Upgrading the Newsroom:
An Automated Image Selection System for News Articles

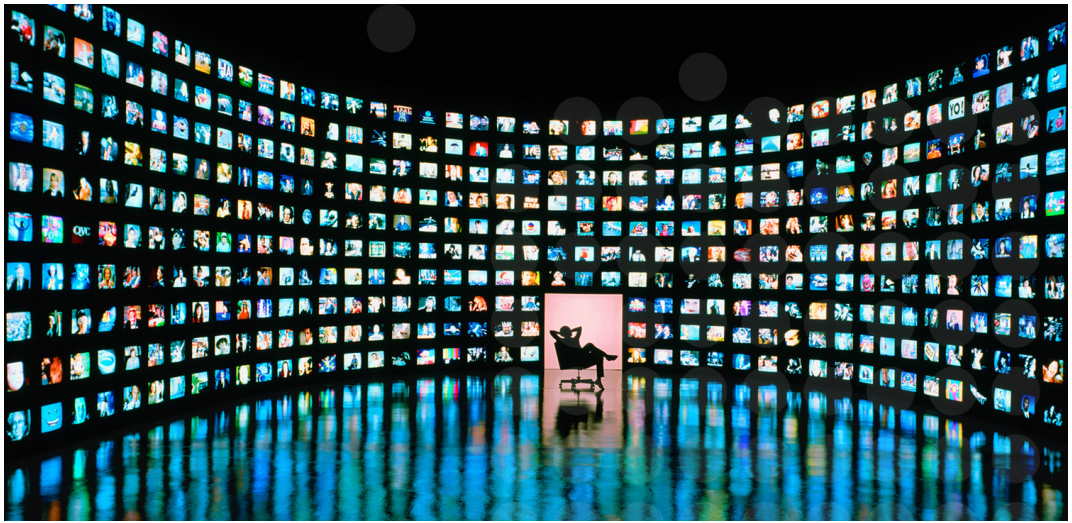
Fangyu Liu
Undergrad @UWaterloo
Intern @LSIR.EPFL

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Some Backgrounds

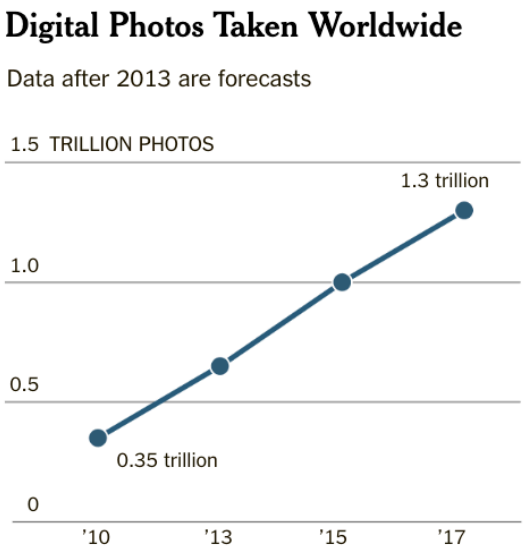
The Photo Libarian

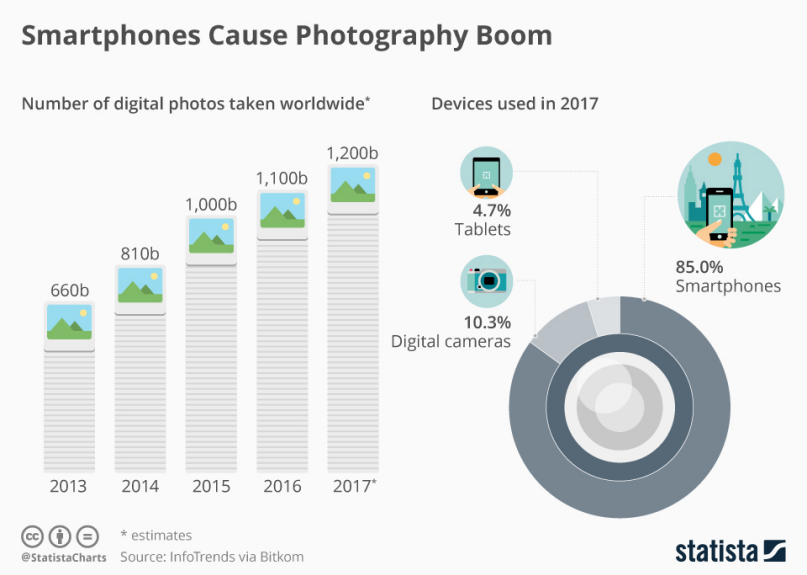
*A **picture editor**, sometimes known as a **photo editor**, is a professional who collects, reviews, and chooses photographs and/or illustrations for publication in alignment with preset guidelines…Sometimes photo editors are in charge of an image data base ; they are called photo librarians.* (wikipedia)



The Photo Libarian

New technologies and the omnipresence of images nowadays have drastically changed the way picture editors work. ***They do their research mostly on the Internet, and have to browse a never-ending flow of images.*** (wikipedia again)





("Photos, Photos Everywhere", The New York Times & statista)

The Photo Libarian

("The Newsroom", HBO)

It should no longer be human 👩🏻 👨🏻 's job to remember a database.

