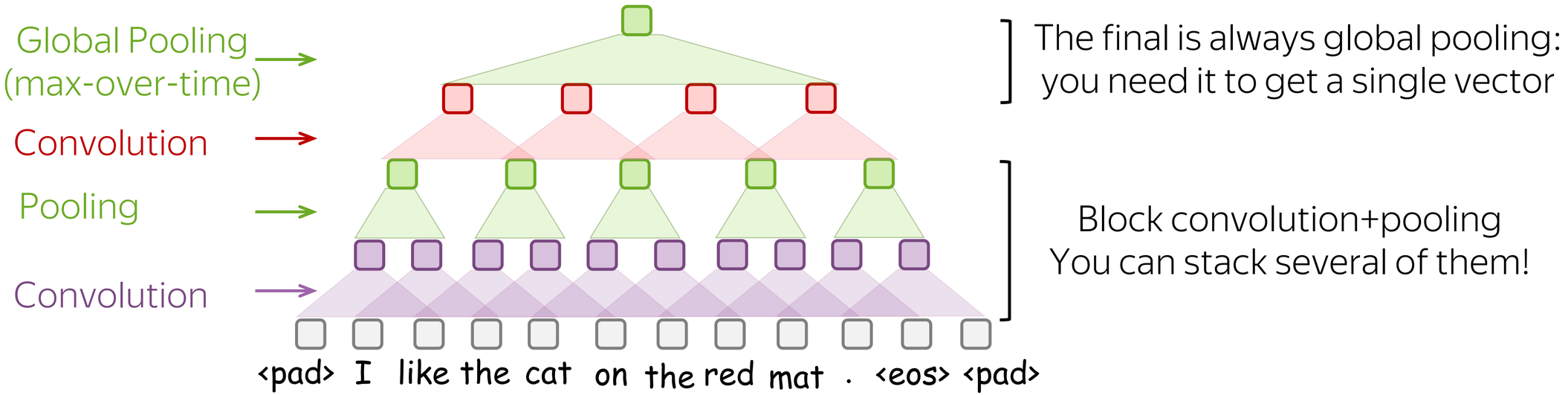
Convolutional Models for Text Classification

- Several Convolutions with Different Kernel Sizes



Convolutional Models for Text Classification

- Stack Several Blocks Convolution+Pooling



Convolutional Models for Text Classification

- Stack Several Blocks Convolution+Pooling

This idea was used, among others, in the paper
Character-level Convolutional Networks for Text Classification.