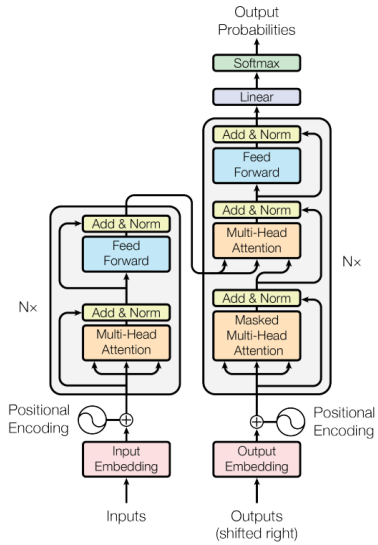
# Transformer

# A universal deep learning architecture




## Learning goals

- Understand the initial use of the Transformer
- Grasp the other application fields since then

# MACHINE TRANSLATION

- Sequence-to-sequence task
- Already served as a motivation for introducing the "ordinary" Attention-mechanism ▸ Bahdanau et al., 2015
- Crucial, that the decoder has access to the whole input sequence
  $\rightarrow$ This is very well solved by cross-attention
- Good contextualization in the encoder improves translation quality
  - (Bidirectional) RNNs/LSTMs are only (concatenated) unidirectional architectures
  - Transformer-Encoder layers are bidirectional by construction
  - Stacking them on top of each other makes this bidirectional contextualization even "deeper"

# WMT 2014 EN-TO-DE AND EN-TO-FR

## Parallel training data:

Parallel data:

| File | Size | CS-EN | DE-EN | HI-EN | FR-EN | RU-EN | Notes |
|---|---|---|---|---|---|---|---|
| Europarl v7 | 628MB | ✓ | ✓ | | ✓ | | same as previous year, corpus home page |
| Common Crawl corpus | 876MB | ✓ | ✓ | | ✓ | ✓ | same as previous year |
| UN corpus | 2.3GB | | | | ✓ | | same as previous year, corpus home page |
| News Commentary | 77MB | ✓ | ✓ | | ✓ | ✓ | updated, data with document boundaries |
| $10^9$French-English corpus | 2.3 GB | | | | ✓ | | same as previous year [md5\|sha1] |
| CzEng 1.0 | 115MB | ✓ | | | | | same as previous year, corpus home page (avoid sections 98 and 99) |
| Yandex 1M corpus | 121MB | | | | | ✓ | corpus home page; v1.3 now in original case |
| Wiki Headlines | 7.8MB | | | ✓ | | ✓ | Provided by CMU. The ru-en is unchanged from last year. |
| HindEnCorp | 25MB | | | ✓ | | | Collected by Charles University |
| ~~The JHU Corpus~~ | | | | ~~✓~~ | | | This is fully contained in HindEnCorp, so not made available here. |

# THE BLEU SCORE

- Based on n-gram overlap from candidate and reference sentence
- Precision (for each n-gram) calculated as follows:

$$P_n = \frac{\sum_{C \in \{\text{ Candidates }\}} \sum_{n-gram \in C} \text{ Count }_{\text{clip}} (n- \text{ gram })}{\sum_{C' \in \{\text{ Candidates }\}} \sum_{n-gram' \in C'} \text{ Count } (n- \text{ gram }')}$$

- Finally, the BLEU score can be computed as

$$\texttt{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log(P_n)\right),$$

- where $BP$ is a "brevity penalty" to penalize short generations, $N$ is the number of n-grams & $w_n$ the weight for each $P_n$ (usually $\frac{1}{N}$)

# TRANSFORMER FOR MT

The Transformer ..

- .. outperforms the previous SOTA models
- .. at a lower number of required training FLOPs