We use a Neural Network 😇😇 to do it for us!

To train a Neural Network we need a dataset

title

lead

image

Image caption

article

«Ich hatte das Gefühl, keine Luft zu bekommen»

Roger Federer erklärt seine Niederlage im US-Open-Achtelfinal gegen John Millman mit den extremen Bedingungen.

Start nach Mass
1|7 Im ersten Satz läuft alles nach Plan: Federer haut seinem Gegner die Bälle ins Feld, Millman ist klar unterlegen.

Bild: Keystone/Jason DeCrow

Roger Federer, was war heute das Problem und wie sehr haben die Bedingungen Ihre Leistung beeinflusst?
Es war sehr heiss und einer der Abende, an denen du das Gefühl hast, du bekommst keine Luft mehr. Ich hatte Probleme mit den Bedingungen. Warum, weiss ich nicht, denn das ist mir schon lange nicht mehr passiert.

(https://www.20min.ch/sport/tennis/story/-Wenn-du-dich-so-fuehlst--klappt-nichts-mehr--15893498)

A dataset collected from Swiss News Medias by @Tamedia https://www.tamedia.ch/en/brands.

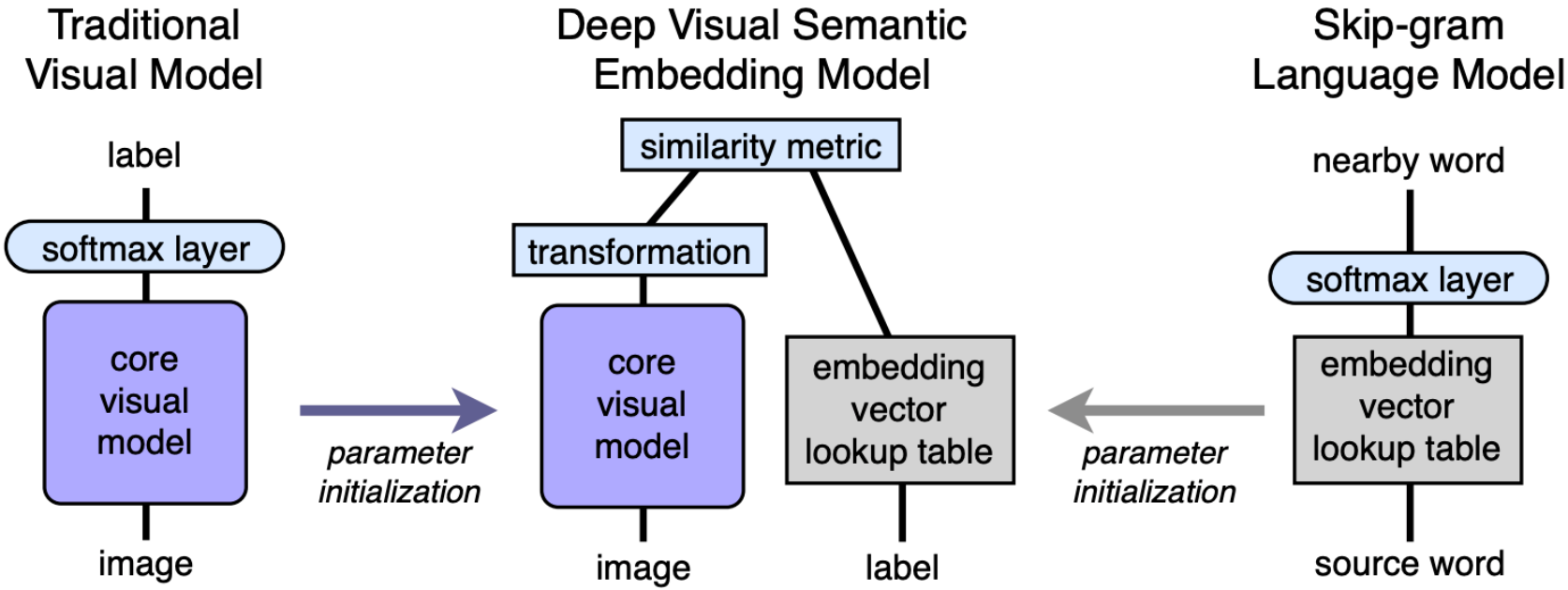
Multimodal:
We call a tuple of (image_caption,article,title,lead,image) one sample.

Multilingual:
350,204 German samples
178,270 French samples
~66.3% de and ~33.7% fr

Table of Contents

Unimodal Models

Basic Architecture



("Devise: A deep visual-semantic embedding model", Frome et al., NIPS 2013)

Unimodal Model

raw text → preprocess → list of tokens, size l

fastText initiated word & subword embedding

word embeddings, size (l, w)

Multi-Head attention

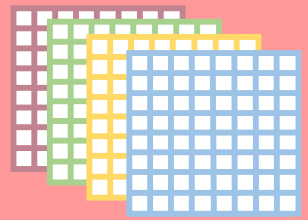
text features, size (l, w)

Pointwise CNN

text features, size (l, d)

Global Max Pooling

text encoding, size $(1, d)$



Multihead attention maps, size (k, l, l)

image

ImageNet pretrained ResNet152

image features, size $(1, d')$

Linear transformation, (d', d)

image encoding, size $(1, d)$

cosine similarity

Rankings!

trainable

frozen

Unimodal Model

Two major enhancement comparing to prior works:

- Word & Subword embedding
- Multi-Head Attention

Word & n-gram Subword Embedding

Reular word embedding:

```
if word in vocab:
   word_vec = vocab.word2vec(word)
else:
   word_vec = vocab.word2vec("<unk>")
// vocab.word2vec maps a string into
// its vector representation
```

ie. A word is either in our prebuilt vocabulary and has a corresponding vector representation or it's out-of-vocab and would be represented by a uniform "unknown" vector.

Word & n-gram Subword Embedding

Word and n-gram subword embedding:

Original word: <lausanne>

4-grams: <lau, laus, ausa, usan, sann, anne, nne>

5-grams: <laus, lausa, ausan, usann, sanne, anne>

6-grams: <lausa, lausan, ausann, usanne, sanne>

"lausanne" would be represented as the **sum** of word vectors of all of them:

{<lausanne>, <lau, laus, ausa, usan, sann, anne, nne>, <laus, lausa, ausan, usann, sanne, anne>, <lausa, lausan, ausann, usanne, sanne>}

(Emprically, speaking 4 to 6-gram works best for German.)

"Enriching Word Vectors with Subword Information", Bojanowski et al., ACL2017

Word & n-gram Subword Embedding

Subword embedding brings three potential benefits:

- tolerates typos
- works for compound words
- opens a door for tansferring knowledge in nearby languages

Tolerates typos

de: *Präsidenttttt Trumpet spricht vor der Presse.*

en: *Presidenttttt Trumpet speaks in front of the press.*

works for compound words

Stockwerkeigentumswohnungen???
(Yes, it means “condo” in English.)

de: *113 **Mietwohnungen**, 64 **Stockwerkeigentumswohnungen** und 48 **Alterswohnungen** sollen unter anderem auf einer Landreserve der Firma Hoffmann Neopac im Südwesten von Thun gebaut werden. Bald darf sich die Bevölkerung dazu äussern. (Symbolbild)*

en: *113 **rental apartments**, 64 condominiums and 48 **senior apartments** are to be built, inter alia, on a land reserve of Hoffmann Neopac in the southwest of Thun. Soon, the population may comment on it. (icon)*

Transferring knowledge in nearby languages

Name of entities are usually very similar in terms of morphology in nearby languages.
And entities are usually the most informative in news image retrieval.

de: *Der Präsident trifft sich mit seinem Kabinett.*
fr: *Le Président rencontre son cabinet.*
en: *The President is meeting with his cabinet.*

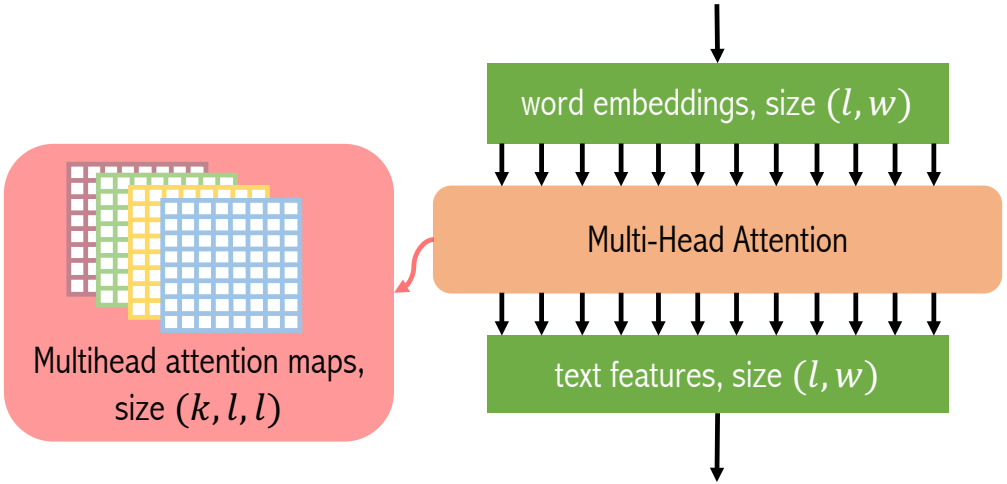
Multi-Head Attention

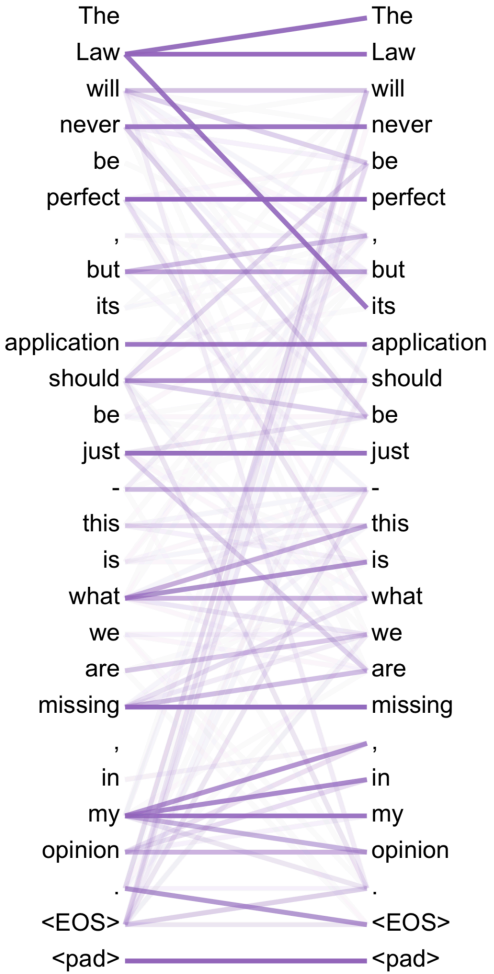
Intuition:

RNN is built for modeling an ordered sequence. But order might not be important for our task.