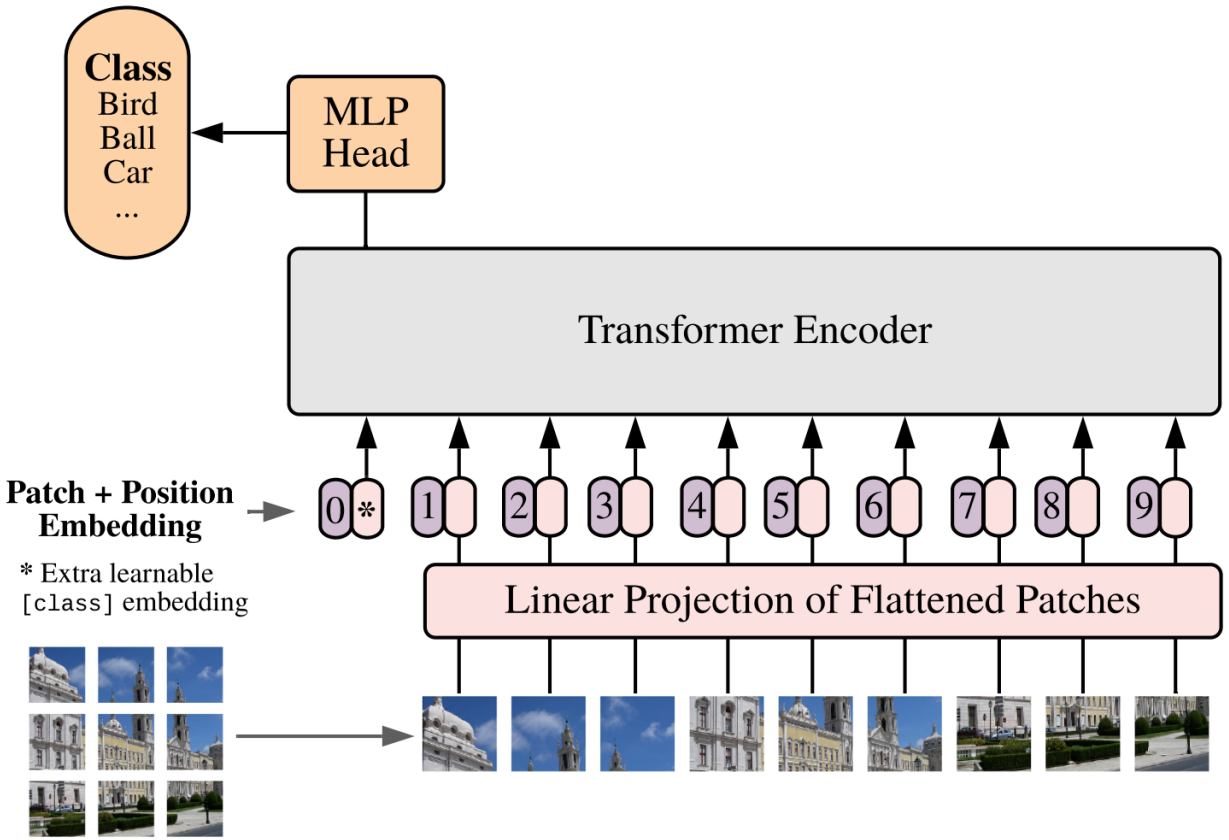
Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

| Model | BLEU | | Training Cost (FLOPs) | |
|---|---|---|---|---|
| | EN-DE | EN-FR | EN-DE | EN-FR |
| ByteNet [18] | 23.75 | | | |
| Deep-Att + PosUnk [39] | | 39.2 | | $1.0 \cdot 10^{20}$ |
| GNMT + RL [38] | 24.6 | 39.92 | $2.3 \cdot 10^{19}$ | $1.4 \cdot 10^{20}$ |
| ConvS2S [9] | 25.16 | 40.46 | $9.6 \cdot 10^{18}$ | $1.5 \cdot 10^{20}$ |
| MoE [32] | 26.03 | 40.56 | $2.0 \cdot 10^{19}$ | $1.2 \cdot 10^{20}$ |
| Deep-Att + PosUnk Ensemble [39] | | 40.4 | | $8.0 \cdot 10^{20}$ |
| GNMT + RL Ensemble [38] | 26.30 | 41.16 | $1.8 \cdot 10^{20}$ | $1.1 \cdot 10^{21}$ |
| ConvS2S Ensemble [9] | 26.36 | **41.29** | $7.7 \cdot 10^{19}$ | $1.2 \cdot 10^{21}$ |
| Transformer (base model) | 27.3 | 38.1 | **$3.3 \cdot 10^{18}$** | |
| Transformer (big) | **28.4** | **41.8** | $2.3 \cdot 10^{19}$ | |