Think about people reading news. For the first look, they always scan it as a whole and spots only a few keywords. Order is likely to be ignored here, but attention is certainly used.

Attention Function





Every text feature of dimension would be a weighted sum of word embeddings.

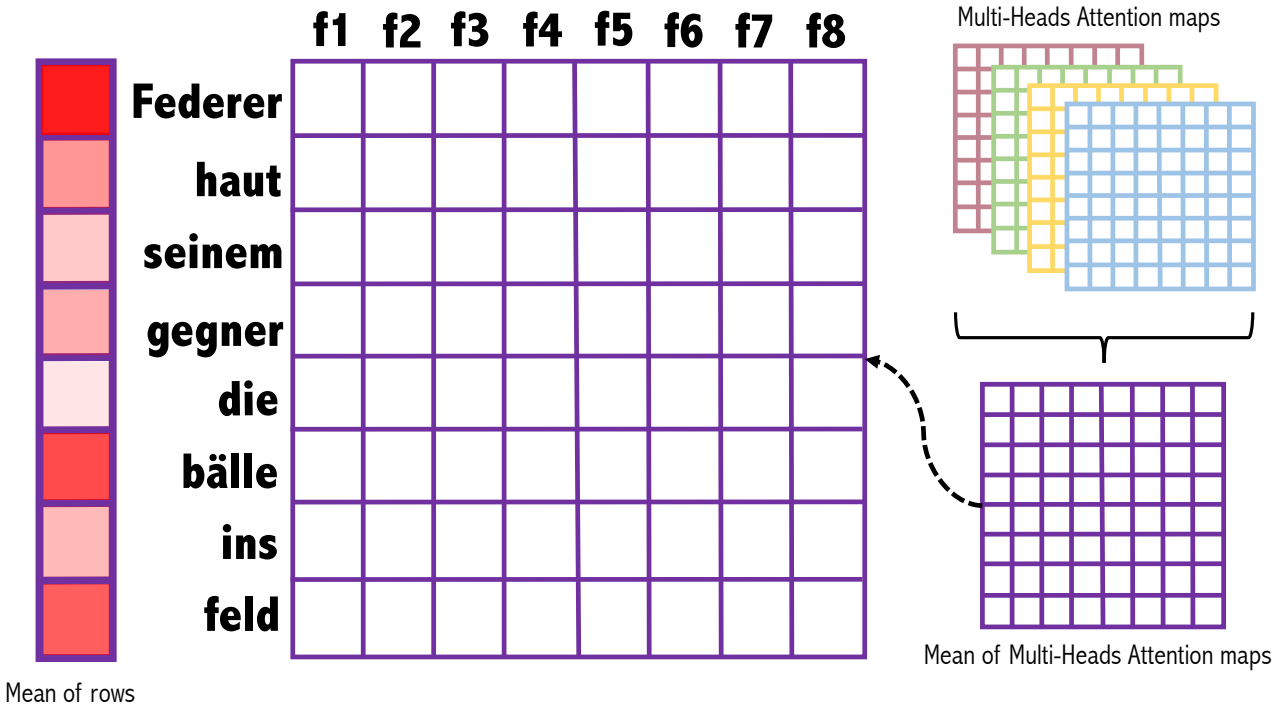
("Attention is all you need", Vaswaniet al., NIPS 2017)

Multi-Head Attention Maps

We argue that the Multi-Head Attention mechanism is acting as a query system.
Important information(words) is(are) preserved and unimportant ones are thrown away.

Multi-Head Attention maps contain the information that's needed for computing *importance* of each word.

Multi-Head Attention maps



Multi-Head Attention works like a query system

A toy example:

Im ersten Satz läuft alles nach Plan: Federer haut seinem Gegner die Bälle ins Feld, Millman ist klar unterlegen. (20min.ch)

How does it resemble a query system? Can we see a visualization of this?
Let's go have a look of some examples: https://modemos.epfl.ch/article

How does the query system work exactly?

Original:

Im ersten Satz läuft alles nach Plan: Federer haut seinem Gegner die Bälle ins Feld, Millman ist klar unterlegen.

Remove highlighted words:

~~*Federer*~~ *Im ersten Satz läuft alles nach Plan: haut seinem Gegner die Bälle ins Feld, Millman ist klar unterlegen.*

~~*Gegner Bälle Millman unterlegen*~~ *Im ersten Satz läuft alles nach Plan: haut seinem die ins Feld, ist klar.*

~~*Satz haut Feld*~~ *Im läuft alles nach Plan: seinem die ins, ist klar.*

Preserve only highlighted words:

~~*Im läuft alles nach Plan: seinem die ins, ist klar.*~~ *Satz Federer haut Gegner Bälle Feld, Millman unterlegen.*

Unimodal Model Quantitative Results

Models	R@10 - text to image				R@10 - image to text			
	image caption	article	title	lead	image caption	article	title	lead
VSE++ (baseline)	39.7	25.2	17.8	27.3	41.2	26.7	17.1	26.7
UVS (baseline)	42.4	32.6	22.2	29.2	42.0	32.8	21.8	28.6
Ours - attention (using GRU as text encoder)	52.8	-	-	-	51.4	-	-	-
Ours - subword	45.8	-	-	-	45.2	-	-	-
Ours	**54.2**	**37.2**	**24.2**	**35.4**	**52.0**	**36.8**	**22.9**	**32.2**

Multimodal Models

The more the better:

Multimodality and Hierachical Attention

Modalities:

title

«Ich hatte das Gefühl, keine Luft zu bekommen»

lead

Roger Federer erklärt seine Niederlage im US-Open-Achtelfinal gegen John Millman mit den extremen Bedingungen.

Image caption

1|7 **Start nach Mass**
Im ersten Satz läuft alles nach Plan: Federer haut seinem Gegner die Bälle ins Feld, Millman ist klar unterlegen.

Bild: Keystone/Jason DeCrow

article

Roger Federer, was war heute das Problem und wie sehr haben die Bedingungen Ihre Leistung beeinflusst?
Es war sehr heiss und einer der Abende, an denen du das Gefühl hast, du bekommst keine Luft mehr. Ich hatte Probleme mit den Bedingungen. Warum, weiss ich nicht, denn das ist mir schon lange nicht mehr passiert.

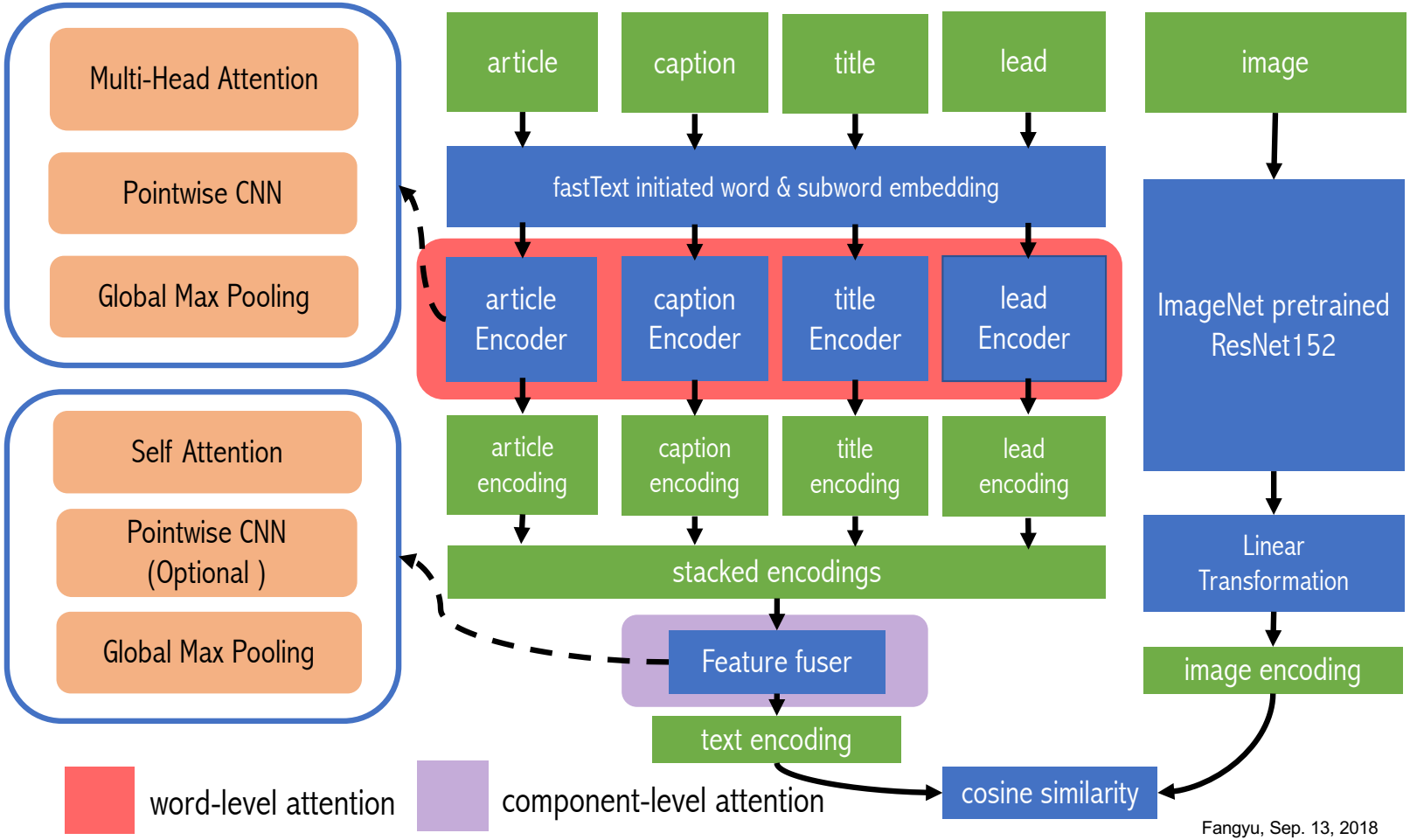
(https://www.20min.ch/sport/tennis/story/-Wenn-du-dich-so-fuehlst--klappt-nichts-mehr--15893498)