▸ Source: Vaswani et al., 2017

# TRANSFORMER FOR COMPUTER VISION

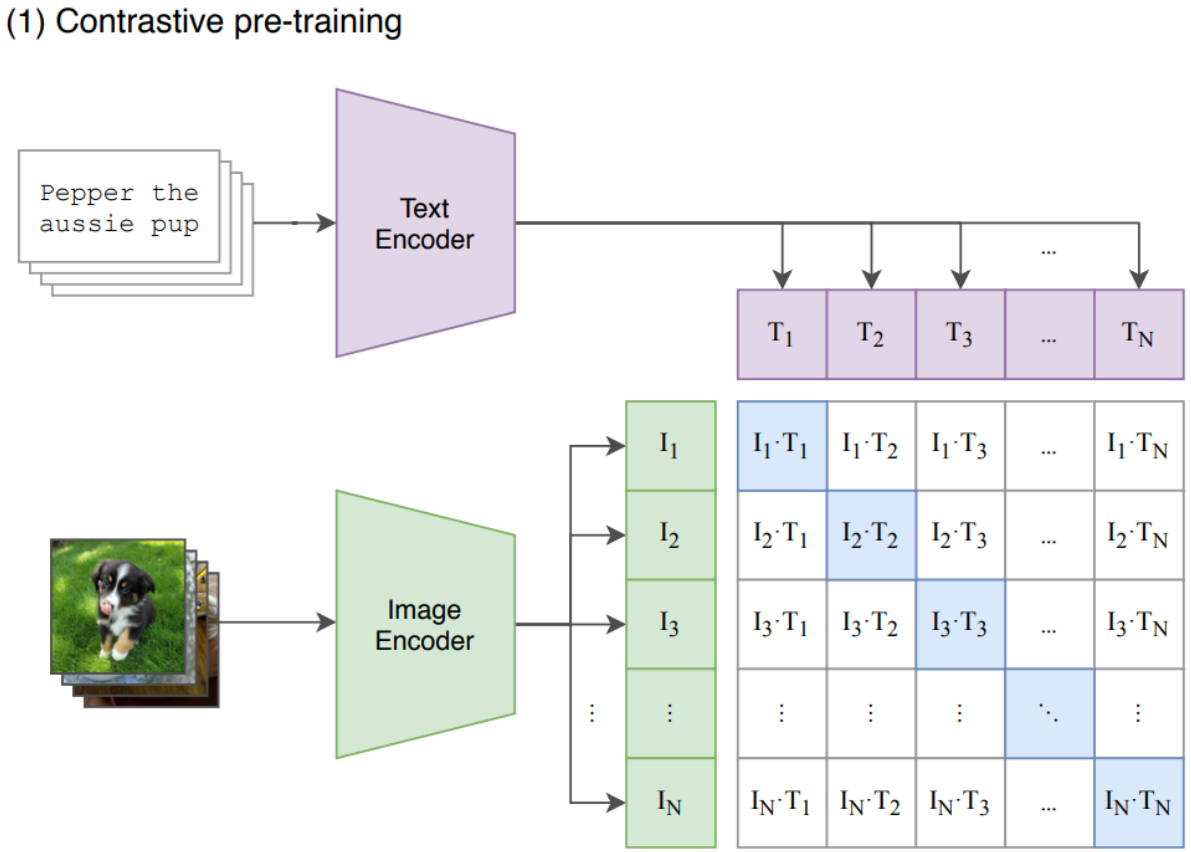
## Vision Transformer




▸ Source: Dosovitskiy et al., 2020

# TRANSFORMER FOR MULTIMODAL LEARNING

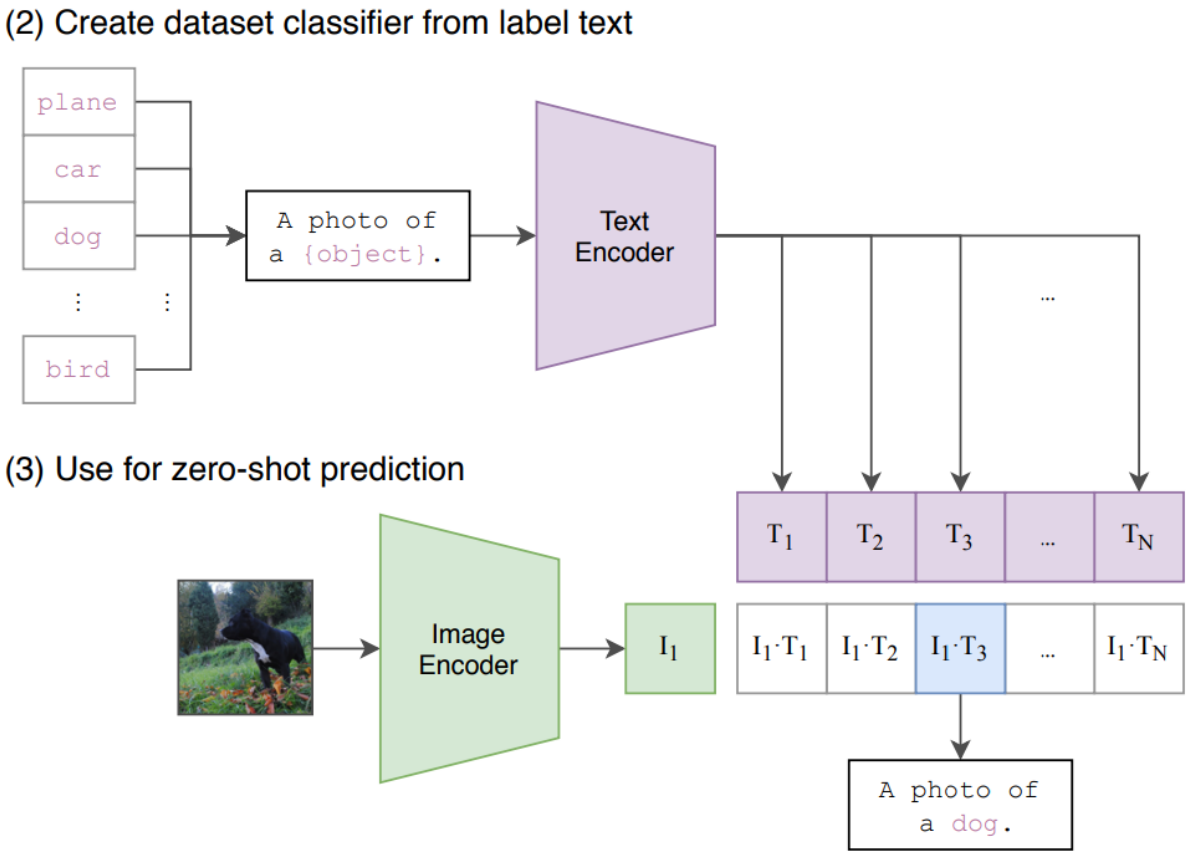
## CLIP




▸ Source: Radford et al., 2021

# TRANSFORMER FOR MULTIMODAL LEARNING

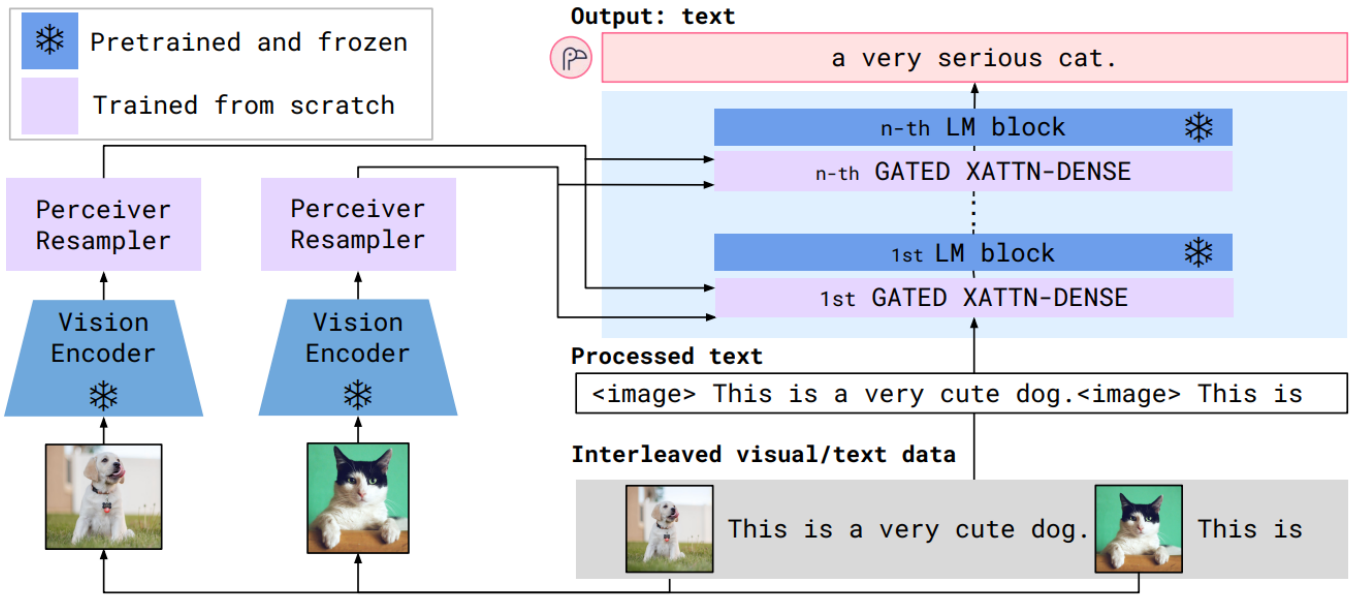
## CLIP




▸ Source: Radford et al., 2021

# TRANSFORMER FOR MULTIMODAL LEARNING

## FLAMINGO




▸ Source: Alayrac et al., 2022

# TRANSFORMER FOR MULTIMODAL LEARNING

## GPT-4

| **Example of GPT-4 visual input:** | |
|---|---|
| User | What is funny about this image? Describe it panel by panel. |
| |  <br>Source: https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/ |
| GPT-4 | The image shows a package for a "Lightning Cable" adapter with three panels.<br><br>Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.<br><br>Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.<br><br>Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.<br><br>The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port. |

**Table 3.** Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question about an image with multiple panels which GPT-4 is able to answer.

▸ Source: OpenAI, 2023