(←) Component-level Attention
& Word-level Attention (↓)

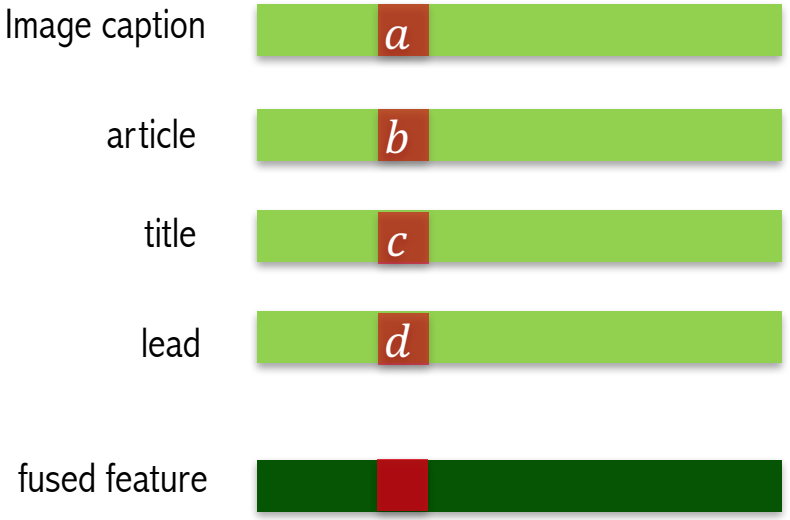
Attention scores

im ersten satz läuft alles nach plan : federer haut seinem gegner die bälle ins feld , millman ist klar unterlegen .

Multimodal Model Architecture



Fusion Methods – fixed policy



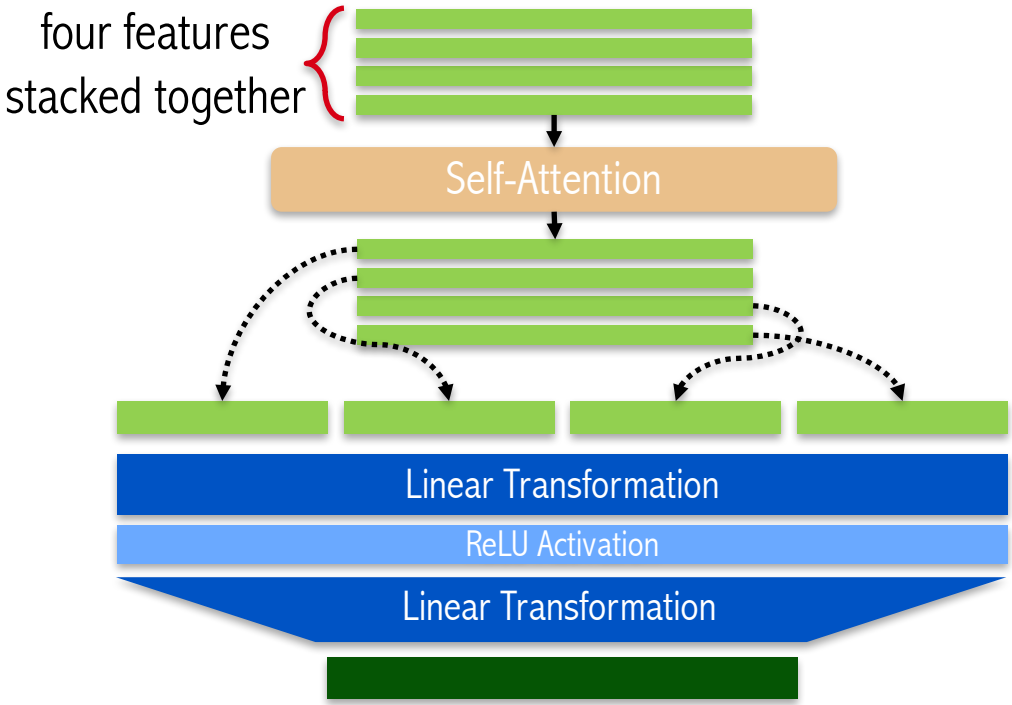
Element-wise Adding

$$\blacksquare = a + b + c + d$$

Global Max Pooling

$$\blacksquare = \max\{a\ b\ c\ d\}$$

Fusion Methods – learned policy



Multimodal Model Quantitative Results

Fusion Method	text to image				image to text			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Global Max Pooling	21.4	50.0	64.3	5.0	19.6	47.4	62.3	6.0
Element-wise Adding	21.2	50.6	64.4	5.0	**20.8**	**49.4**	**63.1**	6.0
Neural Net Fuser	24.7	53.2	68.7	5.0	17.8	47.8	63.0	6.0
Attention Fuser	**24.2**	**54.7**	**69.9**	5.0	18.1	47.1	62.2	6.0

Multilingual Models

Can we build a (very small) Tower of Babel?

- Transferring knowledge across language.

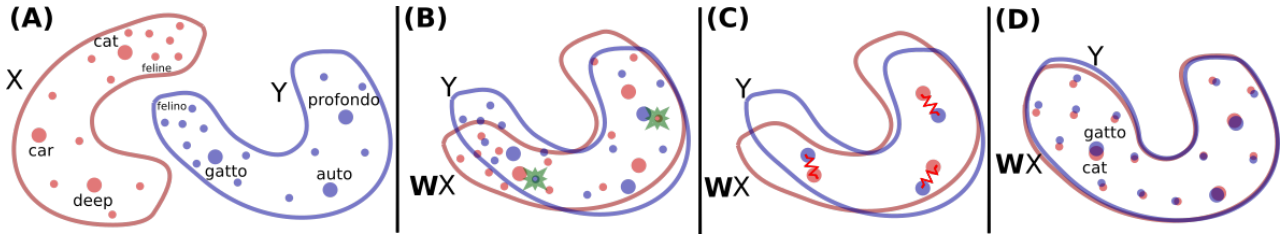
There are two ways of conducting this idea, we call them

One for All and **All for One**.

One for All

– Can we build one model works for all languages?

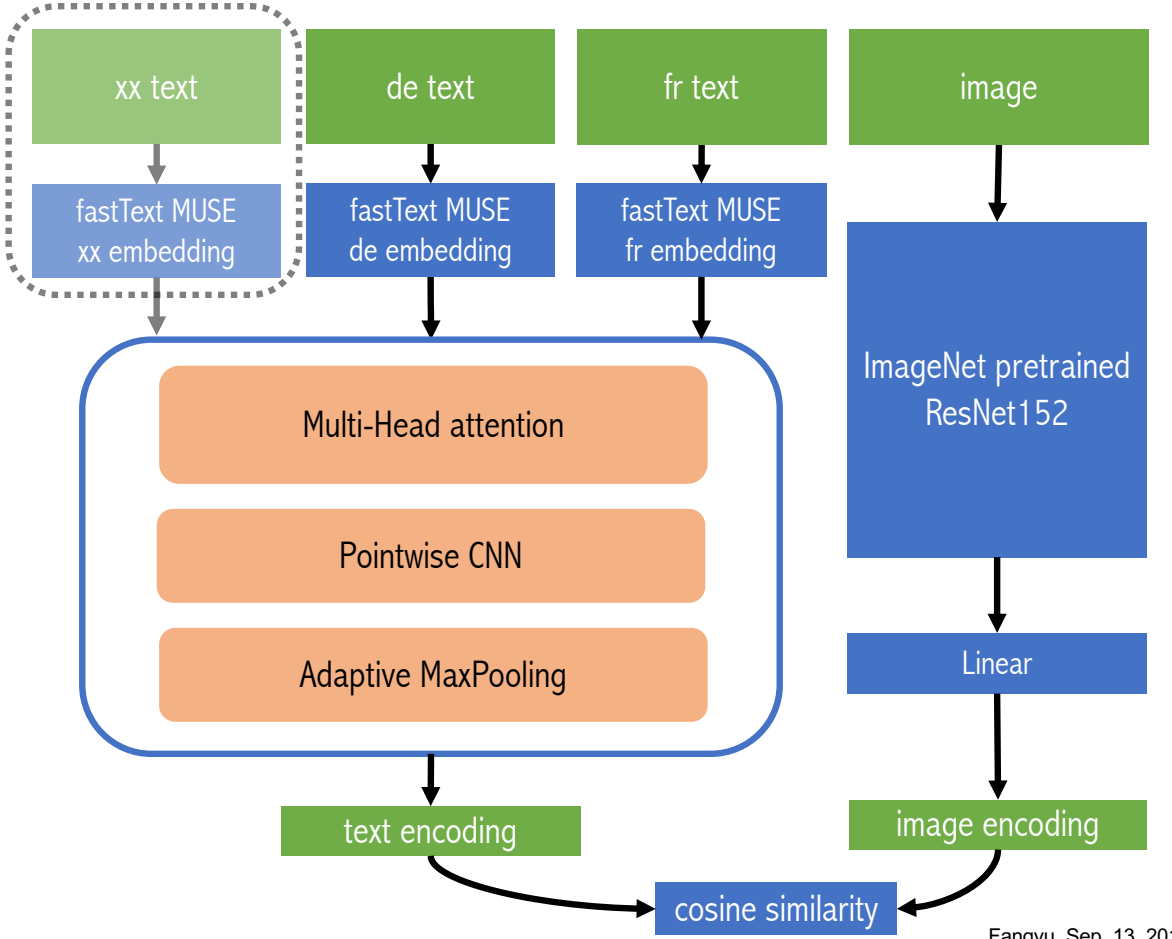
MUSE: Multilingual Unsupervised and Supervised Embeddings



"Word translation without parallel data", Conneau et al., ICLR2018

One for All

In real world application, we maybe able to train one model then only substitute the word embedding to make it work with other languages.



One for All

– Quantitative Results

training language	R@10 - text to image		R@10 - image to text	
	de	fr	de	fr
fr	-	39.4	-	38.2
de	41.1	-	41.4	-
de + fr	**43.1** (+2.0)	**42.0** (+2.6)	**43.3** (+1.9)	**41.7** (+3.5)

All for One

- Can all languages contribute to performance of one model?

We want to buid a model for language B, but would like to also benefit from data in language A…

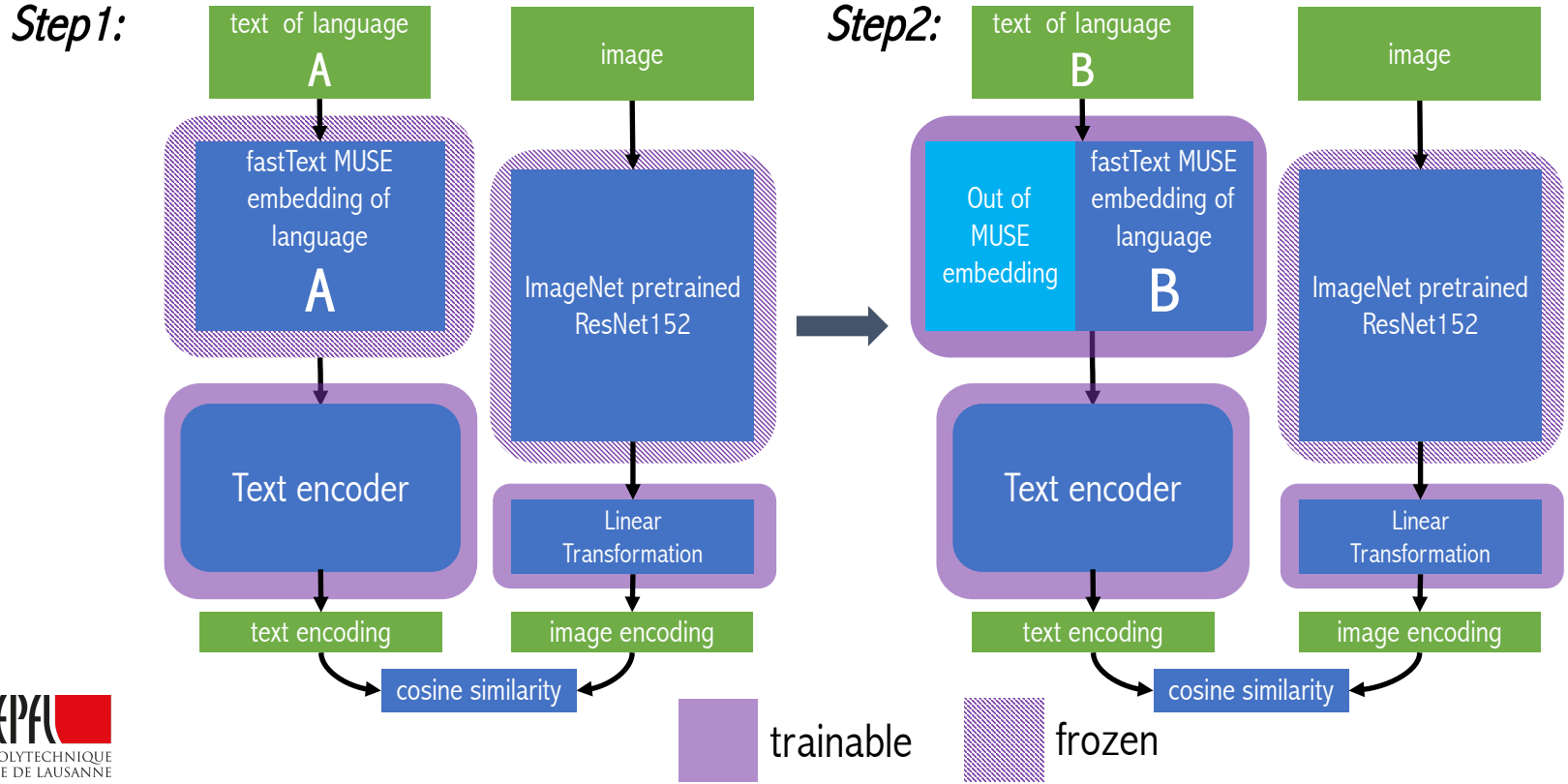
Recipe:

Step 1: Initialize word embedding with fastText MUSE weights for language A. Freeze word embedding. Train model on language A.

Step 2: Substitute the MUSE word embedding with language B's. Keep weights in other parts of the model. Train the model on language B, finetuning the word emebdding.

All for One

- Can all languages contribute to performance of one model?



All for One

– Quantitative Results

training language	R@10 - text to image	R@10 - image to text
fr	41.0	39.9
de	42.1	42.2
de (+ fr)	**49.8** (+7.7)	**48.5** (+6.3)
fr (+ de)	44.9 (+3.9)	43.2 (+3.3)
fr (+ de)	**45.1** (+4.1)	**43.4** (+3.5)

One last thing

Feel free to play with the demos
and give us feedbacks!

https://modemos.epfl.ch/article

Thank you all !